THE COMPLETE
TECHNICAL PAPER PROCEEDINGS
FROM:

100 MILLION PROGRAMS IN 3 CHARACTERS: INNOVATIONS IN SEARCH TECHNOLOGY FOR MASSIVE DATASPACES

Rakesh Barve, Ph.D.
Veveo, Inc.

Abstract

The sheer volume of video a consumer can watch on a TV in the comfort of the living room is becoming overwhelming; unfortunately, user interface design has not kept pace with the UI advancements and innovation we have seen in desktop computing. In addition, TV suffers (as mobile platforms do), from the constraints of restricted screen real estate and sub-optimal input mechanisms (on-screen keymats, directional keys, number keypads, qwerty keypads, etc.). This combination of a dynamically increasing amount of content and sub-par browsing paradigms has accentuated the need for better and more powerful search functionality for the TV experience.

Building a better search solution for the living room TV experience poses a number of challenges for the TV aggregator:

- *Massive growth of programming choices*
- *Multiple data sources to be indexed and merged for one search box/results list*
- *Metadata enhancement & consistency*
- *Easy Input – search-as-you-type, incremental results*

In the past, MSOs have struggled to improve their search capabilities and have accepted limited search functionality, relying on basic techniques such as purely lexical-based matches, (as opposed to relevance-base techniques similar to how Google works). However, most aggregators recognize the value of search and the need for improvement.

There are numerous technical challenges and user experience issues involved in building a better search solution. In this paper we propose a novel search and browse system that is based on the principle of minimizing a user's input, (i.e. the number of keystrokes), to get to the desired content. We outline how the system needs to simultaneously support maximal-recall, field agnostic querying, and incorporate enough of the user's application context that the number of input key-strokes is minimized (by squeezing as much information as possible from each key-stroke).

CONTENT DISCOVERY

Search And Discovery Components

vTap's basic *Incremental Search* paradigm helps users to easily find specific individual videos from a plethora of videos. vTap's *Universal Smart Tag Clustering Engine* creates a huge number of independently searchable collections of videos corresponding to every known person, movie, TV-show, music band, sports team, and 'micro-genre' in the world, by analyzing the text meta-content, (the subject, genres, fields and other attributes), of individual items. Each searchable collection is like a channel; there are a virtually unlimited number of semantic, subject-specific channels. Based on the nature of the

collection and current, (TV, VOD and internet), availability of pertinent items, these listings may dynamically evolve over time when the user browses them. Furthermore, vTap's *Contextual Search* allows easy search and dynamic creation of *playlists* of individual items within the limited context of a specific named collection. When deployed, this feature enhances the user's ability to browse into a particular collection's listing and then refine it further while still in that context, creating a playlist on-the-fly. In addition, vTap also uses a *Learning Engine* that observes the Smart Tags, micro-genres and keywords associated with videos watched by a user, and implicitly infers a stochastic video signature of that user that captures his taste in video, (the smart tags and micro-genres), along with TV/VOD related habit patterns and subscriptions. A *Recommendations Sub-System* then uses the stochastic signature of the user, (inferred by the Learning Engine), and smart tags therein, as well as explicit smart tags set by the user to recommend relevant individual videos encountered in the server's dynamically evolving TV/VOD/streaming video database. A *Usability Personalization Sub-System* uses the stochastic signature, (the smart tags dimensions as well as specific habit related information), to make possible personalized reordering in Search, Browse and Actionable listings that improve usability in significant ways.

Incremental Search – A New Search Paradigm

The way vTap Search enables users to easily find content is via the network-based *incremental search* technique – for each character input by the user, (without hitting a 'return' or 'enter' button), a network search operation instantaneously returns a new set of corresponding results that enable the user to 'converge' to the desired result with minimal text input. vTap's incremental search is fundamentally different from other approaches, which incrementally help users complete search queries by performing a strict prefix-based lookup on a database in which each record is simply a query made in the past. In contrast to these approaches which return completed *query candidates,* one of which needs to be picked up by the user to launch the real search that returns results, vTap Search directly returns *results* which can be meaningfully acted upon.

In vTap Search, each record is not a simple string, but is in fact a composite structure comprising multiple text fields – for example, title, tags, descriptions, source-sites etc. Specifically, vTap incremental search is characterized by the following attributes:

- On any single-word or multi-word query, it simultaneously checks across all fields for prefix-matches with any of the query words.
- Not only are multiple query-words in user's incremental input string allowed to be incomplete, these incomplete query words can each prefix-match a distinct word in a distinct field of a composite record.
- Having multiple fields inside a record allows for calculating more detailed match-scores, based on the matching fields, which in turn helps to better rank results. Field match scores can be based on the semantics or importance of the fields from a search-recall perspective.
- Features such as giving results even when the query has spelling errors.

Incremental Search Intellectual Property

One way to implement the incremental search just described is to simply scan every word in every meta-content record of the video database and then display matching

results. When the database of records is small, (e.g., hundreds of names in a contacts database), naïve techniques will suffice. However, clearly this approach of going through every word of every record in the video database for every query does not scale, (say, for tens of thousands of movies/music artists/cast, millions of music tracks and internet videos etc.). The fundamental innovation in vTap Search is the mechanism by which only a small fraction of items need to be looked up in order to return meaningful results for incomplete queries. All of the following work in unison as part of vTap incremental Search:

(a) the video and cluster/collection popularity computing algorithms for different kinds and types of videos, (vTap uses a variety of heuristics and internet sources to assign popularities as described in the next subsection),
(b) the pre-processing, static indexing steps and the term relevance logic which blend video and cluster popularity with the relative importance of the meta-content fields that contain the term,
(c) the custom runtime incremental search data structures,
(d) the runtime query processing and term relevance re-calculation based on how and which meta-content fields matched the query terms
(e) runtime relevance changes based on scheduled time, subscriptions--availability-dependent, (e.g. if a program VOD window just expired), TV schedule attributes such as LIVE or reruns etc., and user-signature dependent dynamic runtime relevance changes and,
(f) the custom runtime computation steps that bring it all together.

vTap Video and Cluster Popularity

vTap Search applies to text based search for items that may be TV, VOD or internet videos or named smart tag clusters of those videos.

The vTap service's video database of available TV and VOD videos is updated regularly using TV and VOD guide feeds available from various data suppliers. In addition vTap uses a variety of other internet sources to secure richer information on every conceivable movie and TV program. Internet link, modified page rank and rating based heuristics are used to assign popularities to movies and TV programs from various internet sites first. Movies and programs available in VOD and TV listings of the vTap video database are correlated to corresponding entries in a universal database of all movies and TV shows, and these listings then derive a base popularity metric from the corresponding entries of the universal database. They also derive enhanced text meta-content, (more terms for indexing), based on the correlated entries. TV show popularity can be further adjusted by taking into account when the show is telecast, the nature and popularity of the network telecasting it, whether it is a current TV show or not, the cast in the show and its popularity, and so on. Similar rules can be applied to VOD content, except there is no telecast time there -- although there are notions of "recently available."

There are different kinds of internet videos based on their source. Some videos – e.g., news, sports and those from content companies, have a dominant recency component in their popularity computation and are also topical in nature based on meta-content and the popularity and authority of the site producer. Videos from other content companies depend roughly on how well

known or popular the consumer facing content website is. While calculating the popularity of user-contributed internet video, (since most videos do not have external referring links or very few links), the notion of internet page rank or popularity is of limited use. So for user-contributed videos, one must make use of 'social graph' information, internal page ranks among the site's uploading users and subscribing users in various ways, and statistics such as view counts, number of comments and recency. Many popular video sharing sites have content partners uploading videos regularly, and the popularity of these videos is again a function of partner popularity and recency.

The popularity or relevance of specific collections, (Smart Tags), of videos is a function of the corresponding topic's 'popularity' overall and in its domain, and also the temporal properties of the videos associated with that topic.

Term Relevance For Each Term In The Text Meta-content Of A Video

The vTap paradigm of incremental search makes it very important to assign appropriate 'search' weights to appropriate keywords, terms and meta-content in the specific subfields associated with a video. Text meta-content for TV and VOD listings have fields such as title, episode title, keywords and crew members -- correlation with internet sources of movies and TV shows can enrich this meta-content. Internet videos typically have titles, descriptions and tags. User generated videos have very basic meta-content, whereas videos from content companies have editorially created richer meta-content.

Web search engines use information and signals embedded in the keywords and associated with various html tags and page structure at the time of indexing. When indexing videos from TV, VOD and the internet, Veveo can use the fact that text in different fields, (title, description etc.), is of varying importance. The basic advantage of creating a structured database of multi-field records, is that it allows the same record to be recalled using multiple associated names and keywords with different search weights: for example, also-known-as names and, (where appropriate), abbreviations can also be used to index the same record. Additionally, some of the fields -- e.g., running time, video format, dates etc., allow for a filtered or sorted view of search results.

vTap Search 'Number Mode'

When using vTap incremental search from devices such as PDAs that have a QWERTY keyboard, the user's input is unambiguous as each keystroke is mapped to an individual character. However on most phones and on all conventional TV remote controls, there is typically only a numeric 'phone' keypad and inputting a single character using these devices may frequently require pressing the same key two, three or four times. The associated inter-digit timers make the process of inputting a query to a search engine from numeric keypads both error prone and cumbersome. Contrast this with the relative ease of composing SMS or text messages in the so-called 'predictive-text mode.' This ease is a result of the fact that more often than not, the user presses a single keystroke for each character of each word in the composed text message. However, as far as the search application goes, assuming the database being searched is a huge web-scale database, even the 'predictive-text mode' technology used to create text messages from such phones is completely useless in practice. This is because the device resident 'predictive-text mode' technology is premised on the practical assumption that there is a *very restricted list of words* that constitute the 'working set' of words a user

will draw from in order to create text messages. In stark contrast, the set of terms that a user has at his fingertips to launch a search query is a *mammoth set*, comprised of all kinds of noun phrases, all (full, short, nick) names and proper nouns of every entity, (including movies, TV-shows, bands, songs etc), or persons that have been named, 'user ID's of various kinds, words and terms from all languages, including multi-word noun phrases and so on. This renders the 'predictive-text' approach completely unworkable in a web-scale search application.

In order to *dramatically ease* the experience of using vTap incremental search from such numeric keypads, vTap provides a 'number mode' -- based on Veveo's proprietary technology geared towards the incremental search principle -- wherein the user needs only to press a single key, (the numeric digit corresponding to the character), for each character. The ability to input queries quickly, especially without having to pay any heed to inter-digit timers, and then receive results in the vTap incremental search style is what dramatically enhances the search experience from phones and TV remote controls. Even though the user's query is inherently ambiguous, ('227' could mean 'car,' 'bar,' etc.), vTap computes and blends results corresponding to all meanings of the multiple input query tokens according to highest relevance. Even in the number mode, vTap uses all aspects of its incremental search algorithm so that *all of the features mentioned work in the number mode,* including auto-corrected spelling matches. vTap displays matched strings to clearly enable the user to recognize the records of interest among the results.

A Comparison Of vTap's Character-based Incremental Search vs. 'Standard' Word-Based Commercial Search Engine Techniques:

Since the results change with each keystroke input by the user, the user can get to the result of interest with a minimal number of keystrokes, (the query words are allowed to be prefixes of words in the result). This creates a tremendous advantage over the cumbersome character input on devices like phones, TVs and Mobile Internet Devices.

The technology and algorithms used to implement vTap's incremental search are distinct from the ones used in standard word-based search engines. In the latter, the user types one or more words and hits an 'enter' button so that the user's input step and results step consist of *discrete distinct phases*. Because the query almost always contains full words, it can essentially look up pre-computed lists of matching records, (one list for each word), and finds the best records amongst those that are present in all or most of the lists. Moreover, the items that are indexed are not records with fields but unstructured html web pages.

On the other hand, in vTap, the input and results processes are finely interleaved and interactive, so that the user may get to the desired results with a minimal number of input keystrokes. So, at every keystroke, an implicit query is launched to the server and in general, the query may consist of a space-separated list of multiple incomplete prefixes. Using the same approach as the one described above for word-boundary based search would require the pre-computation of one distinct list of records for each prefix of each term in the system. This is infeasible even on huge infrastructures when the number of indexed terms and records in the system is large enough. In order to provide

the incremental effect, vTap's search technology uses a unique set of search data structures and algorithms in a special purpose information retrieval system focusing on instantaneously returning relevant results in response to each keystroke input by the user. The key aspect that enables this system to very quickly compute relevant results for each keystroke even when the search database is massive, is the way it combines offline relevance computation, search data structures, and runtime algorithms; in that way, the result computation process has to examine only a small fraction of the entire search index.

vTap's Universal Smart Tag Clustering Engine

vTap's Clustering Engine provides a practically unlimited number of specific collections of videos which are the rough analogue of conventional 'TV channels.' In order to facilitate very diverse *and* specific interests, vTap provides not just the usual broad categories of TV/VOD/ streaming videos, (news, sports, etc.), but also a huge number of specific and *dynamically evolving* collections of video items. Each collection corresponds to a vTap Smart Tag. Basically, every named person, movie, TV-show, music artist, sports team or entity is tracked by vTap in a Universal Smart Tag database. As and when the system discovers more smart tags, these are added to the UST database. The main motivation is that in the entertainment domain, the user often wants to get to a specific set of items or a topic and then browse through the items all at once, rather than always performing a search for each single item. Furthermore, the UST based clustering forms the basis of capturing the video taste and video interests in the *Learning Engine,* which is used in the *Recommendations sub-system, Personalized Usability sub-system* as well as in analytics. In vTap Search, named collections, (each collection corresponds to a smart tag), and single items are appropriately blended in the search results, so that each collection can be independently retrieved using incremental search based on the associated meta-content.

The clustering engine works by analyzing the meta-content crawled and mined for each video or clip, and then using clustering techniques to associate each video with one or more named specific smart tags, or broader 'genre' categories. Examples of specific smart tags include George Bush, the Boston Red Sox, Tom Cruise, Jay Leno, Herbert Von Karajan, Richard Feynman, Otto Von Bismarck etc. In short, the clustering engine creates specific collections representing meaningful Smart Tags from domains such as news, science, music, sports, movies, TV, as well as topics based on personalities and other named entities.

As a result of the dynamic nature of the TV schedule, VOD asset availabilities and changing websites, the Smart Tags are also dynamically evolving: new smart tags are added as part of the TV/VOD availability changes, crawling, discovery and clustering processes. The relevance of the collections among search results also varies: video items belonging to any given collection may change based on new videos that get discovered and fall into that collection, old videos no longer available on TV or VOD or whose web links become stale have to be removed, and the ordering of videos themselves may be optionally changed when the user decides to browse a collection. Examples include: news oriented topics listing the latest videos on top; sports topics supporting a dichotomy between videos corresponding to scheduled LIVE games, videos featuring highlights of the games themselves, (as opposed to more generic news videos on the same); movie topic previews and interviews, as well as VOD and TV availabilities of full movies and so on.

In vTap, the default organization of videos of a collection may depend on the topic of the collection as described above, and be segmented into TV, VOD and internet video sections. However, vTap can also allow the user to sort and browse the videos in each section as appropriate.

vTap's Contextual Search And User-Specified Dynamic Playlists

As previously mentioned, named collections, (i.e. Smart Tags), of related videos are included as individual search results to a user query. This is useful especially when the user's intent is tentative, which is more typical when the user is in the mood for entertainment as opposed to conducting a typical internet search. For instance, the user may start off with the intent to watch Jay Leno videos in general, but after watching one or two, decide to watch specific Jay Leno videos lampooning George Bush.

Using vTap, a user can quickly get to the Jay Leno video collection within a few characters and then begin to peruse the list of Jay Leno videos in the collection. After watching one or two, vTap allows the user to very intuitively conduct a *contextual incremental search* within the collection by typing in more characters. So, for instance, the user could start typing in 'Bush' and vTap would then pull up Jay Leno videos that mention George Bush in their meta-content. Going further, the user can easily play this list of George Bush videos one after another as a *video playlist*. Effectively, this is a user-specified dynamic playlist created by first tentatively focusing on the full subset of Jay Leno videos, and then on-the-fly doing further incremental search to refine the playlist. Enabling this user experience and fine grained user-control is a capability unique to vTap Search.

The playlist aspect of this feature is more relevant to internet videos that can be instantaneously streamed, but contextual search makes sense even for TV and VOD -- for instance, a user may enter the collection of available Seinfeld episodes, browse through them, and then continue to narrow down to one episode by making a search query restricted to that collection.

vTap Learning Engine

vTap's learning engine analyzes, for each user, the history of videos and collections that have been viewed, the search history, the smart tags, micro-genres and keywords associated with the viewing history, and the preferred TV channels, VOD subscription packages or web sites corresponding to the videos. From this information, it creates a 'stochastic signature' that can be said to represent the video viewing habits of the user. In its base case, the stochastic signature captures the smart tags, (personalities, TV shows, music artists, sports favorites etc.), micro-genres and TV channels, VOD subscriptions and websites that the user has a tendency to view when it comes to video. In a more advanced deployment, the stochastic signature can even capture the time of day and day of week that the user typically views videos associated with each smart tag or TV channel etc. The Learning Engine architecture is general enough that given adequate information, the stochastic signature can capture specific TV/VOD viewing habits of the user such as whether he watches HD or SD, paid or free VOD, and other habits that the service provider deems worthwhile.

The stochastic signatures across all the users can potentially also be leveraged to answer analytics related queries for the Service Provider -- for instance, what is the correlation between users who watch paid VOD movies and users who watch subscription sports channels on TV?

vTap Recommendations Sub-System

vTap Recommendations for discovery have two distinct flavors, explicit and implicit. In 'feed'-based recommendations, the vTap backend system allows users to explicitly specify a list of smart tags they are interested in, and whenever vTap comes across any new TV, VOD or internet video in its backend database that is associated with the user's smart tags, those videos are sent across among the Recommendations. This allows vTap to flag user videos pertaining to the user's explicit interests, and enhances the probability that the recommended videos will be watched.

vTap's Recommendations sub-system also uncovers information that is more implicit in nature, such as information about unviewed videos that share the user's stochastic signature characteristics, and are available in the user's subscribed TV channels, VOD package or somewhere on the internet. In addition, the Recommendation sub-system can employ collaborative filtering techniques to cluster 'similar' programs/videos, (or 'similar' users), based on their viewing habits. The Recommendation sub-system can also recommend to one user other interesting videos 'similar' to the videos or TV shows from his stochastic signature, or similar to the stochastic signature of other 'similar' users.

vTap Usability Personalization Sub-System

The Usability Personalization Sub-system is simply a system meant to apply the stochastic signature, (computed by the Learning Engine), to personalize and reorder various Search, Browse and Action listings in various contexts.

For instance, the stochastic signature tracks the list of TV channels, (part of the smart tags), as well as the list of TV shows, crew etc. derived from the user's viewing history. This means that to present a user with a list of (implicitly computed) Favorite Channels or Favorite TV Shows etc., all that is required is a lookup in the stochastic signature. On the other hand, a personalized reordering of the Movie channel listing in the TV guide application, is a matter of taking the original list of Movie channels and then hoisting above other channels all the channels that appear in the user's stochastic signature. Personalization for search re-ordering, (e.g. search followed by a one-click action to tune to a channel), enables the particular user who simply types a 'c' to bring up 'The Church Channel' before the globally more popular CNN. This is achieved by taking the first few pages of search results and ensuring that results corresponding to the stochastic signature are hoisted up. However, each reordering needs to be done carefully, to avoid hoisting search results that have low relevance terms matching the query.

If the stochastic signature is rich enough to capture the HD/SD preference of the user, then whenever a channel or program appears on multiple tuner numbers, the default one-click tuning action can be made to resolve to the tuner that matches the user's HD/SD preference. If the stochastic signature is rich enough to capture attributes such as free versus paid VOD content, it can be used to reorder VOD movies browse listings to match the user's preference. In general, several other usability personalizations can be achieved by ensuring the stochastic signature captures the relevant preferences, and by looking up the stochastic signature at the time the user is actually searching, browsing or about to select an action.

INTELLIGENT AND SCALABLE WEB VIDEO CRAWLING

Veveo crawls videos from thousands of different websites and presents these in its search play list results. Currently, vTap's

growing video index consists of at least 250 million internet videos. Crawling the web for videos is fundamentally different from crawling the web for web pages for the following reasons:

- In the web search application, when the user types any set of words, the entire web page is the result. In video search, only the specific video is the intended result, and this is tedious when the page contains multiple videos and the crawling and information extraction system must ensure that the meta-content corresponding to one video is associated only with that video and so on.
- Moreover, the crawling system must be able to semantically interpret specific strings in the meta-content, unlike in traditional crawlers. Therefore 'today' in the date field must be converted to the date value on the day of crawling: similarly, clip attributes, e.g. tags, various clip dates, user-IDs, descriptions, genres, and various statistics need to be extracted with a deeper understanding of the content layout and offering, sometimes specific to the site.
- Very often, the crawling system involves the ability to analyze scripts/programs to automatically detect the presence and play-attributes of video links.
- The video link that one comes across could be permanent or transient -- again a deeper understanding and automated analysis of the site is required to ensure the stored URL is permanent and not transient.
- Re-crawl systems have to be optimally and specially designed to ensure that the 'liveness' or 'staleness' of a video URL is known correctly.
- A re-crawl process design is also necessary to track various statistics associated with the video and other videos linked from and to it.
- Where available, the nature and number of uploading users is information that is crawled and stored.
- When there is a change in the crawled page's format, there has to be a real-time change in the crawler to adapt-- this is particularly important, because unlike generic web crawlers, video crawling requires site-specific and semantic interpretation of a web-page's content.
- We need to deploy proprietary techniques of generic, statistical, template-based 'visual' crawling, specialized blog crawling techniques and also other proprietary tools that have been developed to incorporate site-specific understanding of web-content in a rapid and scalable fashion.
- In the context of social networking and sites supporting user-generated video, one important metric for the "value" of the crawling system is the breadth/depth of video clip coverage. Since any crawl system may never get to know the absolute state of the system/website it is crawling, (the actual number of videos present in the crawled site), a statistical estimate of coverage is needed to quantify the breadth/depth of coverage. This statistical coverage, (video clip coverage and user coverage), is measured on an ongoing basis, and the results are used to synthesize crawl schedules and maximize coverage.

In the context of term relevance and recallable information for internet video, it is

interesting to consider the impact of terms emanating out of clip attributes such as user comments, data from subtitles/closed caption of the video, and video-speech to text etc. Such information has the potential of introducing 'search term noise' -- misplaced emphasis on words and terms that are not likely to be used by the user to recall the specific video. On the other hand, there are some situations where these terms may distinctly add value, for example, detecting either 'adult' or 'pornographic' content. Similarly, when one can compare audio and video signatures, speech processing techniques also help to identify and mitigate the proliferation of copyright content. When one considers long-tail content, speech processing is a double-edged sword -- in case of long tail videos that lack accompanying meta-data, speech processing in the audio track of the clip can sometimes enhance the search/recall function by enhancing the meta-data. However, it can also add irrelevant meta-content corresponding to sets of spoken words that are insignificant to the clip. The semantic understanding of media for "popular' material is, in the best case, useful for "research" purposes, (a journalist trying to find a specific utterance somewhere in a five minute clip), and in the worst case introduce lot of search noise.

V360 – A 360 DEGREE VIDEO ANALYTICS SOLUTION

The explosive growth in online video consumption provides a unique opportunity for Service Providers, (whose access networks are the source of this traffic), to gain a new understanding of consumer video behavior that can yield strategic as well as tactical actionable intelligence. What Service Providers have is reams and reams of video (TV/VOD/Internet Video) related tune-streams/click-streams. Fundamental differences in internet browsing on the one hand, and consumption of video on linear TV, VOD or internet streaming on the other, require specialized analytic tools that go far beyond standard internet analytics such as sites visited, time spent, uniques etc. to obtain actionable intelligence regarding video consumption habits. An analytics product must enable its users to very easily, meaningfully and flexibly zoom in on arbitrary subsets and views of the database, (in this case, the database of video tune-events, click-streams, users, programs etc.), get meaningful statistics about that subset, mine correlations and relations between different subsets and do all of this in the most user-friendly, intuitive and scalable fashion. The more domain knowledge incorporated into the database and analytics model, the more meaningful are the tools and results presentation in the analytics product. The semantic knowledgebase contributed by the Universal SmartTag Technology, coupled with tremendous information from the clickstreams representing end users' content navigation behavior, constitute the basic pillars of a semantic video analytics product that provides true insights into consumer behavior.

Unique Positioning And Expertise For Video Analytics Product

The V360 platform's uniqueness is that its video-specific analytic tools are based on a deep understanding and proven expertise in two principal aspects of video consumption by users.

The first aspect consists of the specifics of the service provider *video consumption environment*: Veveo's analytics model incorporates the different consumption modes, user-habits and business relationships involved in the video delivery system. Video can be consumed by a user via basic linear TV, time shifted DVR, specialized channel packages of linear TV, PPV, free or paid VOD, as well as many sites on the internet

via PCs, Mobiles or internet set top boxes. Temporal patterns of video consumption based on time of day, week-days versus week-ends, as well as time-shifted and broadband video are the outcome of decades of conditioning to a dominant television culture and more recently convergence, time shifting mechanisms and internet video as a phenomenon. The influence of an interactive TV guide design on the accessibility of a program is significant. Video consumption is also significantly influenced by access networks and behavior exhibited by consumers related to the device platform -- short clips are preferred on mobile platforms, interactivity is the preferred embodiment for desktop computers, and passive consumption dominates TV. In the context of TV, VOD, and internet streaming, new business relationships between content creators, content distributors and service providers need to be, and are currently being, established.

The second unique aspect of V360 is its fine-grained characterization of what videos a user is watching. In this respect, the V360 analytics platform provides a continuously evolving *fine-grained semantic taxonomy* potentially at an individual level, of the hundreds of millions of video assets available in archives, on air, and available from on-demand servers, across internet sites. The semantic classification of TV/VOD programs, movies and internet videos into micro-categories and millions of tags and clusters is based on Universal Smart Tag (UST) technology and the V360 content database engine, and goes well beyond the limited TV-program segmentation that stems from classic TV analytics regimes. A fine-grained classification engine based on millions of Smart Tags becomes necessary especially for analyzing and understanding consumption in destination sites such as Youtube, DailyMotion and so on -- each of these can be deemed a 'video internet' in and of itself, and each of these sites' millions of uploaders is potentially a distinct channel with its own flavor. The Universal Smart Tag technology is a core component of Veveo's commercially deployed vTap internet product, which services millions of video streams and page hits to internet-enabled phones on a daily basis.

Applying Video Expertise In V360 Analytics

As mentioned previously, a business analytics engine must arm the user with the most meaningful domain dependent tools in order for the user to naturally filter and sift through raw data and mine actionable information out of that data. V360 tools based on the video consumption environment enable a focus on specific subsets of programs based upon: which media companies (with whom the service providers need to negotiate) produce the content, channels airing the content, channel schedule attributes such as airtime (time of day, day of week), LIVEness, VOD attributes such as paid or subscription, usage based filters (programs viewed by at least some number of viewers) and viewer subsets (programs viewed by consumers who spend on paid VOD content.)

V360 also provides ways to filter programs based not only on the program meta-content (i.e. genres, cast members and types of programs available in TV and VOD listings), but also on the Smart Tags associated with the programs. Enhancing the meta-content enhances the tools and filters for selecting programs using V360. In addition, V360 enables sifting through and understanding at a higher level the typical habits of users, for example, do people who buy subscription packages on TV typically watch new and paid movies on VOD? It also enables understanding clusters of users whose viewing habits are similar, (watching similar TV programs or movies), but in two distinct partitions – one cluster using VOD/Pay per

view/Premium Subscription and the other cluster using only free/basic subscription; identifying clusters of users who can be targeted in a marketing campaign that promotes certain kinds of subscription packages based on their affinity to their long tail interests, etc.

Arbitrary subsets (clusters) of users and programs can thus be easily specified in V360 and the relevant statistics for those subsets can be browsed. Additionally, V360 can correlate events in different subsets to uncover new relationships between entities.

The V360 platform provides tremendous flexibility in defining business rules. It also allows Service Provider personnel to pose specific pointed queries based on details of their own network's video consumption environment, and on rich fine-grained UST-based semantic classification, a capability that cannot be replicated in general purpose analytics tools.

Scalability And Trading Off Accuracy And Response Times By Sampling

The V360 analytics correlation engine is based on a distributed computing backend architecture that scales to peta-bytes of data, and then briskly sifts through vast tables of Service Provider video consumption data to deliver the desired analytics information. In addition to enabling video events to be visualized and comprehended at a higher level, the V360 platform also detects various patterns-- examples include Content Affinity Clusters, (related videos watched by users if they watched a specific topic), User Affinity Clusters, (a cluster of users with similar video behavior), Language/Ethnicity Affinity clusters etc.

One of the unique aspects of the V360 platform is its ability to support analysis of the data in a completely deterministic fashion, by considering the entire data set, or in a stochastic fashion by considering a subset of the original data set. The subset is sampled from the original data set using a sophisticated set of statistical sampling algorithms that are fine tuned to the idiosyncrasies of the underlying domain, while guaranteeing accuracy within a specified bound.

Although the analytics engine is highly scalable, sometimes speed or response time is more desirable than exact accuracy. V360 enables the user to choose among various progressive modes that trade accuracy for speed by using statistical sampling techniques that guarantee bounded errors. The statistical sampling techniques provide a way for the user to obtain a quick answer about a very large data set, (for example, two years worth of viewing habits for a very large metro), while a more accurate answer can be scheduled to be computed at a later time.

V360 Insight: Intuitive And Hands-On Analytics

Combining multi-touch development platforms such as the Microsoft Surface, with V360's domain specific database and analytics platform, Veveo is building an exciting product, **V360 Insight,** that enables executives and business owners to directly and intuitively interact with user behavior data on multiple dimensions and get responses to ad hoc arbitrary queries. V360 Insight empowers executives with real-time access and control over massive customer data, without requiring the IT department to create new reports each time a new query needs to be answered. Queries can be dynamically composed and analyzed through intuitive visualization, without any third party involvement. Veveo believes such an interactive and direct hands-on system will dramatically enhance an executive's understanding of customer viewing behavior

with respect to programs, channels/networks, media companies etc.

Platform Aspects

The V360 product is a platform that enables several diverse applications that can be built by third parties based on analytics data and information. The API exposes the semantic information at various levels of granularity across Space-Time for applications such as: Data Visualization, TV/VOD promotions, TV and VOD Recommendations, TV Program popularity and ratings based on exact counting, Targeted and Personalized advertisement insertion, and so on.

Conclusion

As connected multimedia devices proliferate in the next several years, the ability to easily and seamlessly discover and consume media from the TV and VOD world on the one hand, and the internet video world on the other, will be one of the primary factors that determine end user adoption of videos and video services. Veveo believes its vTap technology will enable service providers to meaningfully engage the 'on-demand' consumer generation as it taps the plethora of video across a multitude of input and display constrained devices.

256-QAM FOR UPSTREAM HFC

Robert Thompson, Jack Moran, Chuck Moore, Mike Cooper, Robert Howald, Ph.D.
Motorola Home & Networks Mobility

Abstract

Increased throughput demands, driven by applications like Peer-2-Peer file sharing and social networking, has intensified the demands placed on upstream spectrum. Those demands have been met with advanced DOCSIS tools like SCDMA and Channel Bonding. Additionally, plant architectures are evolving towards fiber-rich networks with reduced RF cascades, potentially improving overall plant performance and creating opportunity to support higher-order modulation schemes.

The benefits of advancing modulation to 256-QAM over 64-QAM is well-understood for the downstream. For example, a 33% throughput increase would also apply to the upstream. As previously explored for downstream spectrum, this throughput increase comes at the expense of increased sensitivity to noise, distortion, and interference. However, the upstream spectrum hosts a different class of impairments as well as DSP tools available to overcome them including equalization, forward error correction, spread-spectrum techniques, ingress cancellation, and interleaving.

The goal of this paper will be to identify the critical engineering requirements for supporting 256-QAM in an upstream environment and the implications for the HFC network performance.

INTRODUCTION

Upstream Service Growth

Growth patterns in Hybrid Fiber Coax, HFC, upstream have been described using Moore's law in [1] to show that service rates increase by a factor of 10 every 5 to 7 years. Specifically, demand for today's service rates, which are in the range of 2-10 Mbps, will increase to approximately 20-100 Mbps in 5 to 7 years. DOCSIS 2.0 links will support the lower end of the 5 to 7 year projections. However DOCSIS 3.0 technology, with channel bonding was designed to help cable operators deliver a 100 Mbps service rate. It was shown in [1] how S-CDMA could help cable operators use spectrum below 15 MHz to achieve this goal. This paper proposes another possible solution through the use of fewer, but more bandwidth efficient signals that will leverage modulation schemes such as 128-QAM or 256-QAM.

HFC Evolution

Technological enhancements and increasingly competitive pressure on cable operators to deliver more capacity is resulting in fiber-rich architectures with reduced RF cascades, shown in [2]. These developments may create opportunity to use higher-order modulation schemes.

Assuming identical upstream RF amplifiers, a cascade's signal-to-noise ratio, SNR, based upon noise figure, NF, of the RF amplifiers could improve by approximately 3 dB with reduction to half as many cascaded actives. The effects of non-linear distortion, specifically composite-intermodulation-noise, CIN, may also be reduced by approximately 7 – 9 dB under similar circumstances.

However, upstream performance is not necessarily limited by cascaded performance, but rather the upstream noise introduced at both intentional and

unintentional entry points. Noise comes in multiple forms, including ingress, impulse, etc. and ultimately places a greater limit on upstream SNR due to the high input levels of the upstream hybrids.

Shorter cascades reduce the impact of noise funneling, improving upstream SNR perhaps further. It has been seen that reducing the cascade by half could reduce the number of actives fed from a node by 4 or more. Generally speaking, SNR changes can be described as a function of the total number of actives in a given node and the typical performance of one of those upstream actives. Thus making the standard assumptions including, everything else being equal, and monitoring at a common point, the SNR should reduce as the total number of actives are reduced. However, noise funneling due to the cable plant itself has been found to be only a small contributor to total SNR performance.

The downstream benefits of reduced RF cascade discussed in [2] are applicable in the upstream as well. Reductions in ingress, common path distortion, CPD, interference, impulse noise, linear and nonlinear distortion can all be expected in the upstream. Overall, less opportunity exists for corrosion, poor connectivity, water seepage, etc. because of less coax, components, and connectors in the upstream path.

Shorter cascades should also have less variation in RF levels. There is typically no automatic gain control in upstream RF amplifiers and there can be significant gain changes across a long cascade. Cutting the cascade in half would reduce the gain variation of the return plant due to causes such as seasonal change of temperature.

Distributed Feedback (DFB) laser or digital return (DR) upgrades from older generation Fabry-Perot (FP) lasers may have been a necessity for some cable operators wishing to deploy 64-QAM DOCSIS signaling in the upstream spectrum. A comparison of laser technologies has been documented in [3] et al. A 5 dB improvement in optical link SNR could be realized with upgrading a FP laser with a DFB, thus providing 64-QAM with adequate margin to operate successfully.

It is reasonable to suggest that the previously discussed evolutionary developments could combine to result in an appreciable improvement in upstream HFC performance. Whether or not this improvement in upstream HFC performance could support more efficient, yet sensitive, modulation schemes will be explored in more detail in the following sections of this paper.

Upstream Efficiency

Multiple digital communication references, including [4], discuss the Shannon-Hartley capacity theorem. The following equation from [4] et al. describes the system capacity, *C*, of a channel impaired by Additive White Gaussian Noise (AWGN) and is a function of *SNR* and channel bandwidth, *W*.

Equation 1

$$C = W \log_2\left(1 + \frac{S}{N}\right)$$

Part of the chart from [4], which illustrates Equation 1, has been included as Figure 1. Figure 1 illustrates the modulation method efficiency for multiple QAM scenarios. The bit-error-rate, BER, is 1E-8 for QAM scenarios shown in the figure. The dark blue curve represents Equation 1 or the normalized channel capacity over a range of SNR values. The purpose of this figure is to show improvement in efficiency via the use of 128-QAM and 256-QAM relative to lower-order modulation schemes.

64-QAM offers an efficiency of 6 bits/s/Hz, which translates to approximately 30.72 Mbps for a 6.4 MHz channel with a BER = 1E-8. The theoretical capacity of a channel with the same characteristics is 59.54 Mbps, based upon Equation 1. Table 1 compares efficiency and capacity of 64, 128, and 256-QAM, based upon a 6.4 MHz channel with a BER = 1E-8.

Table 1 - Modulation Method Efficiency and Capacity

M-QAM	Efficiency (bits/s/Hz)	Data Rate (Mbps)	Theory (Mbps)
64	6	30.72	59.54
128	7	35.84	65.91
256	8	40.96	72.29

The increased efficiency of 256-QAM represents a 33% improvement over 64-QAM or approximately 10 Mbps more throughput for a 6.4 MHz channel. The increased efficiency of 128-QAM represents a 17% improvement over 64-QAM or approximately 5 Mbps more throughput for a 6.4 MHz channel.

Shannon-Hartley capacity theorem, from Equation 1, provides useful insight into the limits of today's HFC networks. However, the formula is truly much worse than what has been presented thus far because it was intended more for estimating the entire channel capacity rather than the capacity limits of an arbitrarily divided subset. Therefore, the Table 2 illustrates the capacity associated with a 37 MHz upstream bandwidth.

Table 2 – Theoretical Upstream Capacity vs. SNR

SNR (dB)	Capacity (Mbps)
28	344.236
31	381.067
34	417.919
40	481.651

Compare the results of Table 2 above to a practical system capacity, specifically a channel bonding scenario using 6, 6.4 MHz carriers in Table 3. Note, not all modulation levels represented in Table 3 are part of the DOCSIS specification. This isn't an apples-to-apples comparison because 38.4 MHz of bonded channels should result in an even higher capacity, per Equation 1, than what has been previously illustrated using 37 MHz.

Table 3 - Practical Upstream Capacity vs. SNR

M-QAM	SNR (dB)	Capacity (Mbps)
64	28	184.32
128	31	215.04
256	34	245.76
256	40	245.76

When the Shannon limit was first pushed at V.34 and V.90 limits, the maximum SNR of 34 dB was assumed and in reality 36 dB was the upper limit. V.34 and V.90 being the standards supporting 33.6 kbps and 56 kbps rates associated with applications including dialup data service. The usable bandwidth was 200 Hz to 3,700 Hz or 3,500 Hz. The symbol rate was 3,429 sym/s with an alpha = 0.08. The theoretical capacity was 40.695 kbps and the capacity of V.34 maximum was 33.6 kbps.

V.34 had attained 82.56% of the theoretical limit while DOCSIS 3.0 using 6 bonded channels attains only 58.8% of the Shannon-Hartley limit. It's clear that significant opportunity to improve efficiency for HFC networks still exists.

Fidelity Requirements

The increased efficiency unfortunately comes at the expense of higher fidelity requirements. The following equation from [4] et al. describes the BER for a rectangular constellation, impaired by AWGN. Both a matched filter reception and Gray encoding are assumed.

Equation 2

$$P_B \approx \frac{2\left(1-L^{-1}\right)}{\log_2 L} Q\left[\sqrt{\left(\frac{3\log_2 L}{L^2-1}\right)\frac{2E_b}{N_0}}\right]$$

Q(x) is the complementary error function and L represents the number of amplitude levels in one dimension. Using Equation 2, waterfall curves for 64, 128, and 256-QAM have been illustrated in Figure 2. These waterfall curves illustrate the required SNR necessary to support a given BER. The results presented in Figure 2 assume no forward error correction, FEC, gain.

The waterfall curves show that an additional 3 dB SNR over 64-QAM is required to support equivalent BER performance at 128-QAM. Additionally, 6 dB SNR over 64-QAM is required to support equivalent BER performance at 256-QAM. Specifically, in order to support BER = 1E-8 the following SNR requirements must be met for a communication channel dominated by AWGN.

64-QAM SNR = 28 dB
128-QAM SNR = 31 dB
256-QAM SNR = 34 dB

Signal-to-interference levels for both 64 and 256-QAM have previously been documented in [5] and [6]. 256-QAM was shown to be approximately 12 dB more sensitive to narrowband interference than 64-QAM. The sensitivity was consistent regardless of whether the interfering tone was at the center frequency are at a location consistent with a CTB beat (-1.75 MHz). Additional variation in sensitivity was documented when CTB was generated using a live video. Given a delta of 12 dB in sensitivity between 64 and 256-QAM, it is reasonable to expect that 128-QAM could be 6 dB more sensitive to narrowband interference than 64-QAM.

DOCSIS upstream equalization had been documented to be approximately 2 dB less effective on average for 64-QAM than 16-QAM in [7]. DOCSIS upstream equalization is actually comprised to two distinct parts, transmit pre-equalization which is defined in the DOCSIS specifications, and post-equalization in the cable modem termination system, CMTS. Both processes are driven by estimations made in the post-equalization function of the CMTS receiver. It was simulation of post-equalization that revealed the decreased effectiveness of the equalizer to correct for single dominant micro-reflections of varying delay and maximum amplitude characteristics when comparing modulation levels.

The interaction of the post-equalization process was confirmed in laboratory measurements. However, the magnitude of single dominant micro-reflections being corrected was appreciably higher than what had been assumed by DOCSIS to be present in the HFC environment. For example, [7] showed that single dominant micro-reflections, with a delay characteristic of one symbol period, were corrected at levels approximately 5 dB higher than DOCSIS assumption of 10 dBc, with similar characteristics. Only 6.4 MHz signals were evaluated. The micro-reflection delay of one symbol period is the inverse of the symbol rate or approximately 195 ns. Micro-reflections with such short delay and high amplitude characteristics may be encountered more frequently within the customer premise, because of multiple factors including short lengths of coaxial cable and loss characteristics associated with drop plant.

Given a delta of 2 dB in sensitivity between 16 and 64-QAM documented in [7], it is reasonable to expect that 2 dB degradation in equalization performance when comparing 256-QAM to 64-QAM under equivalent

conditions. However, equalization should still be robust, thus correcting for single dominant micro-reflections greater than what has been assumed by DOCSIS.

Documented phase noise requirements from [5] show how 35 dB signal-to-phase noise ratio or less is required to assure small degradation (1-2 dB) of the BER curve of 64-QAM. Similarly, 41 dB signal-to-phase noise ratio would be required for 256-QAM. Thus resulting in a reasonable expectation for 128-QAM being half the difference, or 38 dB signal-to-phase noise ratio.

Substantial research exists to aid in the discovery of fidelity requirements for both 128 and 256-QAM use for the upstream HFC. This information is useful in focusing laboratory investigation and validation of necessary fidelity requirements.

HFC Evolutionary Considerations

DOCSIS 3.0 has provided cable operators with the option of extending upstream bandwidth to 85 MHz. This upstream expansion may create a greater range of useable center frequencies for 128 and 256-QAM. However, optical links could have an additional loss of 3-4 dB SNR due to sharing optical link dynamic range with at least twice as much upstream bandwidth.

Introduction of enhanced hybrid technology, such as GaN, may slow down cascade reductions. It's possible that the improved downstream reach of these RF amplifiers could encourage the continued use of longer RF cascades. It could also just mean improved reach for fiber-to-the-last-active, FTTLA, or node plus zero architectures, N+0.

The combination of the previous two considerations could negate some of the gains possible with previously explored HFC evolutionary changes. The more pertinent goal of this paper is simply to raise awareness of these HFC changes and the potential for performance improvement of the upstream HFC rather than enumerate permutations and performance estimations thereof.

MER, CER PERFORMANCE EVALUATION

A performance evaluation was conducted to measure modulation sensitivity. The three modulation levels measured were 64, 128, and 256-QAM. Each modulation level was subjected to varying impairment levels that resulted in both a 0.5%, and 1% codeword error rate, CER. In other words, the data recorded reflects impairment levels that resulted in approximately 0.5%, and 1% combined CER (corrected codeword error rate plus uncorrected codeword error rate). MER was also recorded for each data point.

Test Topology

The test topology, shown in Figure 3, was designed to simulate an HFC network comprised of a FP or DFB optical link and an N+6 RF cascade. The combined SNR performance, which includes contributions from the DOCSIS link and HFC, using an FP optical link was equal to 31.5 dB. The combined SNR performance using a DFB optical link was equal to 33 dB. Multiple vector signal generators were used to produce the impairment permutations, which were also measured using vector signal analyzer.

The CMTS was configured such that; (1) DOCSIS transmit pre-equalization was enabled, (2) ingress cancellation was enabled, (3) channel width was 6.4 MHz, (4) center frequency was 25.2 MHz, (5) modulation profiles supported were 64, 128, and 256-QAM, (6) modulation profiles disabled byte

interleaving, (7) modulation profiles supported FEC = 219, T=16.

The CMs were configured such that they were very large packets (4,000-byte) to simulate a heavy usage condition which would result in maximum exposure of codewords to each of the impairment conditions.

For each impairment permutation, the DOCSIS links were allowed time to settle into a steady state, giving the adaptive processes ample time to converge on an estimate of the communication channel impairments. During steady state, FEC and MER statistics were recorded. This process was repeated until the targeted 0.5% and 1% CER were measured. Recordings were made of CER, MER, and impairment contributions.

Impairment Library

A mid-band frequency of 25.2 MHz was chosen because the authors assumed that cable operators would primarily be interested in increasing modulation levels on channels with a known history of reliable 64-QAM performance. Ingress and noise were the most relevant impairments given the above assumption. Below is a list of impairment characteristics. Each value of AWGN was first measured as a baseline, and then combined with only one static ingress case for each impairment permutation.

AWGN

- SNR = 33 dB
- SNR = 31.5 dB

Static Ingress

- Single QPSK modulated carrier, f_c = -1.5 MHz offset, rate = 10 ksym/s, bandwidth = 12 kHz
- Single FSK (2-level) modulated carrier, f_c = -1.5 MHz offset, rate = 320 ksym/s, bandwidth = 400 kHz
- Single FM modulated carrier, f_c = -1.5 MHz offset, rate = 400 Hz, deviation = 20 kHz, waveform = sinusoid
- Three modulated carriers simulating CPD
    - Two outer Global System for Mobile, GSM, carriers at f_c = ±1.5 MHz offset, MSK modulation, rate = 270.833 ksym/s, 0.3 Gaussian
    - One inner π/4 Differential QPSK modulated carrier, 384 ksym/s, alpha = 0.5

31.5 to 33 dB AWGN represents an error free range of operation for 64-QAM. These SNR values translate to BER = 6.5E-17 to BER = 0 respectively. SNR margin ranges from +3.5 to +5 dB, based upon the 28 dB needed to support BER = 1E-8. This margin should make it easy for other digital signal processes, DSP, like ingress cancellation and equalization to function without issue.

For 128-QAM, the same SNR values translate to BER = 2E-9 to BER = 1.7E-12 or an SNR margin range of +0.5 to +2 dB respectively. This represents a comfortable region of operation, which likely introduces small variations into the adaptive DSP systems.

256-QAM appears to be in uncomfortable range with SNR values translating to BER = 1.1E-5 to BER = 3E-7. This represents negative margin relative to BER = 1E-8, specifically -2.5 to -1 dB. It is expected that some of the FEC margin will be consumed in this range. Additionally, this level of noise is expected to introduce appreciable variation into adaptive DSP systems.

A set of single static ingress was selected to make some initial assessment of ingress cancellation performance relative to

ingress most likely expected to show up in mid-band frequencies. QPSK modulated ingress was chosen to represent ingress with appreciable amplitude modulation component. The FSK modulated ingress was chosen to represent a 2nd harmonic component associated with a set-top box carrier. FM modulated ingress was chosen to represent shortwave radio from fire, police, and/or public safety systems. CPD was modeled after samples retrieved from the field. GSM and π/4 DQPSK carriers were selected because their spectral characteristics closely matched that of the CPD field samples.

The goal of establishing this impairment library was capture some reference points for discussion and develop a process of evaluating higher-order modulation performance suitability in the upstream HFC.

Laboratory Measurements

Tables have been included at the end of this paper that tabulate performance for 256, 128, and 64-QAM subject to impairments described in the previous section. Each table represents a modulation level. The left-hand side of each chart represents the combined performance including DOCSIS link, and HFC using DFB return optics. The right-hand side of each chart represents the combined performance including DOCSIS link, and HFC using FP return optics. Level represents the level measured on the vector signal analyzer, which is the same level input to the CMTS receiver. UNCORR% represents the uncorrected codeword error rate, which is the percentage of uncorrected codewords out of the total codewords received in each measurement. Total codewords is the sum of corrected, uncorrected and unerrored codewords. CORR% represents the corrected codeword error rate, which is the percentage of corrected codewords out of the total codewords received in each measurement. The first row of each chart represents the baseline case with AWGN impairment. Subsequent rows identify the type of ingress and the target CER. CER = 0.5% targets were measured prior to CER = 1.0% targets.

In Table 4, note that the -1 to -2.5 dB margin range for 256-QAM with no other impairments is already creating countable codeword errors. In fact, a baseline starting with 0.880% corrected codeword errors would easily exceed 1% threshold if any additional impairment to the network with the FP return optics were to be added. Also note that the delta between 128-QAM and 256-QAM is well beyond the predicted minimum 6 dB based on previous work [5] and [6]. The likely cause for this variation in performance is the baseline BER performance, which is already consuming FEC margin as well as introducing noise variation into vital DSP functions such as ingress cancellation. This range of 256-QAM operation represents a challenging environment for ingress cancellation success.

In Table 5, note that the +2 to +0.5 dB margin range for 128-QAM with no other impairments has no codeword errors. Ingress cancellation performance has degraded with increased sensitivity and decreased margin compared to 64-QAM. FM and QPSK ingress is only 5 dB lower for 128-QAM compared to 64-QAM. This suggests that ingress cancellation is capable of overcoming increased sensitivity associated with increased modulation complexity, provided there is adequate SNR margin.

In Table 6, note that the +5 to +3.5 dB margin range for 64-QAM with no other impairments has no codeword errors. There is negligible difference between ingress cancellation performance at +3.5 to +5 dB margin range. The noise performance

appears to be more than adequate at this range.

It's clear that ingress cancellation performance is affected by bandwidth and modulation characteristics of ingress. The ingress canceller corrected for FM and QPSK-type ingress far more effectively than any other ingress evaluated. FSK and CPD-type ingress represented the most challenging ingress conditions, which suggest increased sensitivity to bandwidth. Considering 128-QAM, the disparity between dBc levels of FM and FSK is appreciably higher than the other modulation levels. With the reduced margin, the ingress canceller had more trouble with the wider bandwidth ingress (320 kHz FSK) than with the narrower ingress (20 kHz FM).

CONCLUSIONS

Various HFC plant improvements may create opportunity for increased modulation efficiency in the upstream. This paper has described some of the critical requirements associated with supporting higher than 64-QAM modulation levels.

Based on the measured data presented in this paper, 128-QAM, with its 5 Mbps throughput improvement over 64-QAM, is well suited to be the next step in modulation level increase. It seems reasonable that ingress cancellation performance comparable to 128-QAM could be achieved with 256-QAM, provided similar SNR margin, specifically SNR = 34.6 dB to SNR = 36 dB.

Future work in this area could more fully develop and explore specific applications leveraging the use of higher-order modulations in upstream HFC. Development for applications such as Cellular backhaul or local public school video applications could drive further refinement of relevant requirements. In these two applications, packet size is much larger (>1000 bytes) than that typically encountered on "normal" internet and Voice over IP, VoIP, traffic situations (<384 bytes). Because the packet size is larger, modulation profiles could take advantage of byte interleaving and reap its benefits to increase FEC performance and counter higher impulse environments.

ACKNOWLEDGEMENTS

The authors wish to thank Michael Aviles and James Weineck for their valuable contributions to this paper.

REFERENCES

[1] Ulm, John, *Leveraging S-CDMA for Cost Efficient Upstream Capacity,* SCTE ET, Washington DC, April 2009

[2] Howald, Rob, *Fueling the Coaxial Last Mile,* SCTE ET, Washington DC, April 2009

[3] Rathod, Vipul, *Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts,* SCTE Cable-Tec Expo, Denver CO, October 2009

[4] Sklar, Bernard, *Digital Communications: Fundamentals and Applications,* Prentice Hall, Inc., Upper Saddle River NJ, 2001

[5] Howald, Rob, *QAM Bulks Up Once Again: Modulation to the Power of Ten,* SCTE Conference Proceedings, June 2002

[6] Stoneback, Dean, *Distortion Beat Characterization and the Impact on QAM BER Performance,* NCTA Show, Chicago IL, June 1999

[7] Thompson, Rob, *Optimizing Upstream Throughput via Equalization Coefficient Analysis,* NCTA Cable Show, Washington DC, April 2009

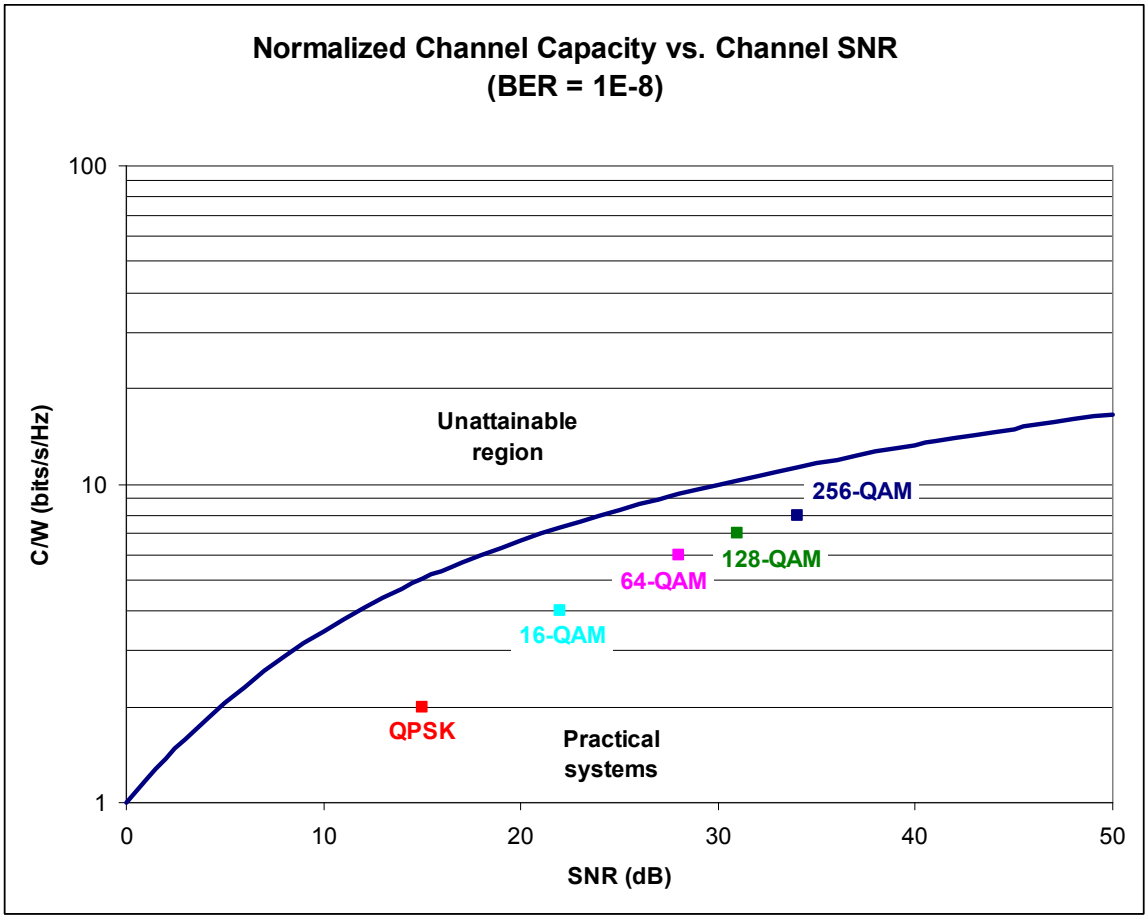


Figure 1 - Modulation Method Efficiency at BER = 1E-8

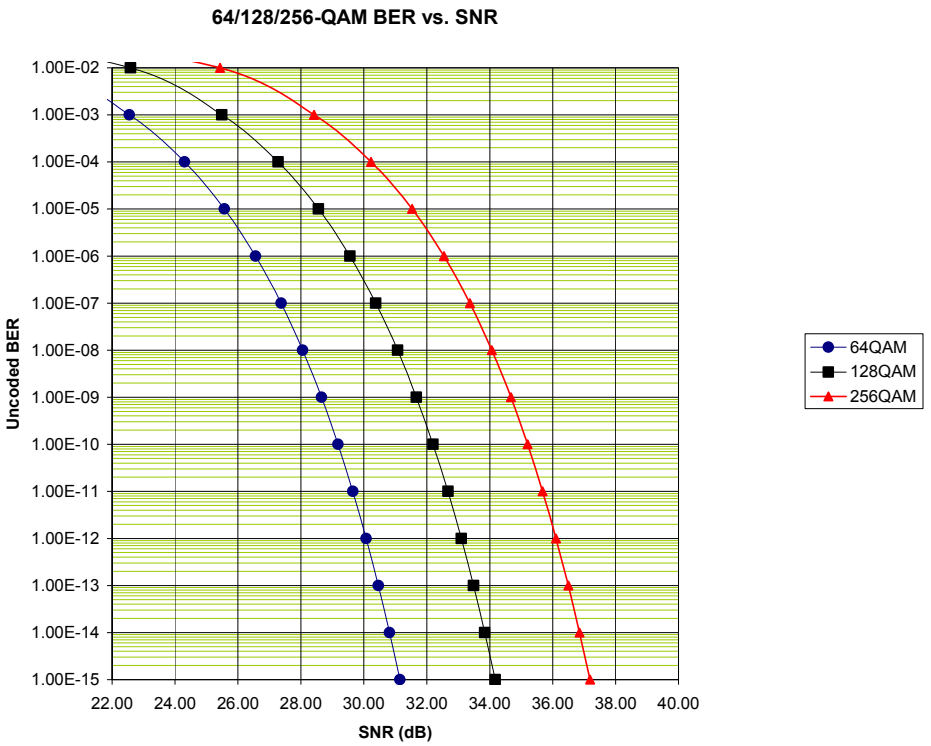


Figure 2 - Modulation Method BER

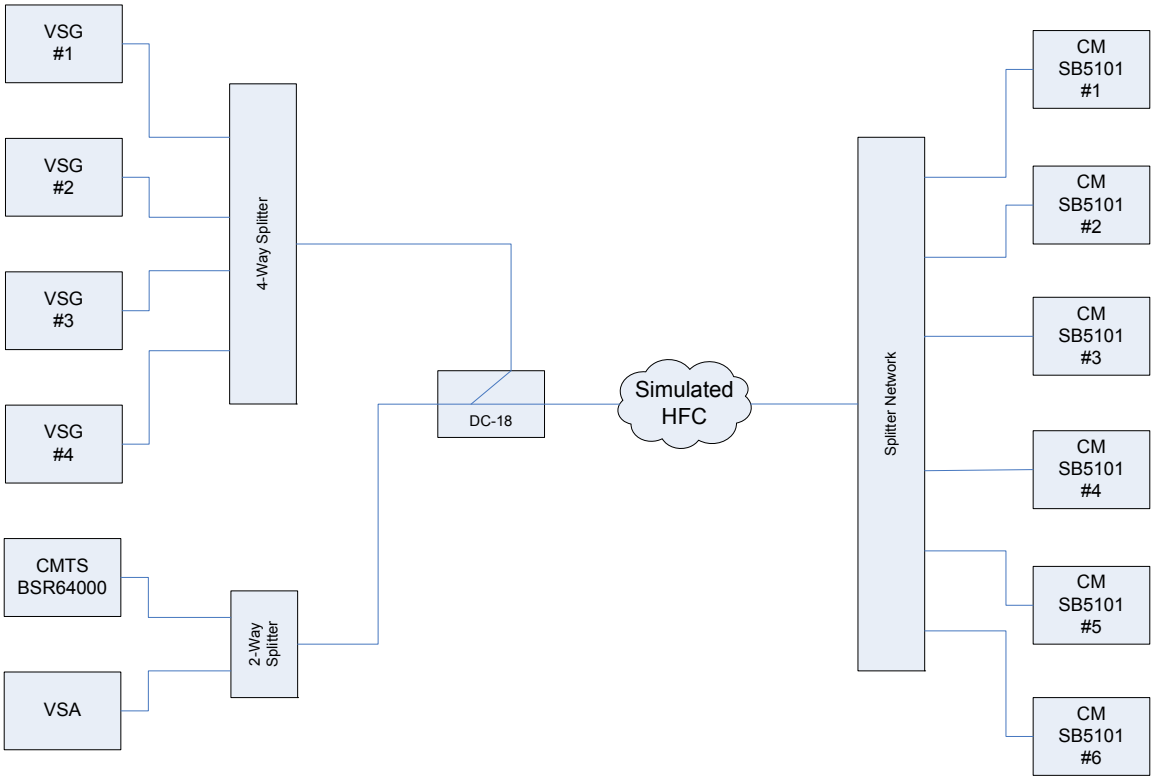


Figure 3 - 64, 128, 256-QAM Sensitivity Test Topology

Table 4 - 256-QAM Performance

	256-QAM							
	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)
Baseline - AWGN	33	0.000%	0.174%	34.20	31.5	0.000%	0.880%	33.30
Single Ingressor Case								
QPSK 12 kHz 0.5%	23.3	0.000%	0.671%	33.60				
QPSK 12 kHz 1.0%	21.4	0.002%	0.911%	33.50				
FSK 320 kHz 0.5%	34.15	0.000%	0.533%	34.00				
FSK 320 kHz 1.0%	21.15	0.018%	1.185%	33.50				
FM 20 kHz 0.5%	27.8	0.000%	0.625%	33.80				
FM 20 kHz 1.0%	22.2	0.000%	0.911%	33.80				
Three Ingressor Case								
CPD 0.5%	37.9	0.000%	0.713%	33.60				
CPD 1.0%	36.8	0.000%	1.034%	33.30				

Table 5 - 128-QAM Performance

	128-QAM							
	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)
Baseline - AWGN	33	0.000%	0.000%	34.20	31.5	0.000%	0.000%	33.30
Single Ingressor Case								
QPSK 12 kHz 0.5%	-1.6	0.014%	0.432%	31.20	0.5	0.004%	0.307%	31.40
QPSK 12 kHz 1.0%	-2.4	0.063%	1.495%	30.90	-0.7	0.009%	0.522%	31.30
FSK 320 kHz 0.5%	16.7	0.058%	0.543%	31.20	17.7	0.013%	0.411%	31.30
FSK 320 kHz 1.0%	15.7	0.072%	0.968%	30.80	15.7	0.053%	1.267%	30.30
FM 20 kHz 0.5%	-0.9	0.119%	0.305%	32.30	0.3	0.125%	0.315%	31.50
FM 20 kHz 1.0%	-2.3	0.331%	0.436%	32.20	-1.0	0.280%	0.449%	31.30
Three Ingressor Case								
CPD 0.5%	24.5	0.172%	0.273%	31.00	26.3	0.071%	0.452%	30.70
CPD 1.0%	22.5	0.575%	0.476%	30.40	25.4	0.214%	0.606%	30.40

Table 6 - 64-QAM Performance

	64-QAM							
	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)	Level (dB, dBc)	UNCORR %	CORR %	MER (dB)
Baseline - AWGN	33	0.000%	0.000%	34.20	31.5	0.000%	0.000%	33.30
Single Ingressor Case								
QPSK 12 kHz 0.5%	-6.4	0.104%	0.312%	28.70	-5.7	0.124%	0.502%	28.40
QPSK 12 kHz 1.0%	-7.5	0.279%	1.090%	27.60	-7.5	0.528%	1.581%	27.60
FSK 320 kHz 0.5%	-3.8	0.029%	0.244%	27.60	-3.8	0.065%	0.347%	27.60
FSK 320 kHz 1.0%	-4.8	0.311%	1.025%	27.00	-4.8	0.329%	1.433%	26.90
FM 20 kHz 0.5%	-4.7	0.229%	0.117%	30.40	-5.5	0.254%	0.152%	28.80
FM 20 kHz 1.0%	-6.3	0.642%	0.246%	30.20	-6.2	0.218%	0.125%	30.20
Three Ingressor Case								
CPD 0.5%	14.6	0.251%	0.341%	27.60	15.6	0.248%	0.340%	27.60
CPD 1.0%	14.1	0.650%	0.784%	27.10	14.6	0.557%	0.719%	27.20

ACCELERATING ADVANCED ADVERTISING: SUPPORTING EBIF WITH CLOUD-BASED SOLUTIONS

Jeremy Edmonds
ActiveVideo Networks

Abstract

The cable industry faces severe challenges in the race to enable advanced, interactive advertising. Cable-led efforts like Enhanced Binary Interchange Format (EBIF), tru2way and Canoe all offer compelling solutions for advertisers, but cable's fragmented legacy infrastructure, particularly in terms of customer premise equipment (CPE), is preventing these standards from fully satisfying advertisers' current needs for targeted, interactive and dynamic advertising at scale.

To generate significant advanced advertising revenue today, cable providers must embrace tools and technologies that provide advertisers with opportunities to engage viewers that are similar to those that exist on the Web. This paper will examine the cable industry's current advanced advertising challenges and provide information that can help the industry deploy cloud-based solutions that leverage Web technologies and standards.

INTRODUCTION

Advertisers are demanding the ability to deliver Web-style television advertising that's targeted, interactive and dynamic. They want to pinpoint receptive audiences with the right messages, engage them through the point of purchase, and measure viewer activity at every point in the process through new levels of technology development. The cable industry is responding to these needs, but, despite its best efforts, it's running into significant technical challenges.

At best, the cable industry is leaving advanced advertising revenue on the table; at worst, it is losing that revenue to competitors, which are using their support for Web standards to gain the early high ground in advanced video advertising.

Cable's strengths, including its unparalleled subscriber footprint and its superior video delivery infrastructure, remain noteworthy and compelling for advertisers. However, a near-term need exists for cable to muscle more quickly into the nascent advanced advertising area with a platform that both leverages the capabilities of and supports a migration to EBIF.

Solutions are now available that enable operators and advertisers to leverage existing Web platforms such as DoubleClick to support advanced ad delivery. Through the use of a Web-based platform and cloud-based transcoding of Web content to MPEG, cable can quickly gain market share in the advanced advertising space. In addition to offering advertisers access to cable's subscriber base, the best programming and unsurpassed video quality, the operators gain access to an existing advertising ecosystem and would be able to draw on countless advertisers and agencies that would be familiar from Day One with the tools necessary to develop and manage advanced advertising content.

This paper discusses existing challenges to mass cable operator rollouts of advanced advertising, and reviews ways in which operators can capitalize on existing Web

tools and technologies to support their advertising efforts.

ADVANCED ADVERTISING: OPPORTUNITIES AND RISKS FOR CABLE

Traditionally, television advertising has been a passive endeavor for the consumer. Ads, typically 30 seconds in length, are broadcast to all viewers of certain channels or programs, with the hope that one or more target audiences are viewing them. Any calls to action associated with the ad require the user to get up and do something: visit a store, call a phone number, log on to a Web site. However, with the advent of DVR, VOD, sites like Hulu and other products and technologies, many of these linear ads are now being viewed by only a fraction of the audiences they once reached.

As a result, advertisers are now seeking to establish more compelling and personal connections with their target audiences. Today's consumers clearly demand increased choice and control, even over advertising. Younger consumers, in particular, want a more Web-like TV viewing experience.

Advanced television advertising basically mirrors Web advertising, in that it enables the user to participate in much more active and even impulsive activity. These focused and targeted ads are more integrated with the individual's user experience, with the ability to engage the viewer all the way through the "purchase funnel," from introduction to a product to the point of purchase. For example, the viewer can click on the remote for more information, access a microsite devoted to the product or service, talk about the product or service through social networking functionality, and even make a purchase, all from the comfort of the living room couch. Advertisers desire these kinds of interactive ads because, like those on the Web, they can provide interested viewers with additional information, measure viewer activity at multiple points in the process, and deliver those measurements to the advertiser.

Such advanced ads present significant revenue generation possibilities for cable systems operators, as well. Operators uniquely have a broad base of subscribers that can be targeted geographically, demographically or by interest; they provide a video environment that is far more stable and of higher quality than the Web; and they have a broad range of content that is delivered directly to the television, which remains the dominant viewing device in the home.

But although personal, interactive experiences are the order of the day, and would seem to be a good fit for cable, the industry faces a significant challenge in terms of the difficulty of most existing cable set-top boxes to meet this demand. Advertising dollars are already starting to shift to the Web, where targeting and interactivity are more easily achieved. The rise of Internet-connected televisions presents advertisers and CE manufacturers with another chance to deliver "over-the-top" content and advertising to consumers. If cable doesn't respond to the need for interactive advertising, other parties are well-positioned to grab those dollars.

THE CURRENT CABLE ENVIRONMENT

Advertisers want advanced advertising to be part of the normal viewing experience, rather than an "interactive TV" application. There is precedent for this. When an "interactive application" gains consumer traction, it exits the perceived realm of

"interactivity" and becomes part of the "normal" viewing experience. Examples of such interactive applications pioneered by cable include the electronic program guide (EPG) and video on demand (VOD). Both have ceased to be considered "interactive TV" applications, and have passed into the realm of the "normal" viewing experience.

However, applications like EPG and VOD, as "normal" as they are, are still "destination-based." To access VOD, for example, consumers must "go to" a place to find and order titles. The same is true for EPGs, which exist as separate menu destinations. These applications are separate from, not seamlessly integrated with, the normal viewing experience.

Immersive interactive video applications such as advanced advertising strive to bring the desired content "to the viewer," not make the viewer search to find a "destination" in an unnatural way. This "delivering the content to the viewer" (versus "destination-based interactivity") can be found on many video streaming Web sites, such as YouTube, where the activity of viewing any given video stream is augmented by metadata links to several other video assets (as well as non-video applets).

The cable industry, to its credit, is fully aware of the challenges advertisers face with the current subscriber environment. It has created a number of new standards and initiatives designed to provide advertisers the interactivity and targeting they desire, notably EBIF, tru2way and Canoe.

The EBIF specification was created by CableLabs to deploy interactive applications over a two-way video plant to all existing and new digital set-top boxes. The cable industry is working to deploy EBIF nationwide, but that goal has not yet been reached.

With tru2way, application developers can create customized interactive services that can be deployed seamlessly to millions of cable customers. It offers "write once, run everywhere" Java-based programming capability to developers. However, the programs will only run on set-top boxes and other devices that support tru2way. Deployment of such devices is minimal at this time.

Canoe Ventures, backed by prominent operators, is working with CableLabs to develop EBIF templates that advertisers can use to deliver interactive ads to major operators around the country. Canoe also offers backend services that will allow for campaign management and reporting across operators. This effort, however, is still in development.

These activities are promising for the future, and show that cable is working to offer advertisers a national, ubiquitous platform for advanced ad development and provisioning. But although cable continues to devote considerable energy to them, these standards can't effectively deliver advanced advertising today.

Significantly, none of these standards currently can reach all of today's deployed cable set-top boxes. This is due to cable's fragmented infrastructure, characterized by its various models of customer premise equipment and head-ends from multiple manufacturers, and of different ages and capabilities. This infrastructure has grown organically over time among the nation's operators, and has served it well. But this diversity is hampering cable's efforts for a national advanced ad platform.

Let's take a closer look at the issues that are impeding national cable rollouts of advanced ad campaigns.

The Workflow Issue

Cable's ability to provide advanced advertising is hampered by a workflow impediment. It lacks an automated systems infrastructure to connect the "sales order process" with the "creative process" to the "content management and provisioning process" and finally to the "delivery process."

For the traditional multichannel video subscription business, this workflow is well established. In its simplest form, movies and TV shows are produced, licensed to an aggregator (e.g., NBCU), wholesaled to an operator (e.g., Comcast) for distribution, and then retailed to the consumer. The advertising and subscription models are well established for this process.

The important point is that there are automated systems (encoding, content protection, "billing systems," trafficking systems, royalty payment and settlement) that support this model so these businesses can scale.

With respect to interactive applications, this "workflow" does not currently exist in any uniform, scalable way. The current ecosystem of extant and desired interactive video applications and services relies on a patchwork of business systems and creative tools, all of which are delivered to a heterogeneous population of operators with no "billable event tracking" except by sneaker-net and swivel chair operations. Without the "back-end" tied to the "front-end" via an automated workflow that generates invoices, tracks payments and respects copyrights, it will be very difficult to build a scalable business around a popular interactive application.

A specific example of the workflow conundrum is the notion of the "bound" application, which executes synchronously with a program or advertisement.

Current cable solutions to this problem are still to come. EBIF's strength is its overall potential reach, which could be the entire installed base of digital cable set-tops. The EBIF specification defines the client execution engine and the data formats for sending applications to the client. Such definition is critical and necessary, but for EBIF-based "bound" applications to become mainstream, a necessary scaffolding of workflow must emerge. The purpose of Canoe is to create ubiquitous end-to-end workflow for advertisers, and to shield advertisers from the need to deal with multiple cable operators and their separate workflows. But Canoe implementation remains on the horizon.

Advertisers, however, require workflow scaffolding today. They require a known, easy and repeatable method for creating advertising applications, and for applying quality assurance mechanisms to ensure those applications behave at their best on all set-top boxes. They need data collection to fulfill the application's intent, and to feed any primary or third-party billing mechanisms.

Consider an advanced advertising application that allows the viewer to click on a widget associated with an ad to receive more information on the product. From a workflow perspective, the following requirements are critical:

1) *Creative*: What should the widget look like? Who builds creative for

the campaign, and to which template, and using which authoring tool(s)?

2) *Application provisioning*: Operationally, the interactive application must be provisioned on to the network. Its widget assets must be transferred for playout, and its availability parameters must be fed into the traffic/billing system.
3) *Stewardship*: All ad campaigns follow general and specific rule sets, such as: competing products may not be shown within the same ad pod; time parameters to protect children from inappropriate content, etc.
4) *Data Collection*: After playout, data associated with the spot needs a method to flow into the aggregation engines feeding national and local campaigns.
5) *Billing*: Any additional revenue associated with the interactive spot needs a feed into operator billing systems.
6) *Reporting and Settlement*: Automated mechanisms must be available to operators and advertising constituencies, etc., to create reports both for advertising effectiveness and contract fulfillment purposes.

While efforts such as EBIF and Canoe are underway, the author does not know of any available solutions that will connect the traditional day-to-day business of advertising sales to the operators' broadcast and unicast streaming platforms. Individually and combined, workflow gaps prevent the business from scaling and impede the ability for multichannel video providers to build both local and national advertising revenues.

The Challenge of the Installed Base

Digital cable set-tops, as a category, are beyond their 15th anniversary. Until fairly recently, they've existed as "thin clients" that lag behind the Moore's Law trend of computing devices. Compared to PCs, digital set-tops have long been dismissed as devices lacking sufficient processing power and memory to enable immersive, media-rich applications. In short, what's thick today is thin tomorrow, and, for digital cable boxes, it's always tomorrow.

However, it is not in the best interest of operators to deploy new set-top boxes every year. There is a good reason why there is so much legacy customer premise equipment in the field. It takes a lot of time and money to replace a set-top box. All of those "legacy" boxes in fact have value. They save operators money on truck rolls to deploy new equipment, and they save customers the time of waiting for those trucks to show up with new equipment. As a revenue generating unit for "the bedroom" or other non-living-room locations in a household, it is very hard to justify replacing them system-wide with more advanced boxes.

Because of today's accelerated advances in technology, it would be impractical for operators to roll out new boxes with great frequency. Even the latest and greatest set-top box becomes a "legacy" device within months. Consider that when you roll a new car off a dealer's lot, it immediately becomes a used car. That doesn't make the car any less useful. It would be impractical to "upgrade" a car every six months. The same holds true for set-top boxes.

The installed base of digital set-top boxes presents a "lowest common denominator" problem for application development and software version control. Building applications only for high-end boxes reduces potential reach; building applications for all set-top variations reduces the application's attractiveness to the lowest common

denominator of graphics chips, processing power and memory.

These issues of potential reach and the attractiveness of applications are of paramount importance for advertisers, as seen in Figure 1.

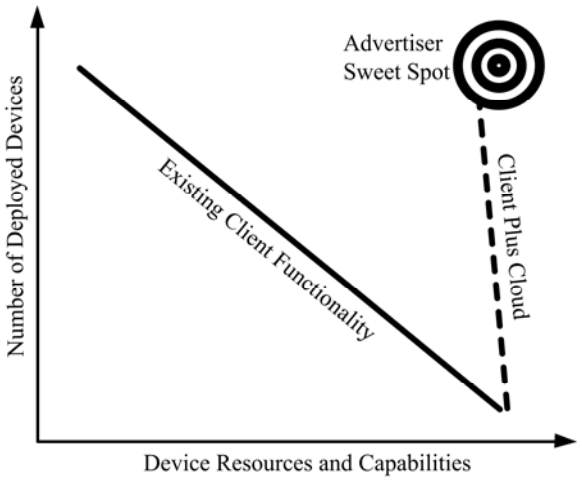


Figure 1 – Functionality Compared to Deployed Boxes

From the perspective of the advertiser, a very small footprint of accessible devices is very hard to target or monetize. With more scale and reach, the advertiser has a greater likelihood of effectively reaching its target audience. At the same time, the better the advertisement looks and functions, the higher the impact the ad will have. Given the deployed base of cable set-top boxes, these two options are mutually exclusive; the more features and functionality a box can offer, the fewer of them there are in the field. Adding cloud rendering and processing to the equation minimizes the device-centric resources from the equation. This is shown with the dotted line. Put another way, operating interactive applications solely upon the limited capabilities of the aggregate set-top base, and without the benefit of network server resources, means the wealth of capabilities in the newest units is eclipsed by the care-and-feeding needs of the oldest units.

The Challenge for the Application Developer

Advanced advertising application developers face prohibitive time and cost outlays in the cable environment. The problem is not unique to advertising applications; developing any software for cable, such as EPG software, requires a substantial round of testing and compliance to ensure that the software works with all of the set-top boxes in deployment.

EBIF development presents a similar challenge. The result is the likely pruning of platform features to the lowest common denominator.

The greatest expense associated with investment in client software technologies often is in targeted development, integration and regression testing across dozens of different CPE platforms—each with its own performance characteristics, graphics display capabilities and consequent impact on the viewer experience. While tru2way, EBIF and other standard client software platforms are certainly a great improvement, they do not solve the pervasive cross-platform compatibility issue.

This is not an issue just for the cable world; even Web browsers have differing capabilities across Macs and PCs, as well as across different OS installations on those devices. It is necessary when developing a robust Web site to test it against all significant browsers in use.

In cable, however, a new release of the GuideWorks-based EPG and VOD menu software package can take between one and two years for development, testing and certification before it is made available to

operators for deployment across a family of set-top boxes.

EBIF is a write-once, run-anywhere platform that provides cable operators with an efficient, lightweight, well-managed set-top environment. It offers interactive applications, graphical overlays and instantaneous responses via an immensely deployed base of existing cable set-top boxes. However, while it opens the door for new kinds of interactivity, EBIF by design was not intended to offer more capabilities than the most basic set-top boxes could handle. It is essentially a generic system that provides some basic capabilities for interactive TV on legacy set-top boxes. In addition, EBIF is not yet ubiquitously deployed.

Simply put, it will be prohibitively expensive and time-consuming to deliver the levels of interactivity and functionality that advertisers require into the wide variety of digital set-top boxes currently in the field. While EBIF, tru2way and Canoe are worthwhile endeavors for cable, there is a need for a server-based, or cloud-based, solution to enhance those advanced advertising efforts. The idea is to push as much programming, application logic and processing into the network cloud as possible, and communicate with digital set-top boxes through simple MPEG streams, for example.

THE CLOUD-BASED ADVANCED ADVERTISING SOLUTION

The language of the computing "cloud" is typically associated with the Internet, even though the term itself pre-dates the Internet by at least two decades, when computer scientists recognized the need to share processing workload over clustered computers. Cloud computing has become a staple of the Web world, with applications such as Web-based email, YouTube and countless others removing the burden from client devices and leveraging the power (and storage capabilities) of the cloud.

Now that we live in a fully digital television world, the cloud concept applies just as much to TV, as we're essentially dealing with data and nothing more. The cloud TV concept can remove workflow gaps and bridge the application and media processing requirements between head-ends and set-tops. A cloud approach also enables developers of advanced advertising applications to use familiar, Web-based development tools that are similar to or the same as those used to provide interactivity within a Web site.

Server-side functionality, such as large-scale data manipulations (for example, deep keyword searches on hundreds of thousands or millions of records), recommendation engines and ad decision engines, already exist for Web applications. These technologies can be applied as-is to the cable infrastructure. Familiar client authoring functionality, such as DHTML, JavaScript, CSS, Ajax and JSON, can be used in existing deployments by moving much or all of the "client" processing into the cloud.

EBIF at the Core of Advanced Advertising

EBIF is specifically designed to enhance broadcast video with prompts ("call to action" graphics) that entice the user to engage with what they are watching in new ways. This ability to embed interactivity in layers is key to the advanced advertising initiative. The base layer is the broadcast advertisement itself: If a user does not have EBIF capabilities, this is all that is seen. The second layer is the "call-to-action" graphic, which the User Agent blends on top of the broadcast video advertisement according to

the instructions in the bound application, as exemplified in Figure 1.

Figure 1 – Call to Action Overlay

Once the user has responded to the call-to-action, there are various possible next steps for the advertiser. A common option at the core of the Canoe initiative is the RFI (Request For Information). Leveraging the customer's personal information on file with the cable company, a brochure, coupon or other information is mailed to the subscriber's physical address.

Enhancing this basic RFI functionality with cloud-based services can offer tremendous increased impact and opportunities for engagement to the advertiser. Given a simple construct such as cookies, an advertisement can offer different screens or other options if the consumer has already clicked on the RFI in the past, for example. Using simple asynchronous client-server communications (akin to AJAX on the Web), an EBIF application can leverage server-side database repositories to display a tremendous amount of information in small chunks, without overtaxing the available client memory.

Adding MPEG to EBIF

Existing EBIF User Agents have been designed to support simple telescoping functionality as seen in Figure 3. An EBIF application can switch away from a broadcast stream to a unicast stream based on triggers typically caused by the user. The

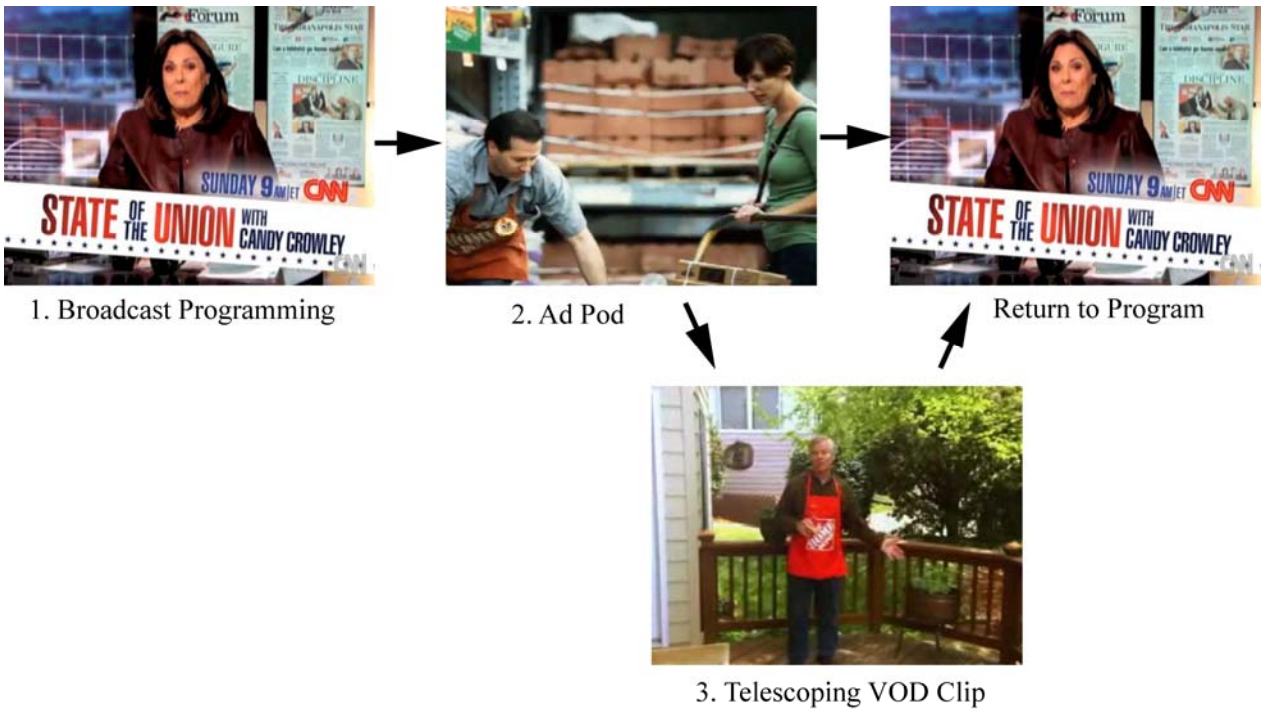


Figure 3 - Telescoping Ad

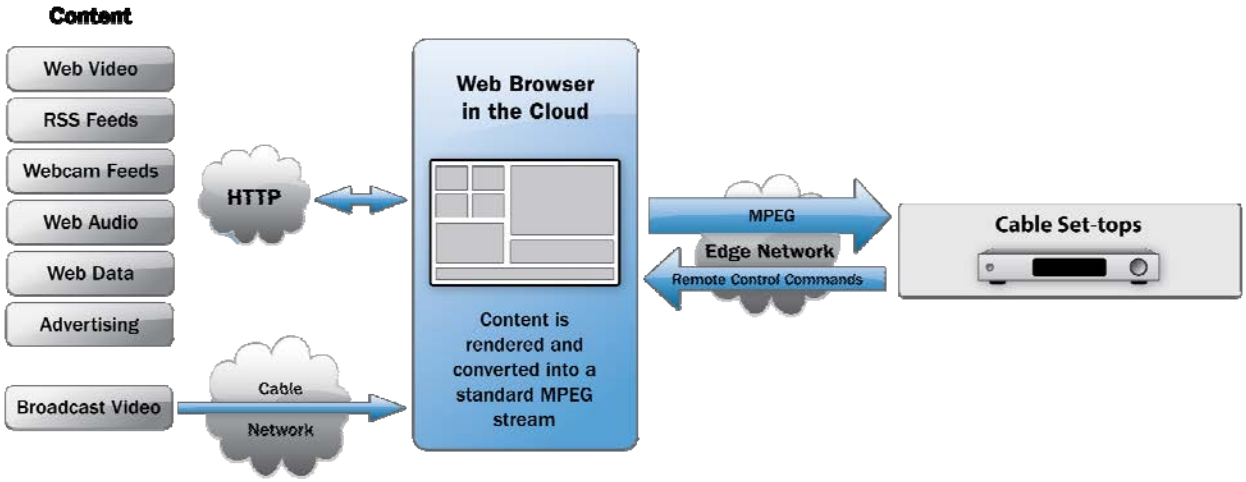


Figure 4 – Web Browser in the Cloud

EBIF application persists during this stream switch, providing a seamless application lifecycle and allowing the EBIF application to send commands and control messages to the server that is generating the unicast stream. In its simplest form, this allows for an advertisement to play a longer video for users who express interest via a call to action on-screen during a broadcast advertisement.

Additional Benefits of a Remote Browser

Using enhanced streaming servers in the network cloud, this basic telescoping functionality allows for advanced graphics and video capabilities to be added to any EBIF application. As seen in Figure 4, the enhanced streaming server is running what is essentially a specialized Web browser in the cloud.

Once there is a framework that allows basic Web browser functionality across all set-top boxes, the door is open for leveraging tools and expertise that have been developed and polished over the last decade or more. Cable companies using client-side browsers as well as streaming-server browsers are already capable of leveraging some very powerful tools. It has been shown that DoubleClick, Atlas and other ad campaign management suites are already able to manage, track and report on interactive ad campaigns running on all deployed cable set-top boxes. It has also been shown that Omniture and similarly powerful tracking and reporting solutions are able to be used in advertising campaigns across existing cable deployments. All of these examples are using existing Web solutions without modification, and with standard commercial agreements in place.

Brand Appearance in EBIF

Another immediate benefit to using this specialized streaming server is that it is capable of rendering images in the entire MPEG-2 color space; more than 16 million colors are available at once, with minor limitations due to chrominance sharing between adjacent pixels. Compare this to EBIF graphics allowed in the most widely supported baseline specification (including the Motorola DCT-2000), which allows 16 colors on screen at a given time, and the Cisco (Scientific-Atlanta) SA-2000, which allows 256 colors on screen at a time. The immediate benefit to an advertiser that is protective of its brand appearance is clear, and because these cloud-based graphics are not dependent on the client functionality, the same brand image appears identically on a

Motorola DCT-2000, SA-2000 and all other deployed set-top boxes.

The Power of Video in Advertising

Beyond simple brand appearance, there is remarkable power and clear benefit to embedding rich video into an ad. It is abundantly clear why so much advertising money is spent on video when compared to static (newspaper, Web banner ads) and audio (radio): The response through brand awareness and feedback is measurable. To an advertiser spending money on television, the expectation is that there will be a large video-based component to its advertisement. When a standard advertisement is enhanced with EBIF, some simple graphics and text are available. Adding telescoping allows for a single long-form video to play on demand. Adding cloud-based streaming suddenly makes a broad range of video and video-centric effects available to the advertiser, for example: multiple video windows playing at once; or allowing the original broadcast video to play in a thumbnail in the corner while the telescoping video plays in a larger window. Many more are possible. Because the rendering of the user interface takes place in the cloud, any existing set-top box can provide this type of interface and brandable real-estate, all leveraging EBIF as the gateway and flexible servers placed in the cloud.

CONCLUSION

A cloud-based approach offers cable operators greater flexibility and implementation of advanced advertising now, the ability to grab market share against competition as quickly as possible, and a solid foundation for the future. It also combines the mass audience of all two-way cable boxes with the features and functionality that advertisers need to meet their brand and marketing requirements.

The cloud approach leverages EBIF's capabilities and supports migration to EBIF, while addressing today's urgent problems head on. It leverages cable's strengths of an unrivaled subscriber footprint and superior video delivery. By using existing Web standards to support advanced advertising delivery, with the ability to transcode Web content to MPEG, it allows cable to lock in immediately to the existing pre-roll ecosystem for advanced advertising, as advertisers and agencies are already intimately familiar with Web content development and management tools. Ideally, a mix of cloud and client-server approaches is optimal.

ACCESS NETWORK BUILD COMPARISONS: FTTH, HFC FIBER DEEP, AND LTE

Tim Burke - Liberty Global
Michael Eagles - UPC Broadband

Abstract

The service characteristics and technology evolution of fiber-to-the-home (FTTH), HFC fiber deep, and 4th Generation Wireless (LTE) will define the next generation of access network and broadband competition. We argue that it is from these developing technologies and delivery platforms that broadband customers will choose the manner in which they receive their future broadband services.

In comparing the alternatives we consider several questions. What will the broadband competition for each alternative look like from a consumer perspective? What factors or trends might influence the outcome? Which access technology can best be optimized for future broadband service? Can the alternatives co-exist? What level of capital would justify the expected services?

The purpose of this paper is to provide a competitive, technology and economic framework for comparing next generation broadband access alternatives from both a greenfield and upgrade basis.

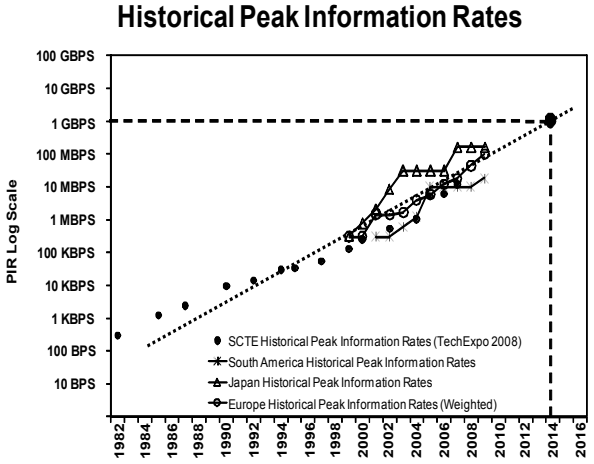
BROADBAND COMPETITION

Broadband competition may be examined along several attributes including: Speed, Price, and Service Types; with all broadband competitors facing what could be called a "capital threshold".

Faster Broadband Speeds

Broadband competition between access technologies is also highlighted by pressure to offer faster peak advertised speeds or Peak Information Rates (PIR). Considering 28 years of historical trends, and subject to regional variations in competition, it is not inconceivable that Peak Information Rates of 1Gbps could be required by 2014!

Figure 1: Peak Information Rates Continue To Grow [1]

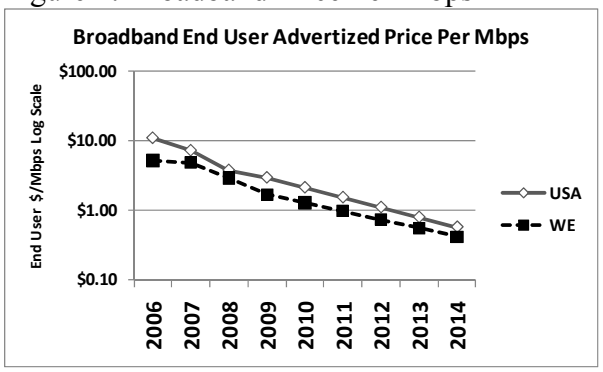


If we assume that, based on historical trends, end users will continue to decide between competing offerings based on advertised speed or peak information rates, this has an impact on the access network technology choices an operator needs to consider in order to remain competitive.

Price and Competition

Broadband prices per Mbps continue to decline over time for fixed line broadband. The declining price per Mbps is a function of increasing information rates and competition. For example, in markets where broadband access competition is particularly intense, or there are irrational competitors, price per Mbps per month declines can be more dramatic. We illustrate with an example of U.S. broadband prices per Mbps below in Figure 2.

Figure 2: Broadband Price Per Mbps [2]



Similar Service Types

With Telco's addressing the need for Video and very high speed data services by moving from Digital Subscriber Line (DSL) to Gigabit Passive Optical Network (GPON), we contend that for fixed line broadband the core "service types" that are offered will be similar to Cable.

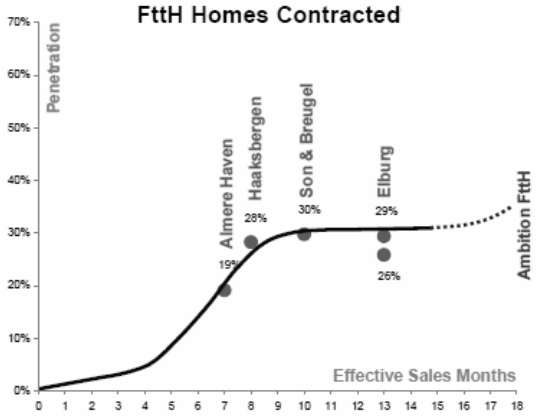
Mobile Broadband or Long Term Evolution (LTE), on the other hand can support Voice and Data "service types" that are also offered by fixed line operators, with the unique attribute of mobility, but will struggle with mainstream Video services[3].

The Customer Base Potential

In a highly competitive environment, we believe it is unlikely that a service provider could achieve more than 50% penetration of addressable homes on average (i.e. the "customer base potential[4]"). Looking at industry examples we see that a typical penetration of addressable homes would be around 30%. For example, noting a 3 year horizon, Verizon stated in Q1 2007 that, "By 2010, Verizon expects to have a 35-40% penetration rate of FiOS Internet customers, and a 20 to 25% penetration rate of FiOS TV customers"[5]. In July 2009, it was reported that Verizon had achieved sales penetration of 28.1 percent for FiOS Internet and 24.6 percent for FiOS TV[6].

Similarly, worldwide other Telcos have stated in that their FTTH pilot results, offering triple play packages of Broadband, TV and Telephony, have met expectations with up to 30% of FTTH homes

Figure 3: Example of FTTH Sales Curve [7]



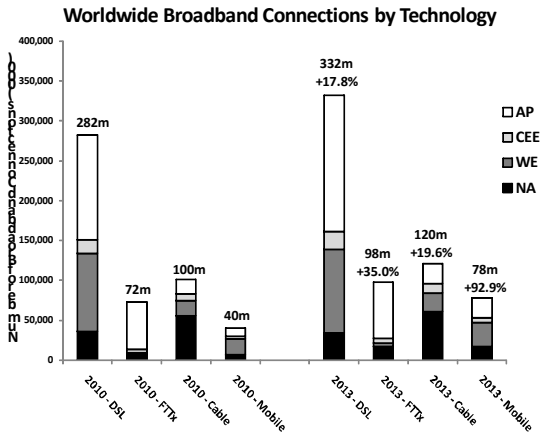
Today, Telcos recognize the effect competition has on penetration potential. In some countries Telcos have stated targets to have one third or 33% of the population connected by FTTH by 2015[8].

Operators that own both Fixed and Mobile Broadband operations, such as AT&T, are also looking at a hybrid model where the alternative access technologies are complimentary rather than competitive by splitting the service types by technology. For example DSL/FTTH could be used for Video and Fixed Data, while LTE is used for Mobile Data and Voice[9]. Using a service bundle, this enables the Telco to maximize the penetration of both technologies by minimizing service type competition between the access alternatives.

Future Competition

Today, Cable and DSL are the most widely deployed broadband technologies worldwide, with Mobile Broadband emerging as a rapidly growing segment. Regional variations are evident, with Cable broadband dominating in North America, and DSL dominating in Asia-Pacific and Europe.

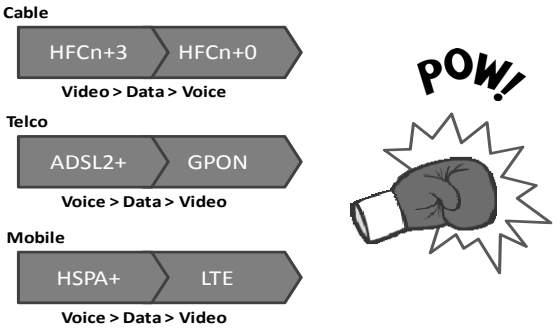
Figure 4: Worldwide Broadband Connections by Technology 2010 vs. 2013 [10]



The points for future competition in the broadband market place are clear: (a) Between future fixed line technologies there are unlikely to be any major "service type" advantages to attract subscribers away from their existing access provider, (b) Without a service type advantage, FTTH/GPON adoption will be driven by Greenfield or future Telco upsell of the existing DSL subscriber base in order to counter higher Peak Information Rates from competing alternatives, and; (c) Mobile broadband offers two of the three service types in the market, which could be expected to place additional pressure on fixed line penetration potential.

If we assume that tomorrow's Fiber access deployments are primarily a Telco competitive response to the limitations of existing DSL technology with its comparatively low data rates[11], rather than a new category of broadband service provider; then three categories of next generation access technology emerge: Cable's HFC Fiber Deep, Telco's Gigabit Passive Optical Networks (GPON), and Mobile Broadband (LTE), as show in Figure 5 below.

Figure 5: Future Broadband Access Competition



TECHNOLOGY

The Telcos and Cablecos use land or terrestrial based technologies via the medium of fiber, coaxial cable or copper cable. Mobile operators utilize the spectrum or airwaves they own or lease to provide these same broadband services. As the offered speeds for broadband access continually increase from 1Mb/s to 100 Mb/s and beyond, the technologies deployed by fixed line and mobile companies must evolve.

The terrestrial based companies will continue to bring the highest capacity and most efficient medium, fiber optics, closer and closer to the customer. Cablecos do this via Hybrid Fiber Coax (HFC) deployments that bring fiber deeper into the network (beyond current Fiber Node locations) so that no regenerators are required beyond the FN. Telcos deploy fiber to Remote Terminals (RTs) or cabinets that contain DSL electronics, called DSLAMS, located in neighborhoods. Because Telcos have much stricter bandwidth (and capacity) limits inherent in the copper plant versus the Cablecos coax, many Telcos have even begun to pull fiber all the way to the home.

Likewise, the wireless operators will need to add more spectrum, make more efficient use of their radio technology and move cell sites closer in, towards their customers' homes.

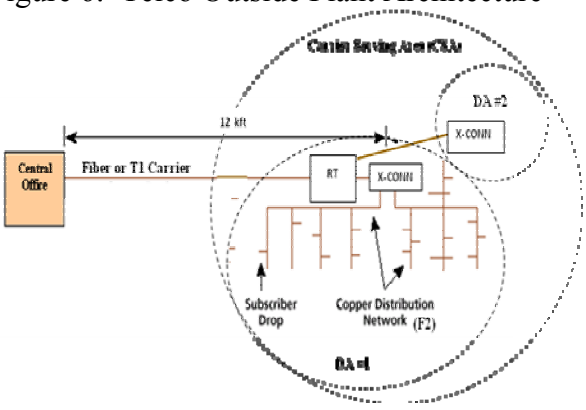
The incremental economics associated with these evolutionary moves are the key to how quickly technology change-outs will occur. The economic challenge for fixed operators has always been the cost/benefit of reusing the existing medium (copper pairs or coax) in the last mile (or ½ mile) to the customer premises versus undergoing the substantial costs and time to rewire the local loop with fiber optics. Similarly, mobile operators need to spend additional capital to build more towers closer to customer locations in order for mobile devices to receive the sufficient signal strength required for high speed services inside homes.

Finally, it makes economic sense to share network elements across as large a group of customers as possible. Contrary to this economic need, network resources are being shared across smaller and smaller groups of customers as the average speed offered to the end users increases and customer penetration levels rise.

Telephone Company Networks

In the U.S. the Telco architecture is quite varied as the number of homes served by a Central Office (C.O.) can range from less than 1,000 to over 50,000 households. Likewise, the distances from the C.O. to the edge of the C.O. area (called wire center) vary from 10,000 to 20,000 feet. Figure 6 provides a visual representation of the Telcos outside plant architecture.

Figure 6: Telco Outside Plant Architecture



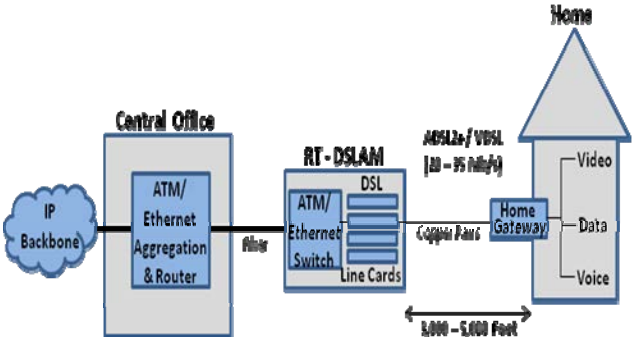
As Figure 6 shows, the wire center district can be broken up into many smaller CSAs (Carrier Serving Areas) that pass from 600 to 2,000 homes, where the furthest home can be easily 12,000 feet from the CO. CSAs are made up of smaller geographic neighborhoods called DAs (Distribution Areas) serving 250 to 300 homes. DAs contain cross-connect points called Feeder Distribution Interfaces (FDI) or Serving Area Interfaces (SAI) where the furthest home is usually between 3,000 to 5,000 feet from the FDI. These cross-connect cabinets terminate the twisted copper pairs that originate in each home and are called distribution pairs. FDI cabinets typically terminate 2 to 3 lines per home, so large cross connects may be required. Historically, these cross connect cabinets were fed by copper coming all the way from the CO where half as many feeder pairs from the CO matched up against the distribution pairs going to the homes. Over time, digital T1s and fiber optics replaced the copper feeder and fiber optic electronics were placed in Remote Terminal cabinets (RTs) right next to the cross-connects

.

Fiber to the Node using DSL Technology

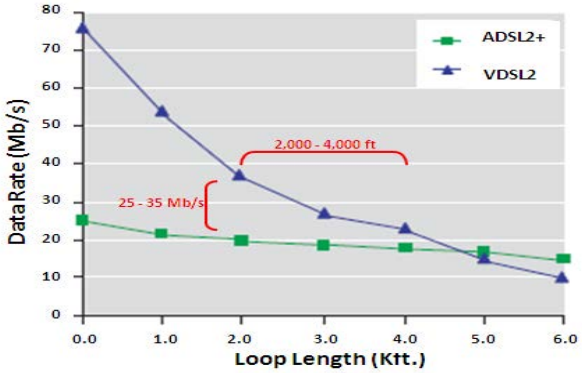
As ADSL and VDSL technology was deployed, remote DSLAMS that contain the DSL line cards, ethernet or ATM switches and fiber optic transmission equipment were placed in these RTs. Figure 7 shows the current ADSL2+ or VDSL architecture used by AT&T (uVerse) and other Telcos.

Figure 7: DSL Network Architecture



The main variants of DSL technology all take advantage of using an increasing amount of the usable spectrum available on twisted pairs of copper wires. VDSL enhancements increased the spectral band plan to 12 MHz from the 2.2 MHz limit of ADSL2+. Consequently, the obtainable speeds increased as long as the quality of the copper plant was very good and the distances from the line cards in the remote DSLAM to the home were less than 4,000 feet. It is important to note that twisted pair copper wires contain a number of impairments such as crosstalk, noise and bridge taps that severely reduce data speeds, even when the distances are short. Because ADSL2+ and VDSL technologies are so sensitive to distances, charts showing data rates versus loop lengths of the copper plant are useful. Figure 9 is a good example of a rate versus reach graph for ADSL and VDSL technologies. Given the Telcos' Distribution Area (DA) architecture, the key range is 2,000 to 4,000 feet which corresponds to maximum speeds of 25 to 35 Mb/s.

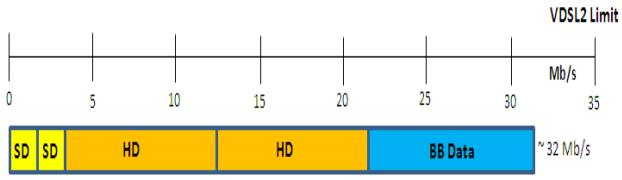
Figure 9: ADSL VDSL Rate vs Reach[12]



When one looks at an overlay of the bandwidths required for triple play services, it becomes apparent that VDSL technologies quickly become obsolete as end user demands increase. Because of the shorter loop lengths in many European countries VDSL speeds are higher and technology lifetimes will be extended.

For instance, Figure 10 shows a household requiring two High Definition (HD) and two Standard Definition (SD) video streams along with 10 Mb/s of broadband data that max's the VDSL bandwidth limits even when MPEG4 SD and HD compression technology are assumed (2 Mb/s and 9 Mb/s).

Figure 10: Triple Play Customer Requirements and VDSL2 Capacity Limits



Fiber to the Home (FTTH)

In the near future most TelCo's will realize they have to deploy fiber directly to the customer premises to meet the growing customer broadband demands as the capacities of twisted pair copper are limited with DSL technology. Additionally, the operational expense of managing many individual copper strands and cross-connect points in the outside plant will continue to be an economic burden.

As video demands are added into the broadband end user speed requirements the Telco decision to extend the fiber to the home will become even more urgent. Some Telcos with a longer financial payback view, such as Verizon, have already reached this conclusion and made FTTH (branded FiOS) a cornerstone of their broadband and video services. Telcos without the financial strength and longer term view have opted to avoid the large FTTH investment by making the copper last longer using ADSL2+, VDSL, satellite and digital terrestrial (DTT) access means for video and broadband services. This is the strategy of AT&T in the U.S. and many European Telcos.

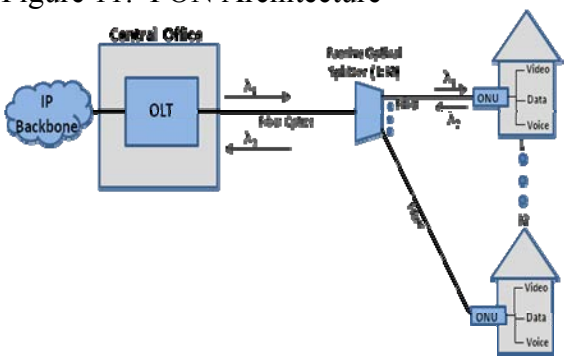
Passive Optical Networks (PONs)

The highest capacity and most economical way of providing FTTH is to deploy PON technology. Although PON architecture requires a complete change-out of the current

Telco architecture, at a huge capital expense, it does solve the capacity and economic constraints inherent with upgrading and re-using the Telco copper plant.

A Passive Optical Network is an end-to-end optical network using a point-to-multipoint architecture containing no active elements at any location in the outside plant. It is an extremely efficient way of providing high capacity broadband services, as the only active (or powered) components are in the CO and at the customer premises. Additionally, the economic benefit of sharing resources is possible as a single fiber optic strand is shared across multiple homes (32 or 64) via the utilization of a fiber optic splitter. It is also possible to configure two tiers of splitters in the network where a 1:4 splitter is followed by a 1:8 splitter closer to the served homes. Figure 11 gives us an illustration of the typical PON architecture.

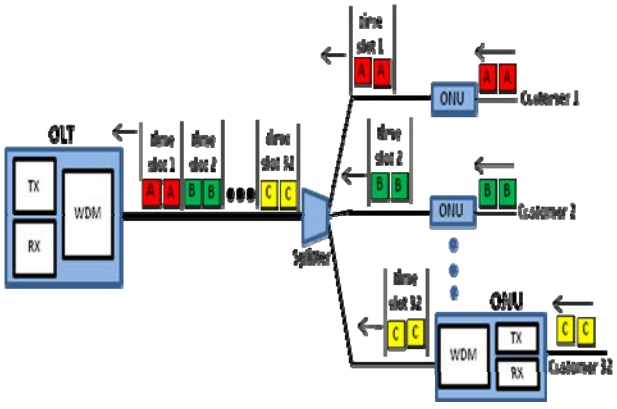
Figure 11: PON Architecture



A key network element shown in Figure 11 is the Optical Line Terminal (OLT) located in the Central Office. Since PON architecture is point-to-multipoint (or multicast) in the downstream direction, the OLT transmits the entire PON bandwidth (2.5 Gb/s for GPON technology) to the PON splitter and all 32 homes receive the packets broadcast by the OLT. The Optical Network Units (ONU) shown in Figure 11 selectively extract the packets from the entire line rate that pertain to the address of the particular customer's ONU. The proper encryption and security mechanisms are implemented in the downstream direction to eliminate eavesdropping and theft of services. Typically, a single optical fiber is used to serve each group of customers connected to a PON splitter, where different wavelengths are associated with the downstream and upstream data flows. Figure 11 designates the different optical wavelengths as λ_1 and λ_2.

The upstream transmission in PON architectures is much more complicated than the downstream. There must be a separation of the information coming from each of the 32 customer ONU's going back to the OLT, as they are all sharing the total upstream bandwidth (1.25 Gb/s for GPON technology). PONs use TDMA (Time Division Multiple Access) schemes that allocate each customer's ONU in the group of 32 to a separate timeslot. The upstream PON technology is quite sophisticated, as it is important for the ONU to have burst mode transmitters/lasers that turn on and off very quickly yet operate at the full upstream line rate. Additionally, the OLT contains advanced receiver technology and performs complex centralized controller functions, as it must be highly synchronized with the ONT's so that upstream timeslots are accurately assigned. Figure 12 provides a visual representation of the upstream TDMA transmission process just described.

Figure 12: PON TDMA Upstream Transmission[13]



PON Standards

In the mid 1990s a group of Telcos formed an association called the Full Service Network (FSAN) to create a PON standard. The outcome of that collaborative effort was the APON specification. APON is based on the ATM transmission protocol and is the reason for the APON abbreviation. Very quickly, the FSAN association upgraded the specification to BPON (Broadband PON) to accommodate higher line rates and interfaces with ethernet protocols while retaining its ATM transmission format. BPON became an ITU standard and was the original technology deployed in Verizon's FTTH FiOS initiative. BPON utilizes a 622 Mb/s downstream line rate and 155 Mb/s upstream speed shared across 32 customers using a single 1:32 splitter.

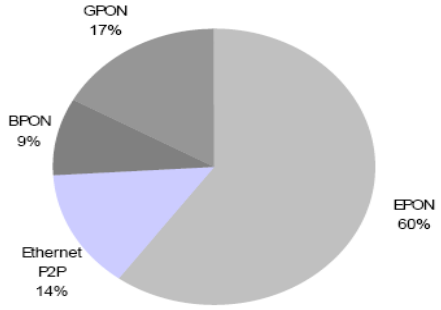
In the early 2000s, Verizon and other Telcos realized that higher line rates and the ability to more easily accommodate Ethernet data traffic was required. As a result of those efforts the GPON (Gigabit PON) standard was born in 2003. GPON was able to accomplish the goals of higher line speeds, more efficient carrying of Ethernet packets and backwards compatibility to ATM and circuit based TDM applications (e.g. - circuit switched voice). Unfortunately, GPON required the creation of a new framing and encapsulation specification within its standard and some very stringent OLT to ONU timing requirements. The result has been greater overall complexity and costs of the network elements. GPON provides 2.5 Gb/s downstream and 1.25 Gb/s upstream line rates and can accommodate either 1 to 32 or 64 split ratios. The typical range from OLT to ONU is 20 Km. and the upstream usually operates at 1310 nm wavelength while the downstream is set in the 1550 nm region. Figure 13 compares the PON standards.

Figure 13: Comparison of PON Standards

PON	Standards Approval	Line Rates		Split Ratios	Typical Range	Enhancement
		Downstream	Upstream			
APON	1995	622 Mb/s	155 Mb/s	1:16 1:32	10 Km	N.A.
BPON	1997	622 Mb/s	155 Mb/s	1:16 1:32	10 Km	N.A.
GPON	2003	2.5 Gb/s	1.25 Gb/s	1:32 1:64	20 Km	10 Gb/s D.S. and 2.5 Gb/s U.S. in 2012
EPON	2004	1.25 Gb/s	1.25 Gb/s	1:32 1:64 1:128	20 Km	10 Gb/s D.S. and 10 Gb/s U.S. in 2011

In parallel to the creation of the GPON standard, an association of equipment manufacturers and Asian Telco operators decided to put together a "pure" Ethernet PON specification that did not make concessions for legacy ATM and circuit based technologies. The resulting specification was ratified by the IEEE in 2003 and became the EPON standard. The key goals of the EPON developers were to combine the simplicity and worldwide economies of Ethernet with the high capacity capabilities of FTTH PONs. As a result of this effort and the worldwide deployments of EPON it has become the most popular PON standard and looks to have increasing market potential going forward. Foremost to its greater potential over GPON is the lower cost of ONUs and higher worldwide volumes primarily driven by Asia deployments. Figure 14 gives us a glimpse of the current PON technology market shares.

Figure 14: FTTH Technology Market Share[14]



The key developmental paths for both GPON and EPON are the increasing line rates. The enhanced EPON standard approved in 2009 will provide 10 Gb/s symmetrical or 10 Gb/s downstream and 2.5 Gb/s upstream speeds.

Commercial chipsets and products will be available in 2011. Additionally, split ratios of 1:128 will be feasible.

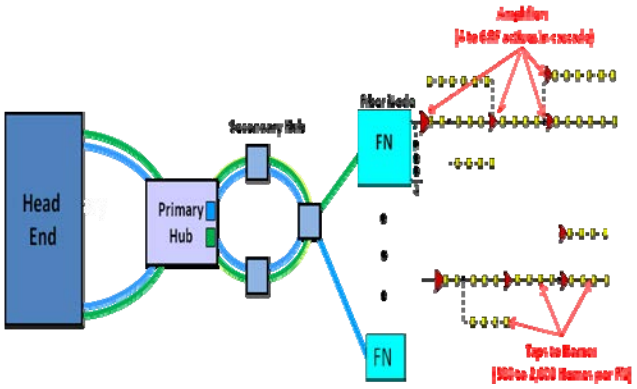
Not to be outdone, GPON will have standards enhancements in 2010 that also provide 10 Gb/s downstream and 2.5 Gb/s upstream line rates. It is expected that commercial products will be available in 2012. At issue for GPON is having sufficient worldwide volumes for chip makers to justify large commercial production levels as North American GPON deployments (e.g. - Verizon) are slowing.

For Cablecos there is an inherent compatibility of ethernet and the IP nature of DOCSIS protocols that makes EPON a stronger future technology choice of MSOs. Because business and commercial customer requirements are typically symmetrical in nature and Ethernet based it is expected that EPON technology for business applications will materialize first for MSOs[16].

Cable TV Company Networks

The other terrestrial based broadband network provider is the Cable TV Company or Multiple Systems Operator (MSO). The MSOs network has evolved in a very advantageous way over the years from both a technology and economic point of view. Figure 15 illustrates the typical two-way Hybrid Fiber Coax (HFC) plant in service today in over 90% of an MSO's footprint.

Figure 15: Modern HFC Network



The Head End (HE) location shown in Figure 15 serves a single or sometimes multiple metropolitan areas covering millions of homes. It contains the video equipment and feeds (satellite and terrestrial) as well as the IP data routers, voice switches, internet and voice network interconnects. Redundantly routed fiber optic transmission equipment is used to transport video, data and voice services from the HE to primary and secondary hub locations. These hubs also serve very large geographic areas of 20,000 to 40,000 homes. In most cases, these hub locations are relatively small unmanned buildings, as they are primarily comprised of optical transmission equipment and Cable Modem Termination Systems (CMTSs).

Over the last ten to fifteen years most MSOs have upgraded their outside plant so that fiber optic strands and equipment is deployed out into the residential neighborhoods. The fibers terminate in small, hardened Fiber Node (FN) cases located either in ducts or on the aerial plant. The FN converts the optical signal to an electrical signal that is transmitted over coax to the household in the FN's neighborhood. As shown in Figure 15, most fiber nodes are designed with four coax distribution segments directed towards the homes in the node. This capability allows for an economical way of adding specific capacity for the various services via service groups. Additionally, this architecture allows MSOs to cleanly segment the fiber node into smaller groups of homes passed without adding new fiber, if demand warrants. A node split or segmentation effectively doubles the available bandwidth per customer by halfing the number of customers served by a fiber node. Typically an FN serves between 500 and 2,000 homes passed (HP) where each coax distribution segment contains between four and six amplifiers (or RF actives) in cascade. Such a configuration is commonly referred to as an N+4 (Node plus 4 amplifiers) or N+6 arrangement. The final network elements to the home are in what is called the

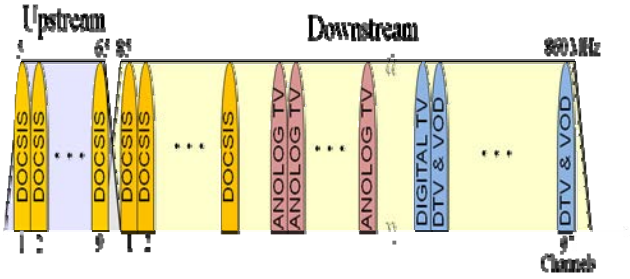
drop portion of the plant and are made up of passive components called taps, splitters and drop cable.

There are inherent advantages to this architecture that lends itself to a graceful and economical evolution: (a) The coaxial cable to the customer premises has a very large capacity that does not limit services or bandwidth in the final mile, (b) The architecture was designed from the beginning with a tree and branch or shared topology that mimics an efficient current day corporate LAN, (c) The plant was designed with a common architectural uniformity so that no matter where you go within an MSOs footprint the structure is similar, (d) A minimal amount of active components are resident in the outside plant as the more intelligent and expensive electronics are either in the hubs or at the customer premises and, (e) Incremental new services (e.g. - video, data and voice) and increasing levels of capacity can be easily added across the existing network elements, so the business scales efficiently.

HFC Customer Bandwidth and Capacity

The capacity of the coax cable portion of the plant has no sharp cutoff.[15] Capacity is limited by the distance from the fiber node to the furthest customer's home and more importantly by the number of amplifiers in series along that particular branch.[16] Additionally, bandwidth is constrained by how much spectrum can pass through the taps and splitter components in the drop segment of the network. Assuming a typical 860 MHz HFC plant common in Europe, Figure 16 pictorially describes the upstream and downstream bandwidth capacities. European HFC networks utilize 8 MHz wide channels (versus 6 MHz in the US) and have 10 more MHz of upstream bandwidth than the U.S.

Figure 16: European 860 MHz Bandwidth



In the upstream direction, a European MSO has a theoretical broadband (DOCSIS) capacity of approximately 270 Mb/s shared across all the homes in the fiber node. This calculation assumes nine usable 6.4 MHz upstream DOCSIS channels operating at a 64QAM modulation (30 Mb/s throughput per channel). Obviously, a clean upstream plant that may have to operate in an SCDMA mode will be required for this capacity scenario. Additionally, substantial capacity gains are possible if operating in a DOCSIS 3.0 mode as the upstream channels are bonded together. Combining 270 Mb/s into one large "pipe" adds a statistical multiplexing gain that is very efficient in the shared LAN type environment of the HFC architecture.

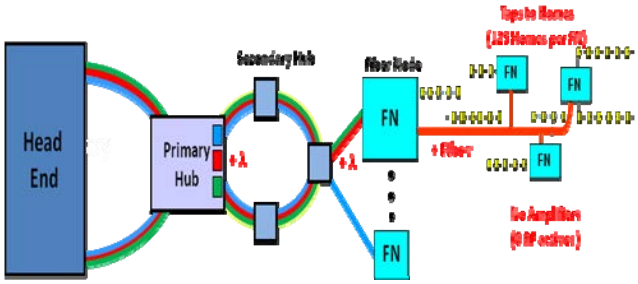
Likewise, downstream bandwidth delivers plenty of capacity in an "all digital" world. Assuming 782 MHz available for downstream traffic in the European scenario of Figure 16, 4.85 Gb/s of capacity is possible. This total assumes 8 MHz channels operating at 256 QAM modulation which provides 50 Mb/s of throughput per DOCSIS channel. Certainly this capacity provides a very future proof capability to support multiple HDTV, VOD, high speed data and voice services. As Figure 16 notes, there need to be allocations for Analog TV requirements and the simulcasting of these signals, so an evolution to the all digital, all IP (including IPTV) world described in the above paragraphs is required.

Hybrid Fiber Coax-Fiber Deep (HFC-FD)

At some point in the life of the HFC architecture, end user demands are so great,

that even in an all digital environment 4.85 Gb/s capacities across 500 homes may not be sufficient. At that point, a further reduction in the proportion of customers vying for the available bandwidth is undertaken by driving fiber optics deeper into the distribution portion of the coax network. This architectural enhancement is illustrated in Figure 17.

Figure 17: HFC Fiber Deep Architecture



Additional optics capacity is added to the hubs and the original serving FN by adding wavelengths to existing fiber pairs (shown as λ in Figure 17). New fiber optic cable is placed in the distribution portion of the plant where formerly the coax and remaining amplifiers were located. In performing this work, the node size is reduced from 500 to 125 HP per FN. The fiber deep scenario provides an added benefit of eliminating the amplifiers and leads to the N+0 terminology, which refers to a node plus zero RF actives. Additionally, having no amplifiers in the HFC plant improves network reliability and operational expenses, as less maintenance support and powering is required.

A critical component to enhancing the HFC plant to a fiber deep architecture is the ability to leverage existing fibers by adding wavelengths to in place fiber. Wave Division Multiplexing is the fiber optic technology that enables multiple wavelengths, each operating at very high line rates, to simultaneously use the same fiber strand. Unique to MSOs, they have deployed the more economical version of WDM, called Coarse WDM. In CWDM systems, the spacing between wavelengths using the same fiber strand is wider (20 nm apart) than other WDM technologies. The large channel spacing was designed to establish a cost effective framework able to accommodate less sophisticated lasers with high spectral width and less stringent temperature and power requirements[17]. This has enabled MSOs to build HFC plants with hardier, smaller, lower power, and consequently more economical, Fiber Nodes.

The higher capacity version of WDM technology is called Dense WDM (DWDM) and allows for very tightly spaced wavelengths (.2 nm apart). Consequently, DWDM systems have extremely high capacity and are usually found in Telcos and long haul transmission systems. MSOs are beginning to deploy DWDM systems where needed in HFC-FD deployments.

As mentioned previously coaxial cable does not have an upper bound at 860 MHz of spectrum. Therefore, when the remaining amplifiers are removed the ability to operate in the GHz frequencies is possible. Fortunately, in many MSOs, passive taps and splitters capable of 1 GHz performance were deployed when the 860 MHz plant upgrades were built. Hence, additional bandwidth can be created from 860 MHz to 1 GHz to be used in either the upstream or downstream direction. The additional 140 MHz of spectrum will create an additional 850 Mb/s of downstream capacity. Therefore, the new capacity allows for 5.7 Gb/s of downstream bandwidth available to the 125 homes in the fiber deep node.

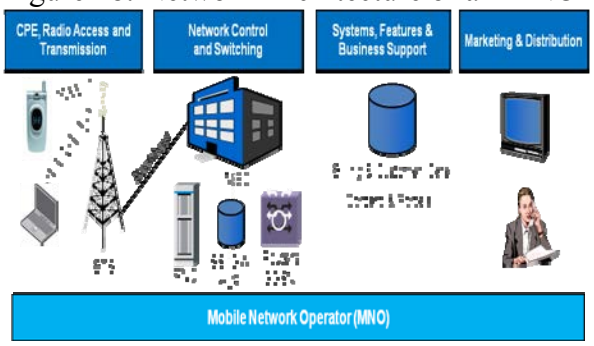
Mobile Network Operators (MNO's)

With the increasing success of MNOs providing mobile based "DSL like" speeds in their broadband offerings, it makes sense for operators to own a mix of wireless and terrestrial based access. Therefore, the existence of standalone wireless or fixed operators, will probably over time, become more and more the exception rather than the rule. Fourth generation (4G) wireless technologies will become the enabler of the

dramatic increase in these end user speeds and mobile network capacities.

Even though the core wireless technologies have evolved over the last twenty years, the overall MNO architecture has remained relatively constant. Figure 18 provides us with a generic layout of a mobile operator's major network elements.

Figure 18: Network Architecture of an MNO

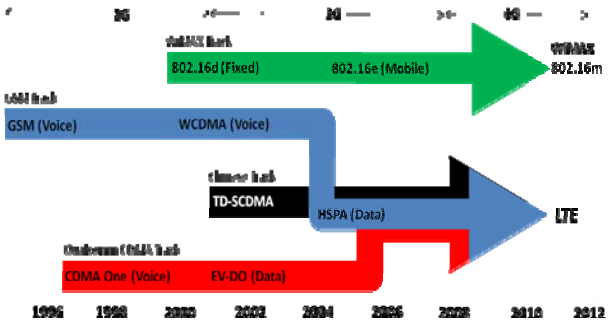


Mobiles provide the end user with wireless connectivity and include traditional voice phones, smart phones and laptops devices. Cell sites are the main infrastructure component and are primarily located on towers and rooftops. The network electronics located at these sites are referred to as Base Transceiver Stations (BTS) and contain antennas, radios and baseband electronics. These elements are both the most expensive and critical portion of the network. As higher speed broadband services are offered, the network component that is gaining an increasingly important role is cell site backhaul. Both microwave and fiber are being used to transport broadband Ethernet back to the main hub location, called the Mobile Switching Center (MSC). The main component in the MSC is the Base Station Controller (BSC) or Radio Network Controller (RNC) that manages the BTSs and the mobility and handover aspects of the network. In 3rd generation systems (UMTS and HSPA), there are various network elements that control and transport the voice and data streams (SGSN & GGSN). Additionally, the voice switch (soft switch and gateways) controls the mobile voice services in a very similar manner as the fixed voice network, the main exception being the role of the Home Location Register (HLR) used to manage subscriber information and roaming mobiles. Fourth generation networks (e.g. - LTE) have simplified the number and complexity of the network elements in the core as they evolve to a flatter, all IP network.

The back office systems shown in Figure 18 have gained increasing importance, as companies, called Mobile Virtual Network Operators (MVNOs), emerge that only own mobile IT systems and marketing functions.

The unrelenting technological progress in wireless has been quite amazing over the last twenty years. Figure 19 provides an illustration of the evolution of wireless standards and technologies since the mid-90s. The progression shows 2nd, 3rd and 4th generation digital wireless standards. First generation mobile technology called Advanced Mobile Phone Service (AMPS) was created in the mid-80s, preceded the standards shown in Figure 19 and was analog based.

Figure 19: Evolution of Wireless Standards

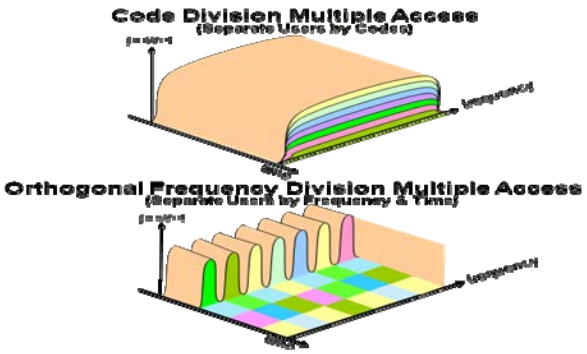


The dominant worldwide standards tract, called GSM, is based on the original digital standard that evolved in Europe. Enhancements have been made to this Time Division Multiple Access (TDMA) based specification to provide increasing levels of capacity and capability. For instance, data service capability was created while continuing to use the TDMA format with GPRS and EDGE technology. A major

upgrade occurred in the early 2000s with the changeover to 3rd generation Code Division Multiple Access (CDMA) based technology called Wideband CDMA (or UMTS). Likewise, the evolution to a 4th Generation technology based on Orthogonal Frequency Division Multiple Access (OFDMA) formats is occurring now with the deployments of LTE. Of particular importance to the evolutionary path is the aspect of backwards compatibility. This means that every new standard keeps the prior standard in place even when major change-outs, such as changing modulation formats, (TDMA to CDMA to OFDMA) occur. For instance, an LTE handset device will have the capability to also operate in the HSPA and GSM mode. Although there have been offshoot technologies over the last fifteen years, such as CDMA One, TD-SCDMA and WiMAX, it appears that the great majority of mobile technology deployments and subscribers are converging to the single LTE standard. Figure 19 shows the other standards as separate evolutionary paths.

The key goal of mobile services is to choose a core technology that uses spectrum efficiently and is also able to effectively separate users (and conversations) within the total spectrum available. First and 2nd generation mobile standards separated voice conversations using frequencies only (AMPS) and both time & frequency (GSM). Figure 20 is a visual representation of 3rd and 4th generation standards (CDMA and OFDMA).

Figure 20: CDMA & OFDMA



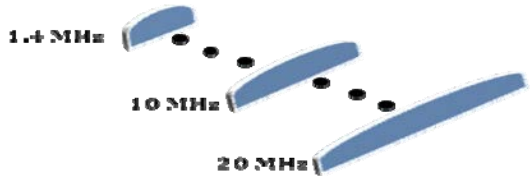
CDMA is quite unique in that it is a spread spectrum technique where every user operates in the same frequencies but conversations are kept separated by the use of unique codes. CDMA operation is best described using the "cocktail party" analogy. Imagine a party held in a small room filled with many couples where each couple speaks only one language which is different from the next pair. Although everyone is speaking at the same time, across the same frequencies, conversations are understandable between a particular couple only. In the same way, unique CDMA codes are like the different languages used by couples in the cocktail party. As in the case of the cocktail party, a key for understandable conversations is the ability to keep the volume in the room low as more and more couples speaking different languages enter the small space. Likewise, CDMA operation requires controlling the power (volume) in the network so noise (adjacent conversations) does not impact the usability and separation of codes.

Although CDMA technology has performed very well over the last 10 years, capacities have begun to reach limits especially for broadband type data services. As a result, the development of 4G technology based on OFDMA technology ensued. Figure 20 illustrates OFDMA's separation of users using the combination of frequencies and time. It differs from 2nd generation TDMA technology in that the frequencies used in OFDMA (shown as peaks in Figure 20) are very tightly spaced and called sub-carriers. The notion of orthogonality is a mathematical way of keeping these close frequencies separate or unique. A conversation between two users would utilize packets appearing across constantly changing frequencies and time slots (shown as the same colors in Figure 20).

Because the number of subcarriers in LTE is variable, LTE allows for a variety of channel bandwidth sizes. This concept is extremely powerful as 3G technologies are restricted to

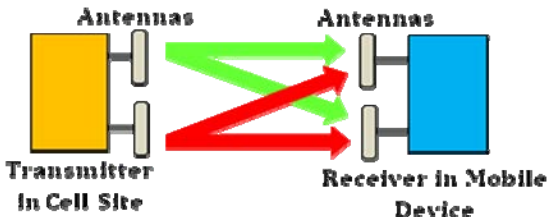
operating at only a 5 MHz channel bandwidth. Figure 21 shows how the LTE specification can operate at various channel bandwidth settings from 1.4 MHz to 20 MHz. A larger channel bandwidth provides the benefit of statistical multiplexing gain.

Figure 21: LTE Channel Bandwidth Options



Another key enabling technology associated with LTE is the concept of Multiple Input Multiple Output (MIMO) antennas. Figure 22 illustrates this concept. MIMO technology provides for the simultaneous transmission of multiple bit streams across the same frequencies at the same time. The result is the doubling of the transmission speed. Figure 22 shows a typical LTE 2 x 2 MIMO downlink scenario. The base station has two antennas and transmitters simultaneously transmitting down to the mobile. Likewise, the mobile device has two receive antennas and receivers that are also simultaneously receiving the data transmission.

Figure 22: 2 x2 MIMO Antennas



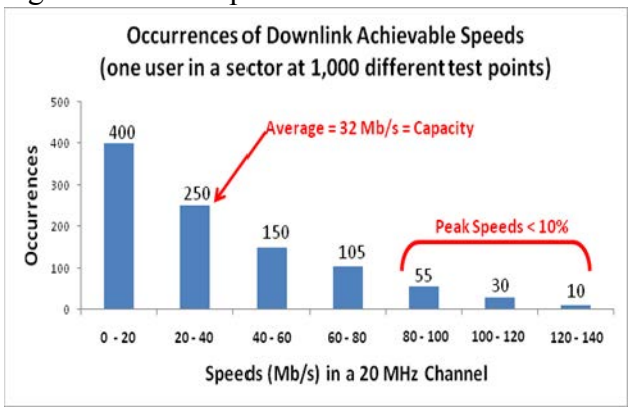
Unfortunately the uplink direction (mobile to the base station) does not employ 2 x 2 MIMO. Because of the high cost and difficult implementation issues associated with multiple transmitters in a small, low power mobile device, only 1 x 2 MIMO is used in the uplink direction. Therefore a single transmitter is used in the mobile and multiple receivers and antennas are used in the base station. This limitation is a contributor to the lower speeds associated with the uplink versus the downlink in an LTE network.

Wireless Peak Speeds

There is a myth that seems to be continually perpetuated in the wireless industry regarding how often peak speeds can be obtained by end users. Peak speeds in 3G and 4G wireless technologies are obtainable only when the maximum modulation mode is used. Unfortunately, these maximum modulation formats (e.g. - 64 QAM) are possible only when perfect RF (Radio Frequency) conditions exist. Wireless technologies differ from terrestrial in that they use variable modulation and error correction formats. Only if the mobile receives the strongest signal from the base station will the highest modulation and most forgiving error correction formats be used, resulting in the peak speeds.

Figure 23 illustrates the 3GPP (Third Generation Partnership Project) mobile standards body simulation results of an LTE device in a 4G data network. In the simulation, 1,000 different test points were assumed in a single sector. Each point assumed differing losses and interference levels and the device having full access to the capacity of the sector without competing for capacity with other users.

Figure 23: Peak Speeds in an LTE Network [18]

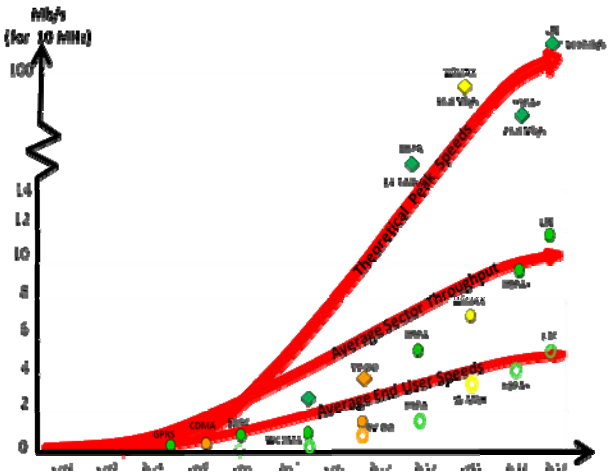


A main conclusion resulting from the data in Figure 23 is that peak speeds are possible in

an LTE wireless network less than 10% of the time. Actual testing performed in operational HSPA networks also validates this result in 3G networks.[19]

Another important conclusion of Figure 23 is the determination of the true capacity of an LTE channel. The average speed of the 1,000 test points is 32 Mb/s. This average is called the average sector throughput and is used by 3GPP and LTE vendors to determine the spectral efficiency of the technology. Since the data assumed a 20 MHz channel, the spectral efficiency is 1.6 bps/Hz (32 Mb/s ÷ 20 MHz). Figure 24 provides us with a side by side comparison of peak speeds, the true capacity of a wireless network (called average sector throughput) and the average end user speeds. These speeds are plotted over time as the various technologies (2G, 3G & 4G) have advanced the speeds and capacities possible in wireless networks.

Figure 24: Wireless Peak and Average Speeds



The uppermost curve represents the peak speeds achievable only 5 % to 10% of the time. The middle curve illustrates the "true" capacity of a wireless technology. It is equivalent to the capacity of a DOCSIS channel (50 Mb/s for a 8 MHz channel operating at 256 QAM) or a VDSL2 line (35 Mb/s at 2,000 feet). The average sector throughput number is calculated using probability or statistical means (e.g. - the Figure 23 methodology). Finally, the bottom curve is the average end user speed that a customer will truly receive. These numbers are determined after the typical oversubscription (or concurrency) calculations are applied to the average sector throughput values.

Technology Comparison

Figure 25 shows a comparison chart of the four main technologies discussed in the technology section of this paper. It is clear that the upcoming enhancements to the fiber to the home GPON and EPON technologies offer the highest capacities on both the upstream (2.5 Gb/s) and downstream (10 Gb/s) directions. The combination of high line rates and a low number of shared users (64) in the PON examples makes it a difficult technology to exceed. HFC-Fiber Deep comes very close to matching FTTH PON and offers quite attractive speeds in an 860 MHz plant (4.85 Gb/s) and could exceed 5.7 Gb/s if a 1 GHz plant is assumed. The biggest issue on the HFC-FD comparison chart is the allocation of this bandwidth across a larger amount of customers (125 versus 64) and the much lower upstream capacity.

Figure 25: Access Technology Comparison

Access Technology	Capacity		Homes Passed (HP)	Access Architecture	Assumptions & Restrictions
	Downstream	Upstream			
GPON and EPON FTTH	10 Gb/s	2.5 Gb/s	64	Shared to home	Build new fiber to the home
VDSL2	35 Mb/s	2 Mb/s	~ 125	Dedicated from node to home	Fiber and electronics built within 2,000 ft. of homes
HFC –FD (Europe 860 MHz)	5.85 Gb/s	270 Mb/s	~ 125	Shared to homes	Fiber to the FN w/o RF actives and all digital, all IP network
LTE	32 Mb/s	15 Mb/s	depends on market density	Shared to Homes	Cell sites < 1Km of homes, 20 MHz channels and 3 sectors/site

Both DSL and LTE offer much lower capacities where the dedicated nature of capacity to a single user makes DSL have higher speeds. If dedicated video delivery is assumed for VDSL, then little capacity (say 10 Mb/s as shown in Figure 10) is left for

broadband services and makes it quite comparable to LTE shared capacities. In fact, both the Frigo and Shankaranaryanan technical papers from AT&T show how a shared 30 Mb/s channel is equivalent to a dedicated 10 Mb/s DSL channel.[20, 21]

DEMAND AND CAPACITY

What levels of demand can each technology support?

We utilized a typical high density city network architecture (HFC n+0 and N+3), unicast service demand profile and cost structure as a specific high density, underground scenario for modeling. We defined two demand profiles, Moderate and Heavy, with the unicast service types including Internet, Voice, and Video on Demand, projected out to 2014 as described in Tables 1 and 2, below. For the purposes of a unicast only analysis, we assume a conservative flat "broadcast floor" of 62 channels (30 analog, 29 digital multiplexes, 3 not usable) across all years.

Table 1: Unicast Service Profile, Moderate *(Digital Max 26% Pen., No 3D-VoD, Internet Kbps Growth 1.6x Per Year)*

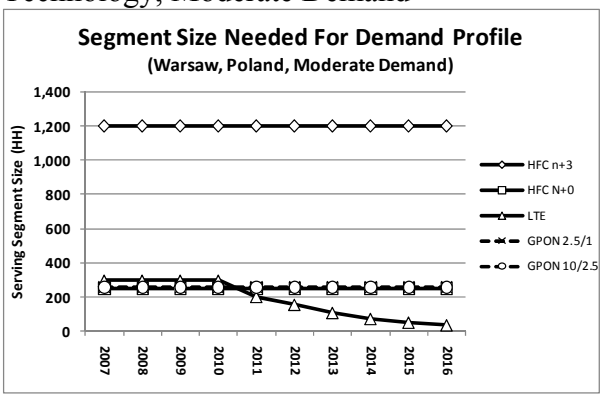
	2009	2014
HSD Pen.	23%	26%
Voice Pen.	12%	19%
DTV Pen.	14%	**32%**
HSD,Peak /Sub	30 Mbps	**1,024 Mbps**
HSD Wtd /Sub	9 Mbps	**70 Mbps**
HSD,Avg /Sub	100 Kbps	**1,050 Kbps**
SD VoD	Yes	Yes
HD VoD	No	Yes
3D VoD	No	**No**
Traffic Per HH	69 Kbps	448 Kbps
Traffic Per Sub	179 Kbps	1,597Kbps

Table 2: Unicast Service Profile, Heavy *(Digital Max 36% Pen., 3D-VoD, Internet Kbps Growth 2.0x Per Year)*

	2009	2014
HSD Pen.	23%	26%
Voice Pen.	12%	19%
DTV Pen.	14%	**39%**
HSD,Peak /Sub	30 Mbps	**1,024 Mbps**
HSD Wtd. /Sub	9 Mbps	**290 Mbps**
HSD,Avg /Sub	100 Kbps	**3,200 Kbps**
SD VoD	Yes	Yes
HD VoD	No	Yes
3D VoD	No	**Yes**
Total Per HH	69 Kbps	1,163 Kbps
Traffic Per Sub	179 Kbps	4,062 Kbps

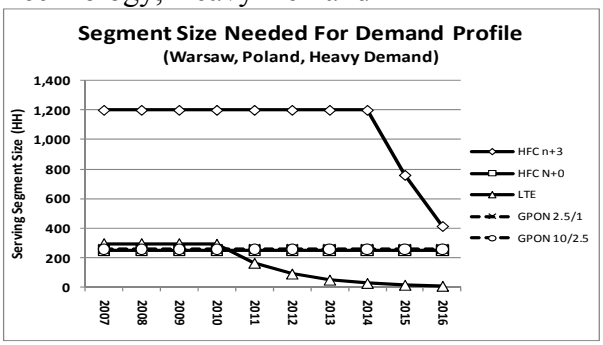
We can see that by 2014, under the moderate scenario each home will demand 448 Kbps of unicast bandwidth, and with the heavy scenario each home will demand 1,163 Kbps of unicast bandwidth. This has a varying impact on the serving size needed by 2014 as noted in Figure 26 and Figure 27 below.

Figure 26: Serving Segment Size by Technology, Moderate Demand



In the moderate demand scenario, by 2014, the 1200 homes passed HPC n+3, HFC n+0, GPON2.5/1 and GPON10/2.5 are able to support the demand profile; while the LTE 296 homes covered per sector reduces significantly to 70 homes per sector. Even with the capacity constraints of LTE, comparable HFC and PON speeds are unavailable.

Figure 27: Serving Segment Size by Technology, Heavy Demand



In the heavy demand scenario, by 2014, the 1200 home passed HPC n+3,HFC n+0, GPON2.5/1 and GPON10/2.5 are able to support the demand profile; the LTE 296 homes covered per sector reduces to a very small 28 homes per sector … that's almost a "base station in every home[22]" !

How many broadband subscribers can each access technology support?

If we assume that the peak Internet speeds required to remain competitive in 2014 are 1 Gbps and that the peak advertized speeds represents historically about 60% of the port capacity, this means we require a port size of about 1.6 Gbps or the equivalent of about 32 channels of HFC capacity would be required. Advances in electronics such as channel bonding can enable an operator, with enabling spectrum, to support higher peak speeds, but as a result also provide segment capacity without the need for dramatically smaller serving group sizes.

Using Shankaranaryanan's 2001 Equivalent Circuit Rate approach[23], We take LTE, HFCn+3/HFCn+0 with 32 bonded channels or 1.6 Gbps, GPON2.5/1, GPON10/2.5, and plot the subscribers supported for differing speeds, including the weighted average product speeds. Since each technology has a different serving group size, we examine at the number of customers supported and consider the penetration level that can be supported, where GPON is assumed to be dimensioned at 4 OLT PONs x 64 ONTs or 256 homes connected. At 30% penetration that translates into 853 homes passed.

We find that where technology advances and spectrum availability allow, HFC n+3 can provide a good fit to the moderate demand profile in 2014, in Table 3 below, with n+0 a good option for supporting heavier demand if needed for additional unicast services, in Table 4 below, while LTE is unable to support 1Gbps speeds; and GPON 10/2.5, far exceeds the demand profile for 2014 in its capacity requirements even for peak speeds.

Table 3: Internet Subs Supported at 70 Mbps Weighted Average Speeds in 2014, Moderate Demand

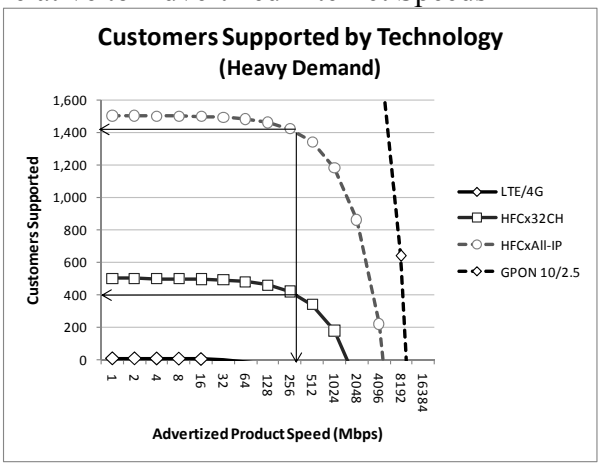
	LTE	HFC n+3	HFC n+0	GPON 10/2.5
70 Mbps Wtd.Avg.	0 Subs	478 Subs	478 Subs	3,178 Subs
Group Size HH	296	1,200	250	853
Max Pen. %	0%	40%	191%	372%

Table 4: Internet Customers Supported at 290 Mbps Weighted Average Speeds in 2014, Heavy Demand

	LTE	HFC n+3	HFC n+0	GPON 1 0/2.5
290Mbps Wtd.Avg.	0 Subs	409 Subs	409 Subs	3,109 Subs
Group Size HH	296	1200	250	853
Max Pen. %	0%	34%	163%	364%

As we can see from Figure 28 below, considering only Internet demand, HFC n+0 would not be required yet, noting the vertical arrow at 290 Mbps weighted average product speed intersecting with 409 customer supported or 34% penetration, with the potential for more where spectrum allows.

Figure 28: Internet Customers Supported relative to Advertized Internet Speeds



With the increased demand profile require to support all unicast traffic, we can see that the increase demand profile results in fewer homes supported, where the 290 Mbps weighted average speed intersects with a reduced 323 homes, or 26% penetration, potentially a candidate for either HFCn+0, or where spectrum allows additional channel bonding.

Moreover, LTE is not able to offer either the peak speeds or capacities of fixed line alternatives. Comparing the fixed line alternatives, Telcos GPON exceeds what is needed by 2014; while HFC has an incremental flexible approach to meet future demand. Advances in electronics are able to leverage spectrum to reach peak speeds, and HFC Fiber Deep is able to be used to reduce serving segments sizes.

COST ECONOMICS

A common factor when considering fiber-to-the-home, HFC fiber deep, and LTE, is that they are all capital-intensive. We compare the fixed upfront cost for each alternative on a 'greenfield' and upgrade basis. Varying assumptions for outside plant environments (e.g. - aerial versus underground) and wireless broadband frequencies and spectrum quantities are analyzed.

Greenfield or New Build Costs

We considered 'greenfield' capital costs associated with each technology, including LTE, HFCn+3, HFCn+0, and GPON/FTTH.

What do we define as 'greenfield' capital costs?

We assume that, with the exception of LTE, each 'greenfield' design will be able to support the heavy demand profile in 2014, noted earlier. We included LTE upgrade costs for comparison purposes, even though it will not be able to support the demand profile or peak end user speeds required.

Included in the 'greenfield' capital costs are the cost to (a) build the distribution network including materials such as optical, coax, splitters, combiners, nodes and amplifiers, in addition to the cost of aggregation electronics such as the BTS, CMTS[24] and OLTs; and labor for ducting or pole mounting; and (b) the materials and labor cost of the drop from the distribution network to the home; We excluded any rights of way costs, NMS, OSS, BSS costs, backhaul costs, headend costs, and Customer Premise Equipment (ONTs are included) costs.

The following cost economics are based on an analysis of actual build costs for high density cities (1,508 HH and 1,754 HH per Km^2)[25,28] Those high density, underground examples were used to baseline labor rates and materials against actual U.S. builds of varying density.

Table 5: Greenfield Cost Per Home Passed/Covered

	LTE	HFCn+3	HFCn+0	GPON
Greenfield High Density Underground	$106	$381	$374	$700
Greenfield High Density Aerial	$106	$124	$140	$231
Greenfield Low Density Aerial	$296	$700	$750	$1,438
Greenfield Low Density Underground	$296	$1,080	$1,229	$1,871

Table 6: Greenfield Cost Per Home Connected

	LTE	HFCn+3	HFCn+0	GPON
Greenfield High Density Underground	$0	$97	$97	$650
Greenfield High Density Aerial	$0	$37	$37	$590
Greenfield Low Density Aerial	$0	$37	$37	$693
Greenfield Low Density Underground	$0	$97	$97	$750

We also considered U.S. public FTTH material in the context of high labor cost economics, including the following cost

outline in Table 7 below, where Jaguar & Hiawatha are rural U.S. deployments:

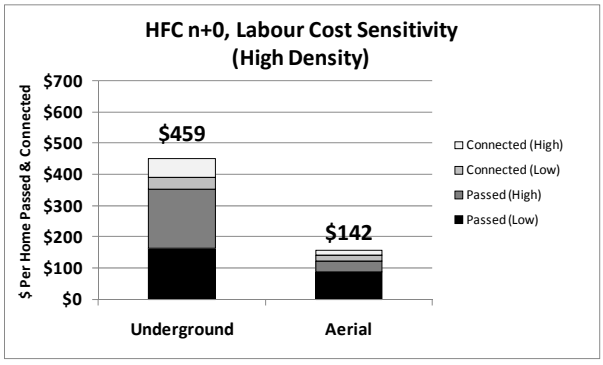
Table 7: Cost To Pass A Home [26]

Service Provider	Cost to Pass	Cost to Connect	Density
Verizon	$700	$650	High
Jaguar	$1,438	$693	Low
Hiawtha	$1,871	$750	Low

Using this analysis we explored several scenarios for new build, considering aerial vs underground plant and high density vs low density conurbations, noted in Table 4 above. Low density aerial scenarios are probably more representative of U.S. topologies.

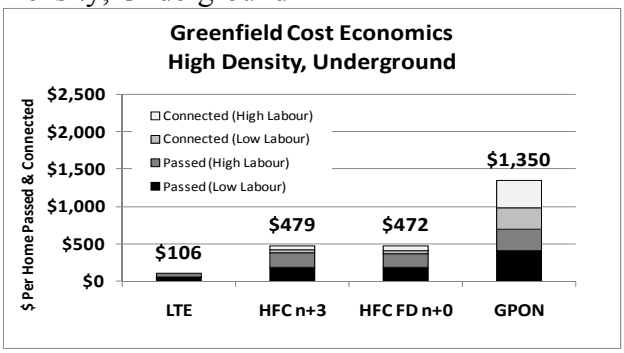
We looked in further detail at the labor sensitivity component for a specific example. In the chart below we show that for HFCn+3, in a high density market, the underground cost per home passed is $105 for markets with low labor costs based on an analysis of HFCn+3 and HFCn+0 vs $361 for markets with high labor rates (i.e. - the U.S). Adding connection costs, this translates into $142 per home connected in low labor cost markets and $459 per home connected in high cost labor markets.

Figure 29: Underground New Build, With Its Substantial Labor Component, Is Sensitive To Individual Market Labor Costs.



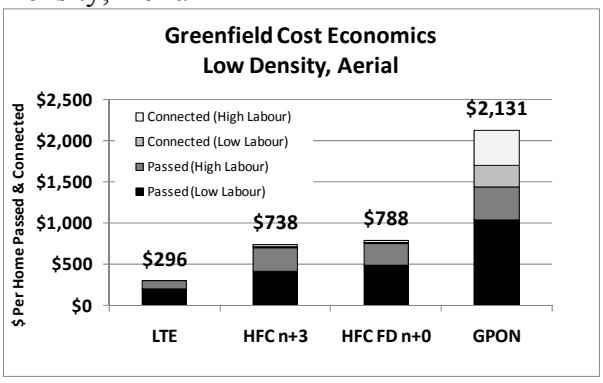
Considering the typical high density, underground scenario across the broadband technology choices, we see in Figure 30 below, keeping in mind LTEs capacity and service type limitations, it is able to achieve cost effective coverage at $106 per home covered, compared to HFC n+3/HFC n+0 at about US$475 per home connected, and GPON at US$1,350 per home connected.

Figure 30: Greenfield/New Build: High Density, Underground



Conversely in a low density, aerial scenario, in Figure 31 below, results, as expected with lower densities, in a higher cost per home connected, at around $750 per home connected for HFCn+3/n+0 and $2,131 per GPON home connected.

Figure 31: Greenfield/New Build: Low Density, Aerial



Upgrade Costs

We also considered upgrade costs from HSPA+ to LTE, HFC n+3/DOCSIS 2.0 to HFC n+0/DOCSIS 3.0 with channel bonding and DSL to GPON. In this way, we believe that the upgrade economic comparison is fair in that all technologies are able to offer end users faster peak speeds after the upgrade.

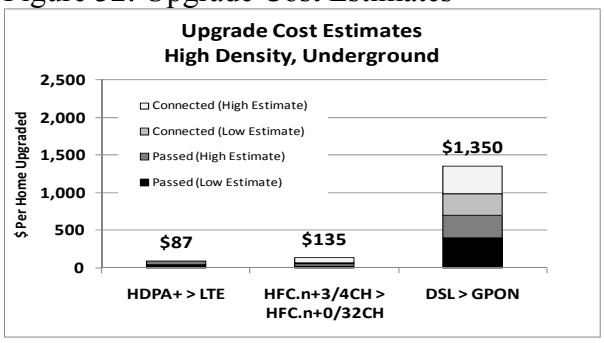
What do we define as an Upgrade cost?

For the purposes of this analysis, we assumed for LTE that a minimum of additional site licenses, cards and radios would be required to upgrade from HSPA+ to LTE, an additional 40MHz of spectrum (2x20MHz) at US$0.03 per MHz per head (low estimate in Figure 32), and in addition, we assume an additional upgrade may be needed from 3 sectors to 6 sectors; and that additional spectrum may cost up to US$0.06 per MHz per head of population[27] (high estimate in Figure 32).

In the case of upgrading from HFCn+3 with 4 bonded channels to HFC n+0 with 32 bonded channels we assumed that CMTS electronics would be required to provide fast speeds and that the technology would be available to support this at US$20 per home passed[28], (low estimate in Figure 32) while segmentation may be required in a high case for the heavy demand scenario. We assumed segmentation "can exceed $10,000 per node split[29]", and we used a range of $5,000 to $25,000 per service group (high estimate in Figure 32).

For an upgrade from DSL to GPON we assumed that, due to the need to replace most of the plant to a completely different architecture, the upgrade cost would be the same as Greenfield, and that low and high estimates are largely a function of the labor cost variations between different markets. We assume $250 for ONT pricing, using HFC's labor drop costs for the low estimate, and using public total drop costs to determine the high estimate.

Figure 32: Upgrade Cost Estimates



It is clear from this comparison that those operators with spectral flexibility (Mobile, Cable) are able to leverage advances in electronics to meet faster peak information rates; where as other operators that lack spectral flexibility (Telco) require a step function in order to move to a new last mile technology (i.e. from Copper to Optical) in order to overcome information rate limitations.

For the purpose of assessing the business model, we assumed that a home passed is a fixed cost, and a home connected is a variable cost that increases as the penetration of homes increases.

BUSINESS MODEL

We provide a sensitivity of the access technology alternatives by market density, broadband penetration and product speeds; and using illustrative unicast service revenues for future broadband services; we look at the Greenfield business model [30].

Table 8: Hypothetical Monthly Unicast Service Revenues Per Subscriber 2014

	Access Network		
	Mobile Broadband	Cable HFC	Telco GPON
Voice	$25	$0	$0
Data	$15	$25	$25
VoD	$0	$15	$15
Total	$40	$40	$40

Investment thresholds

We assume that the hypothetical unicast service revenue for each technology is $40 per month per subscriber, applying the data revenue projection from Figure 2, and assuming that mobile voice has a significant value to the subscriber relative to fixed voice. We also assume that Video on Demand (VoD) revenue for the mobile device will have a low

value to the subscriber relative to fixed VoD services that can be viewed on a large screen. What is the upper limit of capital expenditure per subscriber that may be justified by the operator?

Using an approach described by Friggo, Lannone and Reichmann, AT&T Labs Research in an IEEE Optical Communications paper in 2004[31], we looked at a selection of operators in Table 9 below and concluded that the upper limit an operator could tolerate would be about 15% of revenue in interest payments.

Table 9: Interest Expenses as a Proportion of Revenue in 2009[32]

	Revenues (m)	Interest Expenses (m)	% of Revenue
Telco A	$107,808	$4,209	3.7%
Cable A	$35,756	$2,040	5.7%
Cable B	$11,080	$946	8.5%
Cable C	$6,755	$1,088	16.1%

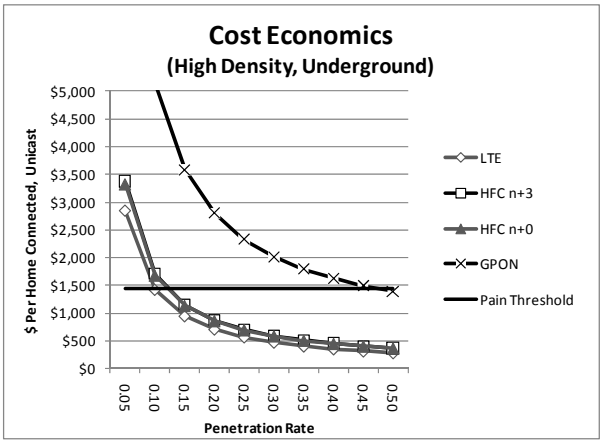
We assumed an upper limit of 15% of revenue for interest payments and an annual interest rate of 5% we deduce that a worst case capital payback time of 3 years or 36 months provides the payback limit. Assuming $40 per month in service revenue we can project that the operator capital expenditure "investment threshold" is $1,440 per home connected.

Under what conditions may the investment related to each access technology be argued to outweigh the economic benefit to be realized by the operator?

Focusing on 'greenfield' cost economics, we considered each technology by penetration rate, applying the fixed home passed associated with unicast traffic and variable home connected associated with uncast traffic for the "high density, underground" and "low density, aerial" scenario's described in the cost economic section above. High labor rates are assumed for both scenarios, and where broadcast services are supported the proportional plant and drop costs are excluded for fixed services so as to consider only the unicast element. We assumed LTE is all unicast, that for HFCn+3/n+0 32 of 91 usable channels related to unicast in 2014, and for GPON we assumed that of 3 wavelengths, 1 is for upstream, 2 are for downstream services of which 1 is for broadcast and the other is for unicast. Upgrade costs are excluded.

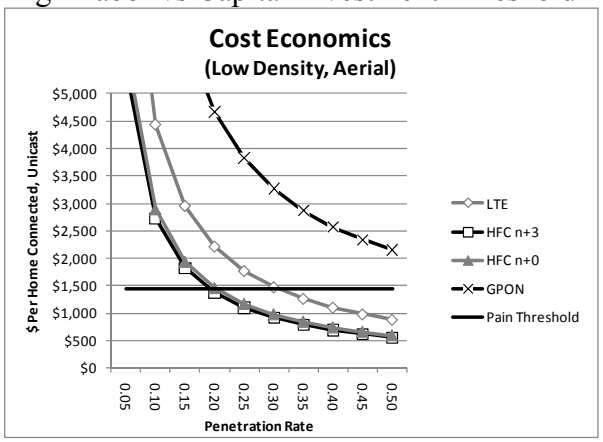
In Figure 33 and Figure 34 below, the investment threshold represents the upper limit of capital expenditure tolerated by the operator. As penetration increases, the cost per home connected falls to an intersection with the investment threshold, identifying the penetration level required to meet the payback period associated with the investment threshold.

Figure 33: Greenfield, High Density, Underground, High Labor vs Investment Threshold



Looking at where each technology crosses the investment threshold (point where technology pays back), we can conclude that LTE, HFCn+3 and HFCn+0 fall below the investment threshold at relatively low penetration levels of 10%, noting that LTE has lower end user speeds, and fewer services. GPON only crosses the investment threshold at a penetration of 50%.

Figure 34: Greenfield, Low Density, Aerial, High Labor vs Capital Investment Threshold



Considering the Low Density, Aerial scenario, we can see that all technologies require a higher penetration to fall below the investment threshold. In this scenario, HFCn+3/HFCn+0 fall below the investment threshold at 20%, LTE falls below the investment threshold at 30%, and GPON falls below the investment threshold at about 80% penetration.

CONCLUSION

The competitive, technical and economic findings are summarized[33].

Competitive: Competition Drives the Need for Faster Peak Information Rates (PIRs), Limits Service Penetration Potential.

Competition is driving the need for faster Peak Information Rates (PIRs) which in turn forces the operator to make technology choices to remain competitive in the market place. However, the corn does not grow all the way to the sun. Operators realize that, in a competitive environment with multiple service providers and similar service types, there may be a constrained "customer base potential" upon which to examine relevant 'investment thresholds". These thresholds may define an upper limit to 'greenfield capital expenditures. Additionally, those operators with compelling upgrade economics that do not require a move to a new last mile technology, are best able to compete.

Technical & Capacity: Because DSL is challenged, HFCn+3/n+0, and GPON will be the key broadband technologies in the future; LTE, on the other hand, cannot provide the same peak speeds, capacity or services types as fixed line alternatives.

Considering the technology alternatives, we note that LTE, while it has a unique attribute of mobility, is not able to support the capacity required, peak speeds (e.g.- 1 Gbps) or all of the service types to the home that will be delivered by fixed line technologies in 2014; and is therefore unable to offer a complete substitute for fixed line services. Fixed line operators face decisions about capacity based demand and competing peak information rates. HFCn+3/n+0 offers great flexibility to meet both varying demand scenarios, and increasing peak information rates.

Economic: In Competitive Markets, There Is No Business Case For Physical Replacement of Plant for Faster PIRs, Low Density New Build.

Cable operators do not have to re-build physical plant in the process of increasing peak information rates, but rather rely on spectral flexibility and/or advances in electronics. This results in an advantage in 'non-greenfield' markets or markets reaching subscriber saturation over operators that require a complete rebuild or plant. Telcos need to totally rebuild their plant to match Cable operators. Mobile operators do not need to rebuild, but they cannot meet the speeds required to compete.

REFERENCES

[1] Tom Cloonon, "On The Evolution of HFC Network and the DOCSIS CMTS – A Roadmaps for the 2012 – 2016 Era", SCTE CableTec Expo 2008.
[2] Screen Digest, Average Broadband Speeds, ARPUs, 2006 to 2009.
[3] Burke & Eagles, Mobile TV Paper, NCTA 2009
[4] Frigo N., Lannone P, Reichmann K., "A View of Fiber To The Home Economics", AT&T Labs, Aug 2004, pS20
[5] Verizon News Center, "FiOS Product Information Sheet, Q1 2007, http://newscenter.verizon.com/kit/nxtcomm/Product-sheet-FiOS-1Q07.pdf
[6] Stacey Higginbotham, Is Verizon FiOS Putting the Hurt on Cable?, GigaOM, Jul. 27, 2009
[7] Update on KPNs Fiber Rollout, 12/09
[8] Swisscom FTTH Council presentation, "Swisscom Fibre Optics or Fibre Suisse", Copenhagen, 2/08
[9] Maisie Ramsay, "Landline Now Optional for AT&T Triple Play", Wireless Week, http://www.wirelessweek.com/News/2010/03/Carriers-Landline-Optional-Triple-Play-ATT/
[10] Screen Digest, Market Statistics,Feb 2010
[11] BroadbandReports, "Verizon: We've Neglected DSL", Jun 11, 2008, http://text.broadbandreports.com/shownews/Verizon-Weve-Neglected-DSL-95186
[12] Aware Inc., VDSL2 White Paper 5.06
[13] Future Directions in the Battle Between Cable & PON, Bernstein & Page, SCTE Paper, 2007
[14] Future Directions in the Battle Between Cable & PON, Bernstein & Page, SCTE Paper, 2007
[15] Modern Cable Television Technology, Ciciora, Farmer, Large, Morgan-Kaufman, 1999
[16] Exploiting HFC Bandwidth Capacity to Compete with FTTH, Werner & Sniezko, NCTA Paper, 2005
[17] Technology to the Rescue-Optical Architectures for Increased Bandwidth Per User, Sniezko, NCTA Paper, 2007
[18] 3GPP web site,www.3gpp.com LTE Spectral Efficiency Calculations
[19] Resavy Research, 2009, LTE
[20] Frigo N., Lannone P, Reichmann K., "A View of Fiber To The Home Economics", AT&T Labs, Aug 2004
[21] N. K. Shankaranaryanan, Z. Jiang, and P. Mishra, "User-Perceived Performance of Web Browsing and Interactive Data in HFC Cable Access Networks," Proc. ICC, June 2001.
[22] Duncan Graham-Rowe, "Why every home should have a cellphone mast", New Scientist, 10 March 2008, Magazine issue 2646, http://www.newscientist.com/article/mg19726466.200-every-home-should-have-a-cellphone-mast.html
[23] N. K. Shankaranaryanan, Z. Jiang, and P. Mishra, "User-Perceived Performance of Web Browsing and Interactive Data in HFC Cable Access Networks," Proc. ICC, June 2001.
[24] Saul Hansell, "World's Fastest Broadband at $20 Per Home", New York Times, April 3, 2009, http://bits.blogs.nytimes.com/2009/04/03/the-cost-to-offer-the-worlds-fastest-broadband-20-per-home/
[25] European Central Statistics
[26] Om Malik, "How Much Will Google's Fiber Network Cost", Feb 11, 2010, http://gigaom.com/2010/02/11/google-
[27] Tim Farrar, "Is the value of spectrum going up or down?", TMF Associates, Aug 25, 2008, http://tmfassociates.com/blog/2008/08
[28] Saul Hansell, "World's Fastest Broadband at $20 Per Home", New York Times, April 3, 2009, http://bits.blogs.nytimes.com/2009/04/03/the-cost-to-offer-the-worlds-fastest-broadband-20-per-home/
[29] Brady Volpe, "DOCSIS 3.0 Introduction", Oct 12, 2009, http://bradyvolpe.com/2009/10/12/docs
[32] Ethernet Passive Optical Networks, Glen Kramer, McGraw-Hill, 2005
[31] Frigo N., Lannone P, Reichmann K., "A View of Fiber To The Home Economics", AT&T Labs, Aug 2004, pS20
[32] Year End 2009 SEC 10-K statements for referenced companies, in $US.

ADAPTIVE STREAMING – NEW APPROACHES FOR CABLE IP VIDEO DELIVERY

John Ulm, Tom du Breuil, Gary Hughes, Sean McCarthy
Motorola Corp., Home & Mobility

Abstract

Adaptive Bit Rate Streaming is a technology being deployed to deliver IP video to personal computers and mobile devices over the internet. This paper provides a tutorial on this technology and its application in Cable IP Video delivery systems.

We will explore the impact of Adaptive Bit Rate Streaming on topics such On Demand unicast services & Linear TV multicast services; transcoding; unmanaged home networks; Ad insertion; impacts on CDN; video bandwidth efficiencies and Migration Strategies.

INTRODUCTION

Interest continues to accelerate for supporting IP Video on a cable infrastructure. Many factors have contributed to this including the exponential growth of over-the-top entertainment quality video, more broadband homes with higher speeds, significantly more efficient H.264 video and AAC audio codecs and the ease of integrating PC and smart phone experiences.

One of the critical questions is how to choose the best video delivery mechanism for all IP delivery. Recently we have seen significant interest in using emerging Adaptive Bit Rate Streaming protocols from the mobile and PC arena for cable IP video delivery. Called Adaptive Streaming for short, it enables smoother playback across a variety of internet-connected devices and is optimized for internet video delivery. But how well suited is adaptive streaming for IP video delivery across all screens including the TV?

This paper will provide operators with an overview of the new adaptive streaming protocols. IP Video delivery has evolved from streaming and progressive downloads to something that's evolved to scale for world wide delivery. Key to this is the use of HTTP for the underlying video transport. Various proprietary ecosystems have been deployed by companies such as Apple, Microsoft and Adobe while various standardization efforts are now underway.

We discuss the strengths and weaknesses when using adaptive streaming for IP video delivery over a cable infra-structure. Managed IP video service must consider both Linear TV and On Demand content delivery. Linear TV is associated with real-time and often multicast delivery while On Demand is stored content with unicast delivery. A critical issue an operator must tackle is where to transcode the IP video into the various formats. Other topics include Ad insertion, delivery over unmanaged home networks, policy & asset management and finally, migration strategies.

These new protocols will also have a significant impact on an operator's Content Distribution Network (CDN), video servers, and edge distribution. Servers must evolve from their current streaming operations and efficiently handle the multiple new formats needed for each asset. We'll also take a look at Trick Mode support and CDN bandwidth and caching.

Finally we will take a look at the video bandwidth efficiencies that we gain with adaptive streaming compared to traditional video broadcast models, including the impact of Variable Bit Rate (VBR) video delivery.

OVERVIEW – ADAPTIVE STREAMING

Background

Traditional Internet Streaming video delivery to PCs was designed with real time protocols and provides the content as you need it with minimal buffering requirements. Some common protocols used include Real Time Protocol (RTP) for the video transport and Real Time Streaming Protocol (RTSP) or the Real Time Messaging Protocol (RTMP) over TCP for the control. These stateful protocols work well in controlled networks including enterprise or service provider environments.

The real time nature provides a responsive user experience with well defined bandwidth usage. However, the real time nature often requires a separate network for video streaming and doesn't work well for distribution over the internet. This approach also does not support standard CDN networks using HTTP caching and has potential scaling issues.

Traditional Streaming also has issues in traversing through firewalls in routers. The real time protocols use ports that are different from traditional web browsing and often require the router/gateway to be manually configured. This is a significant problem for wide spread use in consumer managed home networks.

Progressive Downloads were designed to deliver content over the internet. It works from a standard web server and uses HTTP as the transport protocol. This enables it to scale on a world wide basis by leveraging standard HTTP caching and it eliminates any issues with getting through firewalls since it uses the same ports used for web browsing.

However, there are several significant drawbacks to Progressive Downloads. User experience is impacted with significant latency while the buffer is filled and re-buffering followed by video pauses when there are insufficient network resources.

Progressive Downloads can place added buffering requirements on the user devices and be wasteful of bandwidth as well, especially with un-throttled download speeds. In a very common use case, a user stops watching the content after a short period of time (e.g. channel surfing), but most or all of the video content is still downloaded, consuming excessive network resources.

A Hybrid Approach – The Best of Both

In both streaming and progressive downloads, the video content is encoded with a fixed rate/quality model. If available network bandwidth is reduced, the user may experience starts and stops in the picture or be forced into long delays as buffers fill. If network bandwidth is in abundance, then the user may be viewing content at lower quality than what the system is capable of delivering. Neither protocol adapts well to changing network resources.

So the key challenge is how to deliver great viewer experiences over variable uncertain bandwidth to a wide variety of display devices, not just PCs. Adaptive streaming was developed to capture the best of both streaming and progressive downloads.

Basic Operation - Chunking and Play Lists

Adaptive streaming is a hybrid delivery method that acts like streaming but is in fact a series of short HTTP progressive downloads. It relies on HTTP as the transport protocol and performs the media download as a long

series of very small files, rather than one big progressive download file.

The content is cut into many small segments and encoded into the desired formats. These small segments are often called chunks, streamlets or fragments and typically cover 2 to 10 seconds. A chunk is a small file containing a short video segment along with associated audio and other data.

Adaptive streaming typically uses HTTP as the transport for these video chunks. This gives it all the benefits of progressive download. The content can easily traverse firewalls and the system scales exceptionally well for high demand as it leverages traditional HTTP caching mechanisms.

By using small chunks of video, adaptive streaming also behaves like traditional streaming and is applicable to both live delivery and pre-stored on demand content.

The new twist that adaptive streaming introduces is the ability to switch between different encodings of the same content. This is illustrated in Figure 1. Depending on available bandwidth, you can choose the optimum encoding thus maximizing user experience.

Each chunk or fragment is its own stand-alone video segment. Inside each chunk is what MPEG refers to as a GOP (Group of Pictures) or several GOPs. The beginning of each chunk meets the requirements of a Random Access Point, including starting with an I-frame. This allows the player to easily switch between bit rates at each chunk boundary. This collection of multiple adaptive streams each with different encodings is some times referred to as an adaptive stream bouquet.

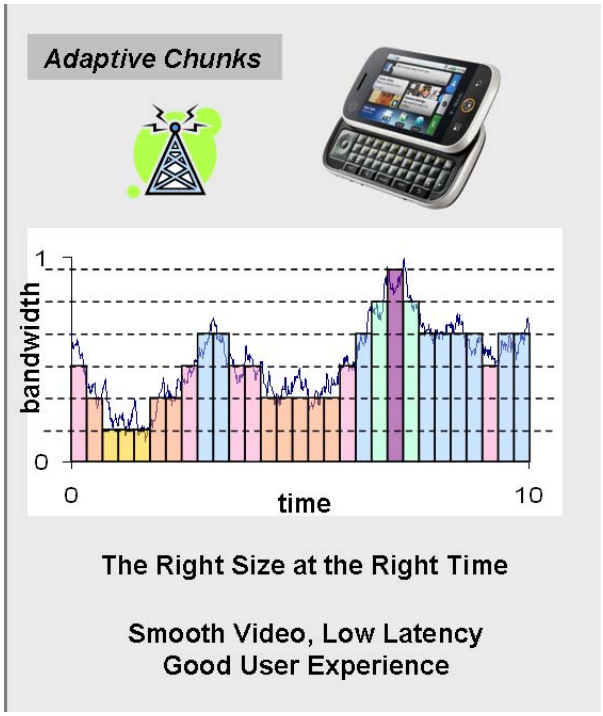


Figure 1. – Adaptive Streaming: Basics

Adaptive streaming also allows a user to start the video quickly by initially using lower bit rate chunks and then quickly switching to higher quality chunks. This provides a straightforward solution for fast channel changes, a feature valued by consumers.

Central to adaptive streaming is the mechanism for playing back multiple chunks to create a video asset. This is accomplished by creating a play list that consists of a series of URLs. Each URL requests a single HTTP chunk. The server stores several chunk sizes for each segment in time. The client predicts the available bandwidth and requests the best chunk size using the appropriate URL. Since the client is controlling when the content is requested, this is seen as a client-pull mechanism, compared to traditional streaming where the server pushes the content.

Using URLs to create the play list also enables very simple client devices using web browser-type interfaces.

Ecosystems and Standards

HTTP chunking is an underlying transport mechanism. To create an end-to-end system for video delivery requires additional components such as video and audio codecs, Digital Rights Management (DRM) and other control plane elements. As of today, different proprietary adaptive streaming ecosystems have emerged from companies including Apple Computer, Microsoft, Adobe and Move Networks.

Move Networks was an early adopter of the technology and showed in 2008 that their adaptive stream HTTP-based media delivery could be done successfully on a large scale. This included broadcast of the 2008 Democratic National Convention using Microsoft Silverlight™ as the client framework [1].used by several programmers for streaming their content over the internet [2].

Microsoft created a prototype HTTP-based adaptive streaming for the 2008 Beijing Summer Olympic games. However, this experience exposed the issues of managing the millions of tiny files that were created during this very large event. Microsoft then introduced Smooth Streaming to overcome these shortcomings by defining chunks as movie fragments stored in a contiguous MPEG-4 Part 12 (MP4) file, using features of the MP4 format to mark chunk boundaries for easy random access. This sub-format is referred to as a Fragmented MP4 file.

Apple refers to its Adaptive Steaming implementation as HTTP Live Streaming and it is used to deliver media to the iPhone and iPod Touch. Quicktime X can also play HTTP Live Streaming, enabling playback on the PC.

Adobe has worked on an extension to RTMP called RTMP Chunk Stream Protocol. While designed to work with RTMP, it can handle any protocol that sends a stream of messages.

All these ecosystems use Advanced Video Coding (AVC, a.k.a. MPEG-4 part 10 or H.264) for their video codec and generally use AAC audio. These modern codecs are valued for their efficiency. However, each ecosystem supports different chunk file formats, typical or recommended chunk sizes, chunk file overhead, number of files to manage at the server and ways of chunk file creation (pre-stored or on-the-fly real time).

Standardization efforts for adaptive streaming are under consideration within several standards bodies and industry groups. Several IPTV organizations are considering adaptive streaming while Apple has submitted a draft of HTTP Live Streaming to the IETF for standardization [3]. At this stage, it is too early to know which efforts will prevail and in what time frame. Some of these efforts may support more than one profile in order to interoperate with one or more of the existing proprietary adaptive streaming ecosystems.

ADAPTIVE STREAMING IN A CABLE ENVIRONMENT

Delivering across the cable managed network

Adaptive streaming uses a client-pull rather than a server-push delivery mechanism and because of this, clients can automatically and dynamically adapt to the available network bandwidth available to them, enabling a smooth video experience albeit with variable video quality. This is extremely

useful for over-the-top unmanaged internet delivery of media services. As such, adaptive streaming provides excellent support for the three screen subscriber experience when they are out of their home and off their cable provider's managed network.

This leads to the question on the role or use of adaptive streaming within the cable provider's managed network. Since adaptive streaming is client driven, each viewing session is unicast and therefore needs its own bandwidth, independent of whatever other subscribers in the neighborhood may be watching concurrently.

A provider can use different approaches to manage the user experience for this environment. These broadly fall under categories of adding sufficient bandwidth, limiting delivery to select devices/subscribers or limiting which content/applications uses adaptive steaming.

One approach is to over-provision the IP network capacity to exceed the combined bandwidth requirement of all the concurrent subscriber demands in a neighborhood. In the near term cable environments, this appears impractical until DOCSIS 3.0 costs come more into line with traditional video costs such as Edge QAM devices. There is also the issue of available spectrum which might require node splits to garner sufficient additional IP bandwidth.

Another avenue is to limit the number of devices or subscribers receiving the new IP video delivery. A provider could limit the IP video service to only PC and mobile devices, or limit the service to only their premium customers. This is another way to ensure that the available IP bandwidth is sufficient for the offered IP video load.

An alternative approach for cable providers is to manage the delivery of IP content between multicast and unicast delivery. The provider can deploy popular content as IP multicast on their networks and reserve unicast for only those services that are uniquely being consumed. This is similar to Switched Digital Video, SDV, today in that services that are not currently being watched aren't transmitted. This provides two significant benefits in that the service provider can guarantee the subscriber experience by selecting their preferred quality for each multicast service while minimizing the total network bandwidth consumed for this delivery since only a single version is delivered to multiple subscribers concurrently.

As an added refinement, the unicast services can take advantage of adaptive streaming which allows the cable operator to better manage their network resources while still providing a good customer experience. This approach with adaptive streaming for unicast services provides the same benefit as SDV in today's video networks with the additional benefit of automatically allowing more simultaneous unicast sessions at lower bitrates or fewer at higher bitrates.

Delivering across unmanaged home networks

Today, most consumer networking equipment, WiFi, or other retail video products do not support multicast delivery. However, providers can support multicast delivery through the gateways, set tops and home networks that they install and manage. Because of this, a short term strategy for providers may be to multicast services to IP set tops in the home via a service provider managed high bandwidth home network such as MoCA and then use Adaptive Streaming unicast services over the access network to other subscriber home client devices such as

PCs or WiFi-enabled smart phones over unmanaged in-home networks.

In smaller residences in sparse neighborhoods with clean WiFi installations and limited concurrent demand on this home network, service quality may be fine. But in larger homes or locations where multiple adjacent WiFi networks are competing for the same spectrum, the end user experience may suffer the usual media interruptions and re-buffering instances that were so common in over-the-top video experience prior to the introduction of adaptive streaming.

One solution is to send the entire adaptive stream bouquet from the network such that the gateway can act as a proxy for the actual PC/phone client and can forward the requested bit rate stream from this bouquet, with the obvious drawback of consuming more of the available DOCSIS bandwidth in the cable provider's network.

Another alternative is to provide one or more real-time transcoders in the gateway that can be used to dynamically transcode the source stream to the available target bandwidth on the fly. This approach adds no overhead to the DOCSIS network, but does add the cost of the real-time transcoder in the gateway device. Note, gateways may need multiple transcoders if they are to serve multiple clients concurrently.

When & Where to Transcode

As we just discussed, adaptive streaming requires content in many different formats, which presents a big challenge. Do we encode as "one size fits all"? Do we encode just High, Mid and Lo quality streams? Do we encode for Progressive Download? Do we transcode on the fly? Where do we do the transcoding? At the core, edge or home?

Creating adaptive streaming services typically requires multiple encoders or transcoders per service depending on the source content format and the desired client device formats and bit rates. These encoders must be tightly synchronized to produce a valid adaptive streaming bouquet where each service instance starts and ends at the same point in time, and with the proper bit stream semantics such that client decoders can seamlessly switch between streams within the bouquet in a seamless manner.

Transcoders that are able to deliver high quality at lower bit rates can be a significant investment, especially since several are needed to produce the appropriate adaptive streaming bouquets for the three screens. Each of these then has a preferred resolution or encoding profile in terms of the device capabilities as well as subscriber expectations in terms of delivered picture quality. This cost favors centralizing the adaptive streaming transcoders into one or two super head ends for larger operators or possibly in a hosted service offering for smaller operators. Transcoding or re-encoding closer to the source also allows the provider to better control the video quality. Offsetting this, however, is that such centralization requires more backbone bandwidth to distribute these bouquets of services around a provider's footprint, and as discussed later, it impacts storage costs for the servers and CDN.

An alternate concept that has been proposed for unicast streams is to use low cost dynamic transcoders at the edge of the network that respond directly to the client's bandwidth requests in real-time. A key advantage of this approach is that such a transcoder could deliver exactly the requested bit-rate to fit the available bandwidth, offering a wide variety of bit rates. This compares to an adaptive bit-rate bouquet that was prepared farther back in the network might only have

three or four discrete bit-rates to choose from at each chunk boundary. Offsetting this, however, is that low cost transcoders require higher bit-rates over the IP access network to deliver the same quality as a higher quality transcoder. And, since the number of unicast sessions can be very dynamic, the head end would need to be provisioned with enough edge transcoders to meet the recurring maximum loads, but would likely be underutilized most of the time.

Ad Insertion

A significant advantage of adaptive streaming is that it enables efficient ad insertion. Since the client device requests content by requesting the next appropriate chunk via a URL in a play list, the play list can be modified dynamically by either the server or client application to substitute the ad play list URLs in that appropriate locations of the media play list based on SCTE 130 signaling. This enables seamless insertion of targeted ads either in the network, or by pre-placing the relevant targeted ads onto a home gateway or DVR client and inserting them locally. In either case, the splice is entirely transparent since adaptive streaming chunk boundaries are always created to allow seamless switching from one stream to another.

By adding this ad substitution intelligence into the network servers, ads can be more dynamically targeted and overall advertising management is simpler since there is no need for a system to pre-place ads into subscriber devices. This server based approach also enables the same ad insertion capabilities across the entire range of client devices including those with storage such as DVRs and those with very little "extra" memory such as inexpensive IP set tops or smart phones.

IMPLICATIONS FOR SERVERS & DELIVERY NETWORKS

Video Delivery

The shift to Adaptive Streaming imposes significant changes to the roles of servers and delivery networks in providing video service to consumers.

As noted earlier, Cable video delivery has relied upon a push streaming model where the server is responsible for maintaining stream pacing. Network transport has typically been based on UDP. To maintain correct buffer behavior the stream is constructed to meet the requirements of the MPEG-2 Systems buffer model. Adaptive streaming however evolved from Progressive Download and is based on the client pulling segments of content over a reliable network transport as it requires them. This shifts much of the burden of pacing and buffer management from the server to the client.

Progressive Download grew out the need to deliver video over HTTP connections, and it is possible to deliver Adaptive Streaming content with simple web servers. This approach is viable for lab trials and small scale deployment. However, to successfully grow to large scale deployments may require that servers and other CDN components have some degree of media and session awareness.

Content Storage

The requirement to store multiple bit rate representations of each content essence creates additional demands for library storage. If the content is stored in 5 different bit rates over a 2:1 range (for example, an SD stream that ranges from 2 Mbps down to 1 Mbps in 0.25 Mbps increments) the storage requirement is 3.75 times that required for the highest quality stream alone. This does not

reflect the overhead of any system support files, such as index or trick mode files, which may also have to exist in multiple bit rate versions. This increased storage requirement will impact library sizing and potentially edge caching and CDN bandwidth requirements.

If simple web servers are used to host the multiple bit rate versions it may be necessary to store each fragment in a unique file. For example, to store one hour of content using the above assumptions and a fragment duration of 2 seconds would require 9000 fragments. When considering a large content library, the number of files quickly becomes excessive and may require special attention to file system tuning and layout.

As mentioned earlier, it is better to use a container file format that allows fragments to be rapidly identified and accessed, such as a fragmented MP4 file, or a stored transport stream with segmentation markers. A media aware server can then respond to requests for systematically named fragments or fragments specified by Normal Play Time (NPT) range by extracting the requested segment from the container file.

Trick Mode Support

Support for VCR style trick modes (that is, scrubbing forward or backward through the content at faster than real-time) has traditionally been a feature of Cable on demand services. In a Progressive Download environment, this style of trick mode operation, when available at all, is restricted to operating on content that has already been buffered in the client, combined with the ability to restart the download, and normal play, at a client specified offset.

In order to replicate the existing Cable On Demand experience, it will be necessary to add mechanisms to provide VCR style trick modes in an adaptive streaming environment. This can be done either in the client or in the delivery network. Some systems have explored alternate ways of displaying trick mode operations that may be more suited to adaptive streaming, for example popping up a filmstrip of thumbnail key frames as a navigation aid

CDN Bandwidth & Caching

One of the attractions of adaptive streaming based on fragmented content is that it is a good fit for the use of a Content Delivery Network to efficiently provide content to the consumer. However caching algorithms designed for traditional web traffic may not result in optimal use of network bandwidth and cache storage. In extreme cases they could result in pathological overuse of resources.

Consider the example of a consumer viewing a movie during a time when available bandwidth is varying. At the end of the session the collection of fragments in the nearest edge cache represents the bandwidth available over the duration of the session. If another consumer requests the same logical content it is desirable to reuse as many of the fragments stored in the edge cache as possible. However that next session may face very different bandwidth availability and will consequently request a different set of fragments significantly reducing cache efficiency. Using a CDN architecture that is media aware would allow for more efficient use of the cache and CDN bandwidth.

Earlier we discussed the application of adaptive streaming to ad insertion by replacing or inserting chunked ad content. A media aware network could move this processing to the edge cache. In this application the media aware edge cache could

function as an Ad Decision Manager in an SCTE130 ad insertion system.

Introducing media awareness into the CDN also helps it protect itself against misbehaving or malicious clients. Media aware servers can place bandwidth limits on client requests and detect access patterns that do not match the content attributes.

EFFICIENCY AND VIDEO QUALITY

Is adaptive streaming as efficient as traditional methods of video distribution, namely Variable Bit Rate (VBR) and Constant Bit Rate (CBR) video? That is the question we explore in this section.

VBR, CBR and P-CBR

VBR is widely used to deliver video because it is capable of producing constant video quality. In VBR methods, an encoder uses as many bits as necessary to achieve a constant target video quality. As a result, bit rate varies freely in time but no bits are wasted, at least in theory.

CBR is also widely used because bandwidth resources can be allocated with virtually no uncertainty. In CBR, an encoder causes video quality to fluctuate up and down so as to achieve a fixed target bit rate.

Adaptive streaming may be thought of as a special form CBR know as Piecewise-Constant Bit Rate (P-CBR) because every adaptive streaming chunk has a constant bit rate over its duration.

At first glance, it might seem that P-CBR would be like its parent, CBR, in the sense that video quality would not be constant. Perhaps surprisingly, P-CBR is just as capable as VBR of delivering constant video quality.

Data that illustrate constant-quality P-CBR is shown in Figure 2. The thin "noisy" line in Figure 2 shows an example of a constant-quality VBR stream. The thick flat line shows a P-CBR stream that would also result in constant video quality -- in fact, the same video quality as for the VBR stream. Both streams produce the same constant video quality because the total number of bits delivered during each piecewise-constant interval is the same for the P-CBR stream and the VBR stream.

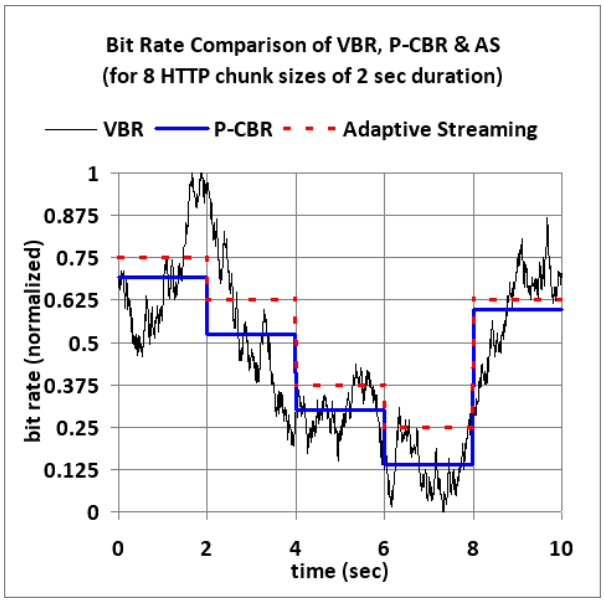


Figure 2. – Adaptive Streaming: Basics

At the end of each piecewise-constant interval, a decoder has all the bits it needs to reconstruct the preceding video. The real difference between VBR and P-CBR is latency, not video quality. P-CBR introduces intrinsic latency because the decoder needs to wait until the end of each piecewise-constant interval to be sure it has all the bits it needs.

Adaptive streaming protocols are also capable of delivering video quality that is as good as VBR, provided the client has sufficient bandwidth. This is a limiting case in which adaptive streaming may be imagined

as a coarse quantization of a P-CBR stream that is equivalent to a VBR stream in all but latency. But adaptive streaming also has the flexibility to reduce bandwidth requirements if needed by switching to a lower resolution or quality version of the content. So, adaptive streaming offers service providers the ability to deliver video quality comparable to VBR, while managing bandwidth in a simple manner like CBR.

ADDITIONAL CONSIDERATIONS

Adaptive streaming protocols are not enough by themselves to enable new approaches to cable IP video delivery. Adaptive streaming introduces new challenges, such as managing the myriad chunks, media fragments, and associated metadata. Fortunately, enforcing media policies and managing assets are not entirely new problems.

Solutions already exist for enhanced asset management systems (AMS) that are designed to package, integrate, manage, and deploy content from many different sources and distribute those assets across multiple platforms. In this adaptive streaming context, asset management challenge may be viewed as an extension or evolution of existing on-demand asset management. While adaptive streaming will come with new business and technical issues, it is reassuring that some of the challenges of dealing with "infinite catalogs" have been addressed already and can serve as a foundation for future progress.

Since Adaptive Streaming changes the underlying transport of video services, this will also generate the need for new tools for service monitoring. Service providers will want the capabilities to be able to measure the Quality of Experience. For managed IP Video services, it will always be critical to maintain video quality.

Cable Migration Strategies

Many cable system lineups today are full with a wide mix of analog, digital and high definition video services in addition to video on demand offerings, high speed data service and telephony service. Meanwhile, providers are under pressure to add additional HD services and upgrade high speed data to DOCSIS 3.0, both of which require significant additional spectrum on the cable plant. Migration to IP Video will put even more pressure on bandwidth needs.

There are many tools available today that enable cable operators to recover existing or gain new spectrum capacity in their systems. These include analog reclamation by moving analog services to digital, enabled by low cost digital terminal adapters (DTAs), Switched Digital Video (SDV), migrating services from MPEG-2 to MPEG-4 video, node splits, and HFC expansion up to 1GHz.

Using some or all of these tools to free up spectrum enables cable operators to deploy additional DOCSIS bandwidth needed for IP Video. Cable operators can deploy additional DOCSIS 3.0 bonding groups and begin using these for adaptive streaming of media. This can initially be deployed to PCs and smart phones in the home via WiFi connections. If sufficient IP resources are available, the provider can also support IP set tops connected to the managed home network. These offerings can be used for both new linear TV channels or additional on demand content and also can take advantage of MPEG-4's efficiency improvements.

During this transition period, many current set tops can be used in a hybrid configuration, using both their QAM and

DOCSIS capabilities to deliver the operator's full suite of services to their subscribers. And, over time, as more of these hybrid set tops or new hybrid gateways are deployed, cable operators can begin migrating some of their traditional QAM VOD and linear services to the IP path.

The hybrid home gateway enables the use of low cost IP-only client set tops elsewhere in the home. Eventually, when all services migrate to IP, even the gateway set top can become an IP-only device and this leads to a simpler overall system architecture that has the potential to support all IP clients across all three screens, the TV, PC, and phone, in the subscriber's home.

CONCLUSION

Adaptive streaming is an emerging technology that is of great interest to cable operators for deploying IP Video. It grew out of internet video delivery and provides the smooth user experience of streaming with the ability to scale economically like Progressive Downloads thanks to HTTP transport. Several proprietary adaptive streaming ecosystems are already in place and standardization efforts are underway. It is an obvious choice for providing services to the 2nd and 3rd screens (i.e. PC and mobile devices).

After the adaptive streaming overview, the paper took a look at using adaptive streaming in cable environment. It holds many promises and challenges. Some of the benefits include: bandwidth efficiency; minimal local storage required in user devices; support for trick mode; simplified synchronization between server and client; and expanded opportunities for targeted advertising.

We took a closer look at the impacts on the video servers and distribution networks. Some of the challenges that need to be addressed are the pressure on increased content storage and CDN bandwidth. Trick mode support and caching algorithms are other important areas that are impacted.

Finally, the paper took a deeper dive into the bandwidth efficiencies and video quality of Adaptive Streaming compared to today's VBR and CBR delivery. Adaptive Streaming holds the promise of video quality comparable to VBR with the ease of bandwidth management like CBR.

Adaptive Streaming will create new high quality multi-media distribution opportunities. It will enable rich user experiences as well as monetization of video delivery. This technology will become a cornerstone of future cable IP Video delivery systems.

Contact Info:

For more information, contact John Ulm at julm@motorola.com .

References:

1. Move Networks, http://www.pr-inside.com/move-networks-announces-microsoft-as-r773575.htm

2. Move Networks, http://www.movenetworks.com/history.html

3. Apple adaptive streaming IETF submission: http://tools.ietf.org/html/draft-pantos-http-live-streaming-01

4. Introduction to Data Compression. Khalid Sayood. Morgan Kaufmann Publishers, San Francisco, CA 2006

SIDEBAR – SOME BACKUP MATH

Efficiency of Adaptive Streaming

It is useful to investigate a limiting case to understand the bandwidth efficiency of adaptive streaming. Consider a scenario in which a provider wishes to deliver video that meets or exceeds a particular operational video-quality level. If the video is delivered using VBR, we could produce a stream such as the one represented by the thin line shown in Figure 2. If instead a P-CBR method is used with regularly spaced piecewise-constant intervals, it would produce a stream such as the one represented by the thick line in Figure 2**.** However, for adaptive streaming the chunks would be able to take on only certain pre-defined bit-rate values, and we would produce a stream such as that represented by the dashed line in Figure 2.

Recall that the number of bits delivered by the P-CBR stream in each interval is the minimum number of bits needed to achieve the target video quality. Thus, in the limiting case, the bit rate associated with each adaptive-streaming chunk must be chosen so that the total number of bits delivered during each interval is equal to or greater than the total number of bits delivered by the P-CBR stream during the same interval. More often than not, the adaptive-streaming chunks will end up delivering more bits than the minimum necessary. Those extra bits are the overhead associated with adaptive streaming.

Bit Rate "Overhead" and Adaptive Streaming

The amount of overhead, or excess bit rate, associated with adaptive streaming depends on the number of quantization steps: i.e. the number of possible chunk sizes. More chunk sizes translate into finer precision and less mismatch between the P-CBR bit rates and the adaptive streaming bit rates.

The mismatch is a quantization error which can be analyzed according to standard methods such as those described by [4]. If we make the least presumptive assumption that the value of the quantization error has a uniform probability distribution, then the average quantization error, B_{mqe}, would be equal to one-half of the difference between chunk sizes, ΔB. Thus we may write $B_{mqe} = \Delta B/2$.

Given a maximum operational VBR bit rate of $B_{\max}$, if our adaptive streaming protocol employs a number of uniformly distributed chunk sizes represented by N_{chunks}, then we may write:

$$B_{mqe} = \frac{1}{2}\frac{B_{\max}}{N_{chunks}} = \frac{B_{avg}}{N_{chunks}}$$

Note that we use our assumption of uniform probability to substitute $B_{avg} = B_{\max}/2$ in the above equation, but the choice of probability distributions does not affect our conclusions in a meaningful way for the purposes of this paper.

The average quantization error, B_{mqe}, is the overhead of the limiting case of adaptive streaming. It is the extra bits that would need to be delivered to a client to match or exceed the video quality delivered by a VBR stream. Note that the average quantization error is inversely proportional to the number of chunk sizes.

In the simplest view, if we were to use 10 chunk sizes, we would expect an overhead near 10%. Five chunk sizes would correspond to an expected overhead of 20%. Eight chunk sizes would be equivalent to 12.5%.

Coding Precision and Adaptive Streaming

The simplest view is not, however, the complete view. The overhead in adaptive streaming represents real data that goes towards improving video quality above that of the corresponding theoretical VBR stream. The more sophisticated view of adaptive streaming is that is not only a form of piecewise-constant bit rate, it is also a form of piecewise-constant video quality. Each piecewise-constant interval of adaptive streaming delivers a level of quality that could be matched by a VBR stream having the same average bit rate over the interval..

For H.264/AVC, video quality is largely regulated by a coding-precision parameter known as QP, which takes on positive integer values up to 51 with lower values corresponding to higher video quality. The H.264/AVC standard is designed so that a change in the QP value by 1 will tend to produce an average bit rate change of approximately 12.5% regardless of absolute bit rate.

Thus, in the example of 8 chunk sizes, adaptive streaming would produce an average video quality boost approximately equivalent to a unit change in the average QP value. The general form of the relationship between the overhead associated with adaptive streaming and the increase in effective coding precision ΔQP_{eff} (video quality) may be written as shown below:

$$\Delta QP_{eff} = -\frac{\log\left(1 + N_{chunks}^{-1}\right)}{\log(1.125)}$$

What the above equation indicates is that an adaptive streaming application that employs 5 chunk sizes, for example, and produces exactly the same average bit rate as a VBR application would result in a loss of coding precision of approximately 1.5 QP values. Ten chunk sizes would alter the effective QP value by approximately 0.8 on average, which would not normally be noticed by a typical consumer.

AN EXTENSIBLE QOS ARCHITECTURE FOR A HETEROGENEOUS NETWORK INFRASTRUCTURE TO SUPPORT BUSINESS SERVICES

Srividya Iyer, Dr. Nagesh Nandiraju, Dr. Sebnem Zorlu-Ozer
Motorola – Access Networks

Abstract

Cable operators have been deploying business services for some time now. The nature of these services demand higher bandwidth, better management and complex service level guarantees. In order to meet the demands of the customer, the heterogeneous networks (DOCSIS, Ethernet, PON, Wireless) deployed by the cable operators have to be integrated to work seamlessly to support these services.

The technologies in use for delivering business services are predominately packet based and hence require different Quality of Service (QoS) mechanisms to ensure proper delivery of services. However, the QoS mechanisms defined for these technologies have evolved independently and interoperation in a multi-technology network environment can be difficult. Often the network architects and engineers managing these networks face tremendous challenges in translating the QoS definitions and rules from one network to another, thus making it difficult to provide a seamless QoS experience for the users.

In this paper, we propose an extensible QoS architecture that will support the heterogeneous network infrastructure. Specifically, we will address bandwidth reservation policies, priority queuing and the relationship between the Layer 2 and Layer 3 (IP) QoS mechanisms in a multi-technology network.

INTRODUCTION

Business Services is rapidly becoming the main growth area for cable operators. As guaranteed services have become commonplace for this customer base, it has become necessary for the operators to address the customer's quality of service concerns.
Quality of service is defined as the collective effect of service performances, which determine the degree of satisfaction of a user of the service (ITU-T Rec. E.800).
Quality of Service in general is determined by the network performance for a given service, measured by throughput, delay, jitter and packet loss. Translating a customer requirement into a Network QoS to satisfy the customer's perception of Quality is illustrated in the following diagram [1]

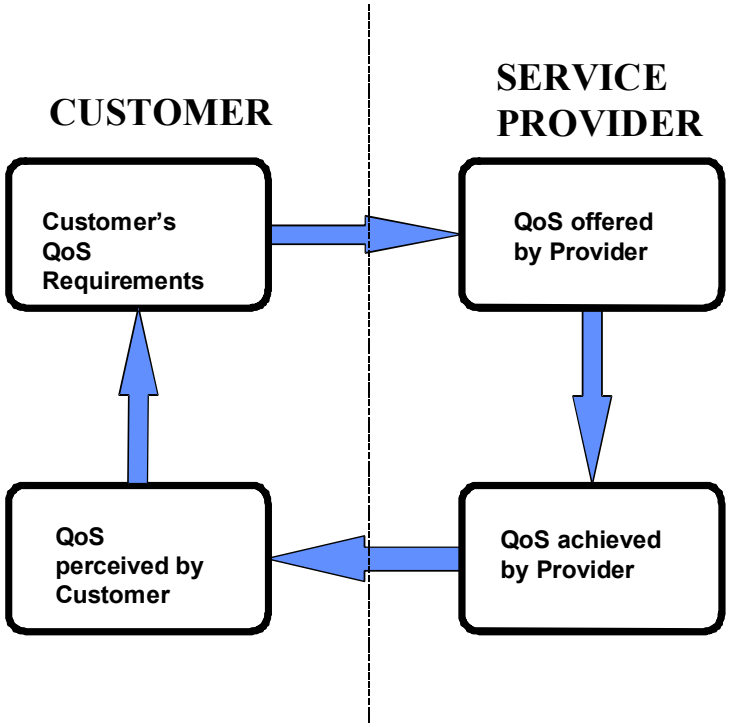


Figure 1-The four viewpoints of QoS

Providing a high level of QoS is possible if the networks are designed uniformly and are under a single management entity. In reality, the cable operator's networks are usually designed taking into account the services

required, cost and distance making the networks heterogeneous in nature. Heterogeneity exists in technology, transmission media, applications, user devices and individual vendor implementations. This heterogeneity makes providing end to end QoS a challenge due to different QoS implementations and lack of communication across these network domains.

The rest of the paper is organized as follows: Section 2 provides a brief overview of the various QoS mechanisms and a mapping between different types of applications.
Section 3 describes the challenges with providing QoS in heterogeneous network architecture. Section 4 describes the solution proposed for providing a consistent QoS across multiple network domains. Section 5 concludes the paper.

SECTION 2 - KEY QOS CONCEPTS

Business services typically consist of varied applications with diverse requirements. Each type of service has a different set of bandwidth requirements and tolerance for latency, jitter and packet loss. The end user requirements, usually defined in subjective terms needs to be translated into Network performance parameters. In order to achieve the necessary performance parameters, QoS mechanisms in packet based networks are implemented either at Layer 2 (MAC) or at Layer 3 (Network or IP). The QoS mechanisms at Layer 2 operate in the individual domains (Ethernet, Wi-Fi, DOCSIS, PON), while the Layer 3 QoS mechanisms provide End to End QoS within the IP domain.

In order to achieve the desired Quality of Service characteristics, some of the QoS mechanisms implemented include Packet classification, Queuing, Bandwidth reservation and Traffic conditioning.

Packet classification – Provides the capability to classify the network traffic into multiple priority levels or classes of service. Packet Classification can be done either at the Layer 2 or Layer 3. Typically a "classifier" is used to classify packets.

Priority Queuing – Usually implemented to avoid congestion in the network by placing the packets in buffers till bandwidth becomes available. Priority queuing uses the various packet classifications and, based on their priority, places them in various queues (high, medium, low etc.) Examples of priority queuing include 802.1p queues and the Diffserv forwarding classes.

Bandwidth Reservation – Usually employed when a network is transporting traffic that requires minimum bandwidth guarantees. In such scenarios, each application will receive a predetermined minimum and maximum bandwidth allocation. The service provider needs to provide the customer the best possible QoS while conserving the network resources. This requires some kind of traffic estimation to prevent either under utilizing the bandwidth or over subscription.

Traffic conditioning – The process of metering, marking, shaping or dropping traffic. Traffic conditioning provides the enforcement of the traffic profiles defined for each of the services.

QoS mechanisms at Layer 2 vary according to the technology implemented. The four commonly used technologies by the cable operators include Ethernet (bridged), DOCSIS, Wi-Fi and PON.

Ethernet uses the 802.1p priority bits to classify the traffic into various classes. The

priority queuing uses these traffic classifications to place different classes of service into different queues.

DOCSIS uses a concept of service flows to classify packets into a unidirectional flow which is then shaped, policed and prioritized according to the QoS parameter definitions. Multiple service flows can exist for each Cable Modem and grouping of service flow properties is also facilitated through the definition of a service class.

The QoS mechanisms defined for 802.11 networks include an additional coordination function called HCF (Hybrid Coordination Function) on top of DCF (Distributed Coordination Function) and PCF (Point Coordination Function). The HCF enhances QoS provisioning during both contention and contention free periods by using EDCA (Enhanced Distributed Channel Access) and HCCA (HCF Controlled Channel Access) respectively. Due to higher complexity and complications caused by overlapping stations and unlicensed spectrum use, HCCA is not an industry choice. The Wi-Fi alliance WMM program is based on the EDCA method and was initiated as the market needed a QoS solution before 802.11e was ratified. WMM defines four access categories (voice, video, best effort, and background) that are used to prioritize traffic to provide enhanced multimedia support. EDCA defines four AC (Access Categories) to differentiate different services. This requires mapping from UP (User Priority) to AC. 802.11e recommends UP to AC mapping, however current approved standards do not provide a uniform implementation for vertical (between Layer 2 and 3) and horizontal (between various network domains at Layer 2) QoS mappings. The ongoing 802.11u standardization effort aims to standardize the "information transfer from external networks using QoS mapping".

In Passive Optical Networks (PONs), Gigabit-PON (ITU standard G.984) introduces the concept of Traffic Containers (TCONTs) to classify the traffic and service them according to the QoS definitions. In Ethernet PON (IEEE 802.3ah/av) multiple Link Layer Identifier(LLID) can be used to classify the traffic into different classes and service them according to the negotiated QoS definitions.

QoS architecture frameworks at Layer 3 consists of either a differentiated services model (DiffServ IETF RFC 2474/2475), where the traffic is classified and the frames are treated with different priority based on information carried in the frame header, or a reservation model (IntServ IETF RFC 1633), where a signaling is used per session to reserve resources.

SECTION 3 - CHALENGES OF IMPLEMENTING QOS IN A HETEROGENEOUS NETWORK

There are several challenges facing the service provider in providing a high level of QoS across heterogeneous network architecture.

First and foremost, enabling QoS has been focused on mechanisms and protocols in individual network domains (eg wireless or cable access) or even individual network elements. This provides a high level of QoS within that domain or network element, but lack of communication between the various domains makes it harder for the QoS implementations to cross boundaries. That leaves the network architects and engineers managing the networks with the task of implementing the disparate QoS policies within each of the devices in each of the

networks. In addition, the devices provide a multitude of options to the network architect. These options might not be the same across networks or even across different vendor devices within a network. This makes the QoS implementation error prone and inconsistent.

Almost all the QoS mechanisms that exist are defined statically to enable QoS and do not automatically adjust to the traffic conditions that exist in the network. To solve this issue, the network architects routinely over provision the network leading to wasted bandwidth. Conversely, where bandwidth is at a premium, the network is oversubscribed leading to lower priority services being denied access. Also, it is to be noted that over provisioning the network doesn't always guarantee required QoS to the end user as evidenced by the peer to peer applications consuming a large portion of the bandwidth and causing degradation of other services.

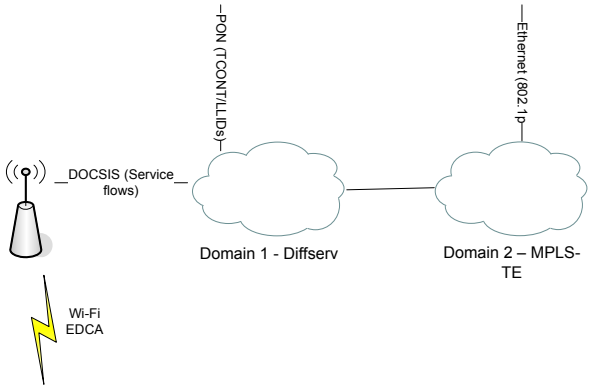


Figure 2 - An example of QoS mechanisms in a Heterogeneous Network

Several standard bodies have defined service classes and their packet classification schemes. However, mapping these service classes to the individual QoS mechanisms in each of the network domains still remains an issue for most service providers. Since there is no fixed number of service classes supported in each domain (Ethernet 802.1p supports up to 8 service classes and Wi-Fi APs support four Access Classes) and no enforced mapping between the service classes and the QoS mechanisms, it is up to the person defining these QoS mappings to select the best possible one, thus leading to inconsistent behavior across the network.

Typically QoS can be provided at the Layer 2 or the Layer 3 level. The current implementations usually mix Layer 2 and Layer 3 QoS mechanisms (For example, a device can implement the 802.1p CoS with a Diffserv AF PHB). The traffic class markings between packets can vary depending on the originating or forwarding device and can be at Layer 2 or 3 or both. The device would then use the Layer 2or Layer 3 class markings depending on which one is available or based on network configuration (trust Layer 2 or 3) . The challenge in this case is to provide a consistent mapping between the Layer 2 and Layer 3 mechanisms. As indicated earlier, they evolved independently and thus there is no standard way of mapping the Layer 2 to the Layer 3 QoS, leaving it to individual implementations.

SECTION 4 - AN EXTENSIBLE QOS IMPLEMENTATION

Several consortiums and research bodies are in the process of defining end to end QoS architectures [2]. These architectures define a method to provide end to end QoS and might benefit certain new network builds. They however require an exhaustive rework and re-architecture to function within a large existing network infrastructure.

Implementing a network-wide end to end QoS architecture is time consuming, requires a lot of resources and is disruptive to the existing operations. In order to address the challenges outlined in the previous section, we are proposing an extensible QoS architecture that would use the existing QoS

mechanisms in each of the domains optimally and with minimum manual configuration, while providing a seamless QoS experience to the user.

Service classification and Traffic category mapping

In any network that provides more than a Best Effort service, the first step for the network architect is to classify the types of services that will be provided over the network. Having a very granular service classification provides the best possible QoS for each class, however in a multi-technology network, this becomes cumbersome and scalability becomes an issue.

The first step is to classify the services into broad categories that share similar requirements. IETF and 3GPP both define four service classes and can be used as a reference, or the service providers can define their own set of classes. The key is to define service classes and the corresponding traffic category mapping that enable a consistent behavior across multiple domains. The following table illustrates a subset of service classes and the recommended (or commonly implemented) Layer 2/Layer 3 QoS schemes available across a sampling of the different network architectures.

Traffic Category	**Ethernet, Wi-Fi DOCSIS,PON**	**Diffserv Layer 3**
Real Time Traffic (Voice)	P bit 6, 7 AC3 UGS,TCONT1	EF
Interactive (Video)	P bit 5 AC2 RTPS,TCONT2	AF
High Availability (Data)	P bit 3, 4 AC1 nRTPS,TCONT3	AF
Best Effort (Data)	P bit 0,1, 2 AC0 BE,TCONT4	Best Effort

Table 1 – Traffic category/QoS mapping

Layer 2 to Layer 3 mapping

With the advent of heterogeneous networks and Layer 2 Services (in order to avoid expensive Layer 3 equipment), several Layer 2 networks will be interconnected between the Layer 3 core networks. This introduces interoperability issues between the Layer 2 and Layer 3 QoS mechanisms.

Although the IETF states that the Diffserv model can be extended to support Layer 2 architecture, there is no defined behavior or mapping between the Layer 2 and Layer 3 QoS mechanisms. In our model, we propose implementing a consistent set of traffic category mapping with the Layer 2 QoS mechanisms available and updating this information across all the network elements. This is in order to avoid inconsistencies that arise from the manual configuration of QoS profiles in each network element. A specific implementation of this mapping is hard to define as there are variations within the network domains and network elements.

Queuing

One of the key service differentiators used in supporting multiple classes of service is priority queuing. The priority queues will usually be implemented separately for both ingress and egress traffic. However priority queuing does not provide bandwidth guarantees and deterministic latency, jitter and packet loss characteristics. In addition, each network element supports two or more types of priority queuing algorithms and different numbers of priority queues thus

causing inconsistent treatment of service classes across various network elements. That being said, priority queuing is widely used and an integral part of providing service differentiation. Two key elements in using the priority queuing efficiently are mapping the service classes to the right queue and selecting the right algorithm. Neither the Strict Priority (SP) nor the Weighted Fair Queuing (WFQ) /Weighted Round Robin (WRR) provide the fairness required for all traffic conditions. In the Strict Priority queue, the highest priority traffic always gets through and, depending on the traffic load, the lower priority queues can be starved for bandwidth. In the WFQ/WRR scenario, the weights that are assigned are static and can either not provide the required QoS or waste bandwidth depending on how many different service classes are accessing the network at a given time.

In our QoS architecture, we propose a scheme where the type of queuing employed at the ingress and egress and the weight assigned to each of the queues is based on the traffic received on each of the queues over a period of time. The time interval for sampling of the traffic can be predetermined by the network element or configured. This eliminates the problem where the high priority traffic always gains unfair advantage, while the low priority queues are denied access. This also eliminates the wasted bandwidth with statically configured queues. In addition, the minimum and maximum rate of each queue will be adjusted dynamically based on the traffic conditions.

Bandwidth reservation

In most services that require high Quality of Service, bandwidth reservation is one of the key factors affecting the QoS and additionally has an effect on the delay, jitter and packet loss characteristics.

IETF proposed IntServ with RSVP to reserve bandwidth for each flow end to end. This approach, while providing the optimal bandwidth for each flow, is not scalable and hence not supported in most network elements. On the other hand, due to its simplicity and scalability, DiffServ is more widely supported. Diffserv only provides bandwidth guarantees for aggregate flows and defines a set of Per Hop Behaviors (PHBs).

Most of the network elements support some level of DiffServ functionality. For example, A DiffServ compliant (DS) node that supports the four Assured Forwarding (AF) traffic classes must allocate a configurable minimum amount of forwarding resources (buffers and bandwidth) to each AF class. The minimum bandwidth is pre-allocated for each AF class and is equal to its corresponding Committed Information Rate (CIR). This pre-allocation of bandwidth has to be done on all the nodes supporting the AF classes. In addition, the bandwidth availability and traffic patterns in each of the node vary. Pre-allocating the CIR bandwidth for each AF class might be too much or too little.
In order to overcome the challenge of pre-allocating bandwidth at every node for each of the AF classes, we propose a dynamic bandwidth allocation scheme at each node based on the service classes defined. This is done as a two step process. Each node maintains a count of aggregate traffic received for each AF class and the Best Effort services. In addition, each node supporting a DS function will query the adjacent node for bandwidth usage to support dynamic bandwidth allocation and avoid congestion.

Communication between multiple network domains

With the evolution of heterogeneous networks, there is a need for multiple network domains to interoperate and provide seamless

QoS experience to the end user. Two of the main issues with interoperability of QoS between these domains is the lack of consistent QoS implementations and communication between different entities. The QoS implementations have evolved independently and it would not be practical to change the existing mechanisms. Providing end to end signaling across these networks is also not feasible due to scalability issues. In order for services traversing these domains to achieve high level of QoS, we propose a communication path between the nodes residing at the edge of each domain. These edge nodes would ensure that the QoS required for a service traversing that domain can be satisfied by the resources available.

SECTION 5 - CONCLUSIONS

In this paper, we propose a QoS implementation scheme that utilizes the existing Layer 2 (MAC) and Layer 3 (IP) QoS mechanisms, while providing a consistent QoS to the end user across multiple network domains. This approach also minimizes the number of static QoS profile configurations that are needed in each network element. The key to achieving QoS interoperability across multiple network domains is to recognize and reconcile the different QoS implementations that exist today. There is also a need to focus on the end user service requirements while optimizing the network resource utilization.

REFERENCES

[1] ITU G.1000 Standard, Communications quality of service: A framework and definitions, 2001.

[2] B. Iancu, V. Dadarlat, A. Peculea, “End-to-End QoS Frameworks for Heterogeneous Networks - A Survey”.

CALIBRATING THE CRYSTAL BALL FOR THE NEXT DECADE OF GROWTH

Dr. Robert L Howald
Motorola Home & Networks

Abstract

Today's HFC plants continue to be a powerful infrastructure for delivery of video, voice, and data services to residential and business customers. It has successfully evolved over time to support a broader suite of services, and these services continue to be enhanced. Where will it go next? While predicting what comes next is always risky, the alternative – moving ahead without a vision – is riskier still. A decade ago, Motorola embarked on a bandwidth projection analysis in order that the industry could prepare for a future full of new possibilities. That projection was published as part of the NCTA proceedings in 2002.

Now, here we are in 2010, with nearly the ten years of "the future" behind us. What predictions were accurate? What misfired? What factors contributed to divergence in estimated growth? This paper will assess those projections. The illuminating conclusions provide important lessons learned about subscriber behaviors, the pace of technology maturity, and how new services come to market. We can use such lessons to better project the next wave of services and technology. Such knowledge is critical to making the next ten years a success.

INTRODUCTION

Predicting technology over a 10-year horizon can be risky, at best. Nonetheless, it is valuable to do so, not only to identify potential game-breakers to fundamental assumptions, but also as part of continual assessments required to keep in touch with the changing dynamics of the industry. Rapid change, relatively speaking, was the name of the game in 2001 when this initial assessment began. A rapid, accelerating pace of change is part of the dynamic as well today as we assess the cable business. The changes in play 10 years ago were primarily driven by early action around new services (primarily DTV and data) and the maturation of key enabling technologies. Today's changes have very similar elements, but with the powerful new variable of wireline marketplace competition brought on by telco triple-play providers. This represents an overriding force of change unaccounted for in any significant way 10 years ago, and that undoubtedly has had an effect on the services evolution MSOs have undertaken. Another important element of the MSO picture today is a renewed focus on commercial services outside of the small business (best-effort DOCSIS) realm. In the assessment done 10 years ago, commercial services were deliberately *not* considered, in order that the analysis could focus on residential services. However, because of increasing service rates, deeper fiber runs, PON, and WDM technologies, synergies have been created between the residential and fiber architectures. The effect could impact residential network evolution through the "pull" of new fiber technologies.

The pages that follow will be organized quite simply:

1) Walk through the predictions that were made in each service area 10 years ago, including referencing the words from the original paper [1] directly.

"All reference [1] statements will be in italics and quotes."

2) Compare that prediction to today's reality, considering explanations and lessons learned from the disparities
3) Quantify and visualize the outcomes of predicted versus "actual" on a before-and-after spectral map
4) Identify areas that completely missed the mark
5) Discuss potential drivers for the next decade of growth

Note that the reference paper for the 10-year analysis [1] was published in the NCTA proceedings in May, 2002. Note also, however, that much of the MSO data gathered to support a 2002 publication was obtained in 2001. As such, it represents EOY 2000 to 2001 data. We describe these steps in order to clarify how a paper in the middle of 2010 captures a "decade" of growth. Arguably, one could consider it 9+ year comparison.

Now, let's get on with the post-mortem!

SERVICE-BY-SERVICE ASSESSMENT

Analog Broadcast

"Analog broadcast service is projected to remain largely unchanged over the next 10 years"

It is difficult to assess this singular statement very critically, as it is certainly true in many cases and most MSOs at this instant. As such, this prediction has objectively turned out to be quite accurate. However, looked at in a fuller context, there is not much in [1] that recognizes the trajectory of some MSOs moving away from analog carriage, and some, such as Comcast, in a very aggressive way. Most have some amount of reclamation under consideration as a minimum, with the question mostly about when. The following statement also appears:

"In a typical network, 14 analog channels are expected to migrate to digital, reducing the analog spectrum required from about 500 MHz to 400 MHz by year 2006"

Thus, while there is a recognition that the principle of reclamation would come into play, [1] does not foresee the congestion-based momentum to exchange analog bandwidth for digital. As we will see, this is mostly due to a significant underestimation of the growth in the area of High-Definition Television (HD).

Digital Video Broadcast

First, some context of "where we were" when we think about the starting point time frame for the 10-year projection:

"Digital video cable is currently in the mass adoption phase. By the end of year 2001 approximately 18 million digital cable set-top boxes were in use by US subscribers. Typical systems offer 10 QAM carriers."

It is impressive to note how far cable has come in DTV since this study! The study noted 10-12 QAM carriers was typical in launching into this prediction for where it was headed in 10 years:

"A net gain of 36 additional programs is expected over the next 10 years. Four additional QAM carriers will be added to cable plants, bringing the total number carrying Standard Definition (SD) content (including migration from analog broadcast) to 16 by year 2011."

This prediction for broadcast SD is off by roughly a factor of two (low), which seems unusual given that at this point in the DTV

transition, we were still on the "hype" curve. Note also, however, that being off by a factor of two over a 10-year period means that the growth was underestimated by less than 10% per year on a compounding basis. An inkling into prevailing thoughts observed at the time and driving the underestimation were two subsequent statements:

"With VOD services emerging ..... and the cable modems competing for consumers' free time, it is hard to see a case for the addition of many new broadcast channels"

VOD was at an even earlier point on the "hype" curve, enough so that it was already bleeding attention away from the "mature" digital TV technology. We will evaluate the VOD piece in a later section. We now know that the addition of HSD to the service portfolio has done little to divert attention away from TV viewing hours. In fact, in the last few years, it appears that the capability of the PC medium to support video has in fact delivered *more* hours to the big screen through the association and loyalties built with broadcast programming through the PC.

Another statement made that caused second thoughts on continued broadcast SD growth:

"Beyond our 10-year period, 2-way interactive broadcast content could be the salvation of broadcast services in a world that is otherwise evolving to total content-on-demand"

The latter part of this prediction is prescient, and will be discussed later. However, on the initial postulate, there was a sense that once interactivity arrived (envisioned as the OCAP effort taking shape), the nature of the viewing experience would change in a way that negatively impacted the pure broadcast experience. This would be due to subscribers finding the interactive channels more compelling, stifling the growth of "POBS" – plain old broadcast services. In reality, while interactivity exists, it is still struggling to find its place in the way envisioned – a way pretty much everyone envisioned at that time. Perhaps also contributing was the general misunderstanding of just how much some of us might enjoy simply being couch potatoes! More seriously, part of the interactive aspect may have been connected to demographic changes, and the anticipation of the emerging behavior patterns of "connected youth." Consider the following statement:

"Consumer interest in interactive TV exists as evidenced by a growing number of consumers interacting with TV programs using PCs"

As we now know, demographic patterns have manifested themselves in multi-media experiences not necessarily onto TVs. Instead, multi-media has proliferated onto the myriad of other devices that advancing technology made possible, and where convenience has trumped performance quality, as similarly seen in the cell phone voice example. The PC, rather than merely an outlet for TV-viewing interactivity, is instead (or in addition to) a popular screen of over-the-top content – a trend noted 10 years ago, but underestimated in speed and magnitude. Its role in MSO-owned and managed content is being defined by many operators at this point.

Finally, the calculation of interactivity did not foresee the desire for the "lean-back" viewing experience likely increasing with HD penetration, which has grown as large screen TVs became affordable.
Consider once again the statement:

"Beyond our 10-year period, 2-way interactive broadcast content could be the

salvation of broadcast services in a world that is otherwise evolving to total content-on-demand"

The perceptive recognition of a "total content on demand" world in this early digital era was indicative of the anticipation of how VOD would transform the industry. While this means of unicast video distribution may not be the core technology around which "everything-on-demand" takes place, there was a general sense of engineering future systems for an increasing range of multi-cast and unicast delivery. Total content on demand is now more broadly captured by the industry mantra of serving the "four any's" – any content, anywhere, anytime, any device.

HDTV Broadcast

"Is the picture quality worth the price of an HDTV? How many consumers viewing a 42 inch screen at normal distances can discern the improvement in HDTV quality relative to DVD or MPEG 2 SD quality?"

Probably the most significant underestimate in terms of bandwidth repercussions was in the area of HD viewing. However, objectively, the error is not actually so large from the perspective of what is in the field in many places today, but more so in the context of where trends are headed this year and next. The above statement alludes to some of the perceived barriers to scaling HD – the high price of HDTV's at the time, and the associated value proposition for normal viewing. As expected, and predicted in the paper, the price for mass adoption did get to a tipping point in the 10-year time frame – relatively recently, in fact. The original analysis did not foresee that with HD would come a new class of display technology and an overall increase in "normal" size screens to enhance the value proposition of HD – the concern alluded to in the above statement which referred to what was essentially the largest "tube" sizes of the day of 42 inches. Associated with this is the increase in sports viewing (NFL Network, Golf Channel, MLB Network, ESPN19 anyone?) for which HD is fuel for the fire.

Finally, a key missing factor was the forces of external competition – specifically satellite broadcasters, who, without a VOD play, found a powerful, profitable refuge in racing to the front of the HD competition.

"With two HDTV programs per carrier, systems will begin carrying 4 to 6 HDTV video programs this year. By year 2011, 16 HDTV channels are forecast. The bulk of content will continue to be delivered in SD resolution."

The underestimation of HD content, because it represents the largest Mbps volume of all current services, adds up to the largest error in the bandwidth maps we will show later. However, note that we are in a period where HD is a relatively rapidly moving target – most MSOs are looking to add HD content in a big way, with physical bandwidth in the way in the near term. Bandwidth constraints revolve around both the Mbps associated with HD, but also with the need to simulcast alongside the SD version.

However, again, while most MSOs expect rapid addition of HD programming in the very near term, the prediction is actually not very far off at this exact moment for many systems. The above statement predicts 8 QAMs of HD, using a two-HD-programs-per-QAM relationship, where 2-3 is normal and content dependent today. As such, the prediction was for 32 HD programs by 2011. In fact, by my count, my home cable system, in a middle tier (top 30) metro area overbuilt by VDSL triple-play services, has between

35-40 HD programs depending on the channel lineup version I review.

Nonetheless, large MSOs today are looking to be HD-competitive beyond the VOD-based HD library often used to boost advertising campaigns. This generally means campaigns to achieve 100-program line-ups of HD, or greater, or approximately 40 QAMs worth (240 MHz) – about one-third of the bandwidth on a 750 MHz plant.

While the HD prediction may have been in fact quite accurate in the purely numerical context of systems today, the underestimating of the trend of accelerated HD deployment leads to an area that was completely missed – implementation of switched digital video (SDV) technologies in cable. Long the "only" realistic solution for telco video delivery via the overmatched xDSL wires, cable has seen this matured technology as another bandwidth tool in the toolbox – such as to increase an HD line-up without having to physically support the complete spectrum required to do so. Exploiting both the shrinking serving group sizes and natural statistical gains of popular, multicast content, bandwidth gains (and thus program count gains) of 2-3x of "virtual" bandwidth can typically be added.

MSOs are at different stages of SDV deployment. Allocations of 4, 8, or 16 SDV QAMs is roughly par for where competitive systems that have deployed the technology will likely be this year, moving towards 20-24 where aggressively underway already. It is well-documented that TWC has been the most aggressive of the large MSOs in North America deploying SDV.

Note that MPEG-4 gains were not factored in as elements of the 2002 analysis – nor has there been significant MSO activity moving towards MPEG-4 based HD for traditional MPEG-2 TS QAM delivery.

Video-on-Demand (VOD)

Again, as a point of reference context from [1]:

"At the start of Year 2002, operators had launched or planned to launch VOD (commercially or in trials) in almost 90 markets"

Thus, VOD service was very much a newly emerging service, and with that emergence, there was much hype in what it could become, and the impacts it might have on the fundamental broadcast-oriented nature of cable video delivery. Nonetheless, and in spite of the hype of the period, the analysis relied on key numbers in ultimately predicting conservative growth. An important factor was noting that VOD was necessarily tied to digital penetration, and there was enough available trend data and market research at that time to have some "expert" opinions rendered on that trajectory:

"Kagan's 2001 annual growth forecast shows digital cable penetration growing to 63% by year 2011"

The above analysis was actually relative to homes passed, so error can be attributed to properly capturing the actual MSO service penetration multiplier in today's more competitive world. It turns out to be a good representation as a function of MSO cable subscribers, however, and so reasonably captures the digital growth trajectory in the context of the cable customer base.

Also, the analysis relied on realistic models and research in viewership behaviors of like-content to that for which VOD would naturally support. To a first order, this

would be the "Blockbuster/West Coast Video" replacement model, followed secondarily perhaps by likely very popular TV shows – those that inspire lifetime loyalty and repeat watching. These would be shows with the popularity of, say, M*A*S*H, Seinfeld, Friends, etc. Of course, the "West Coast Video" model did indeed have a noticeable impact on … West Coast Video!

Using peak-time viewership behaviors while concluding that availability of VOD does not lead to major changes in basic viewing behaviors of that type, the following guidelines were used to model its growth and impact:

"The estimate of simultaneous use during peak hours is 5% today and forecast to increase to 9 % by year 2011"

The 9% value is in line with ranges used today in system engineering of VOD. Peak hour concurrency factors from 5% to 15% are unofficial but observed values used by architects today. The net result of this penetration and peak-time concurrency is 3-6 VOD QAMs predicted by the analysis from 2002. The analysis assumes that node sizes (quite accurately) will be in the 500 hp range, justifying three QAMs. It does not further consider service group splitting of video and data groups, which today may result in sharing of VODs across nodes. Doing so would double the QAM count for the same traffic engineering parameters, resulting in 6 VOD QAMs. And, in fact, 4-8 VOD QAMs is a reasonable count on today's systems, with some MSOs expecting to increase this to perhaps 12-16 in the coming years.

VOD trajectories may grow slowly moving forward. A significant factor in where VOD heads, and recognized in the 2002 analysis, is the impact of IPTV on VOD. VOD, as a unicast video delivery mechanism, represents a natural service type to permit smooth migration to IP delivery from a technology standpoint – opposed of course by legacy infrastructure and HSD architectures built for data. Nonetheless, a prophetic statement from 10 years ago was:

"VOD can be streamed over the Internet using IP and DOCSIS……at these rates, audio and video quality is competitive to that offered over MPEG 2 multi-program transport streams to set-top boxes"

Of course, we now know that the video and HSD service rate relationships reached a watershed moment, shown in Figure 1. There was a separate section in the 2002 paper entitled "VOD over IP," which fits best as a discussion topic in this section of our analysis today. It is hard to have a discussion about video service trajectories and service expectations without devoting some time to video over IP. The impact brought about by the crossing trajectories in Figure 1 brought some inevitability to cable's video evolution path. So, let's dig into this idea of video over the HFC IP pipe, which today is implemented by DOCSIS.

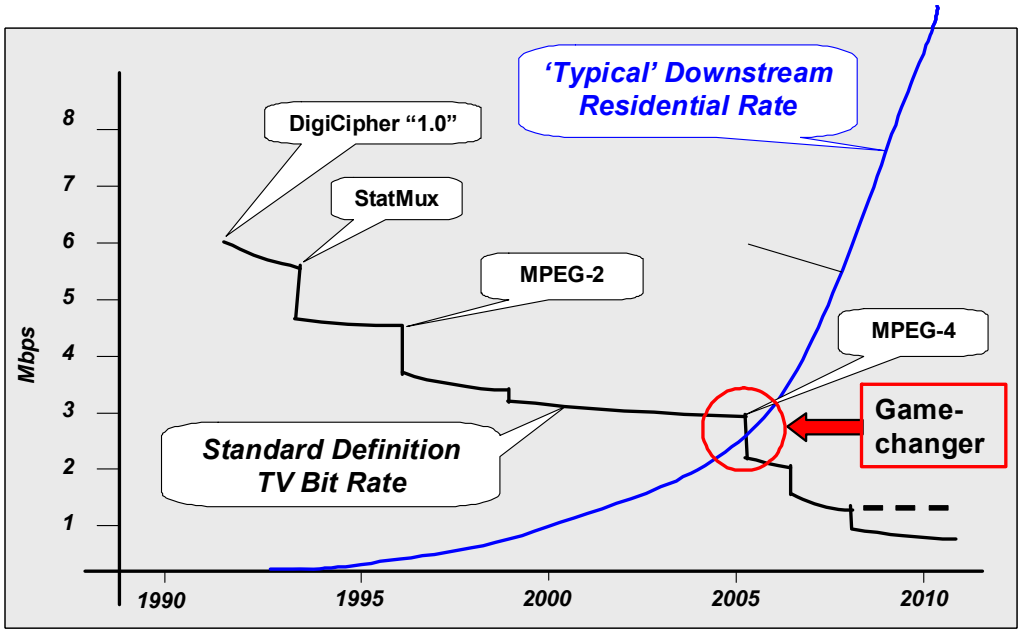


Figure 1 - HSD Access Speeds vs. Video Rates

Note from Figure 1 that the introduction of MPEG-4 encoding (H.264 AVC) brought with it another 50% of average video rate bandwidth efficiency – though this 50% value can be quite content dependent [2]. As a new technology, it would of course involve major infrastructure changes. As such, it naturally conjures up thoughts on how best to introduce it, and suggests a new opportunity to remake and improve future delivery systems, such as with IP delivery.

In addition to MPEG-4, another potential bandwidth efficiency exists, which was not considered when initially pondering IP delivery in [1]. This is the benefits of variable bit rate (VBR) delivery enabled by IP packet scheduling mechanisms. HSD schedulers are designed to efficiently make use of capacity when input data is of varying packet sizes and arrival rates. Coupled with more streams per carrier (MPEG-4 helps with the law of large numbers) and the introduction of DOCSIS 3.0 channel bonding, VBR promises additional bandwidth efficiencies over traditional CBR delivery – likely in the range of 20-40%.

This combination of bandwidth efficiencies is doing two things:

1) Providing a reason to hold off on what might otherwise be continued growth of MPEG-TS based narrowcast delivery for on-demand
2) Making reasonable the idea of "simulcast" of MSO channel line-ups (or portions thereof) for delivery of essentially the same video line-up over IP

Consider that a logical target of IPTV for MSOs initially may be PC screens with modest (VGA-like) resolutions. In this case, the bandwidth numbers come together quite nicely:

- Today's Broadcast Line-Up (MHz) = [FULL] MHz

- IP Simulcast Line-up (MHz) = [FULL][50%(MPEG-4)][75%(VBR)][50%(VGA)]= 18.75% x [FULL] MHz

Thus, an initial IP video offering for the PC could require just one-fifth of the full-service bandwidth being made available. Put

another way, 300 Broadcast SD channels over 30 QAMs could be handled instead by 6 QAMs, or 36 MHz of spectrum…pretty powerful stuff. Of course, this must work its way up to 12 QAMs (72 MHz) to bring SD to the PC, but this could perhaps occur in time as the IP service is gradually rolled out by replacing less-efficient MPEG-2 VOD carriers.

So, as we can see, there are new and compelling reasons to consider the unicast "everything on demand" future as one derived from IP delivery even from an access network standpoint – with the access network being the last piece of the HE-to-end device architecture not already IP-centric.

Note also the issue that struck a chord 10 years ago is still under scrutiny today in envisioning this IP evolution in full:

"……but end-to-end QoS mechanisms are required to support continuous data rates in the range of 2 Mbps to 4 Mbps"

CMTS platforms with prioritization capabilities, and a myriad of IP QoS techniques have been developing in these last 10 years, but the concept of "guaranteed" services for IP services per the Intserv model has given way to simpler and more scalable means to offer statistical guarantees. However, with video services being intolerant to packet drops, and with constraints around jitter, these assurances are still to be proven out in a way that does not require severe underutilization of pipe capacity [2][3].

Another perceptive statement:

"HDTV VOD might be an interesting proposition. Early HDTV adopters are good candidates for higher priced VOD content."

While we have moved beyond "early HD adopters," it is certainly the case that service providers (and over-the-top-providers) have recognized that HD content can be charged at a premium – not because "early adopters" represent a wealthier segment of the population, but because the value proposition of HD video is that strong.

Finally, a still relevant postulate from 2002, mired in some regulatory obstacles:

"VOD has much more potential than just replacing the video store. MSOs could offer server based Personal Video Recording (PVR) capability."

This one still remains to be seen, given the potentially significant implications to CPE, storage, content delivery networks, and access bandwidth.

Internet Access

Downstream

This basic user expectation, which was at the origin of placing cable data services at the forefront, has not changed:

"Users' key expectations are low latency in delivery of web pages and downloads, rapid updates in games and seamless delivery of streaming content. They also expect "always on" service."

Of course, what qualifies as "low latency" and "seamless" has changed, as the bar has been raised in both cases. Additionally, "always on" today really, really means ALWAYS, as opposed to almost always. The most powerful example of this reality is the routine use of remote offices and workforces despite the increase in the quantity of information that is exchanged daily.

Once again, to put into perspective "where we were" at the start, consider this statement from [1]:

"It is estimated that YE 2001 a provisioned gross average bit rate of 21 kbps per subscriber was needed to achieve subscriber satisfaction"

With traffic engineering principles applied in the original analysis, this gross average worked out to a peak service rate of about 1 Mbps. The state of spectrum at this stage was one downstream DOCSIS carrier deployed, serving multiple nodes, which served a larger number of homes passed than today, creating serving groups in the multiples of thousands.

Let's compare where downstream data is today and where it is headed in the short term. Like HD, HSD is, relatively speaking, a moving target. Most MSOs are executing on plans now to add downstream DOCSIS carries, in some cases to add them and to bond DOCSIS 3.0 channels. They are simultaneously working on shrinking service groups on a market-by-market basis based on competitive need, engineering for more bps/sub to keep ahead of the growth trajectories. As such, downstream DOCSIS spectrum, which is moving towards no longer being shared across nodes if not the case already, can be considered somewhere between 2-4 channels today, on average, with plans to increase to 6-8 shortly thereafter where spectrum is available or can be cleared. Some MSOs are looking more aggressively still to execute on the transition to an all-IP architecture, in which case DOCSIS carriers would take on a larger role and consume yet more spectrum than 8 slots more quickly. The trend towards all-IP is undeniable, but the speed at which that can occur for video is hindered by several key legacy factors. Thus, for the purposes of DOCSIS QAM count, to assess the prediction in [1], ranging up to 8 in the 2010-11 near term is a reasonable high end. However, recognize that the accelerated pace of 2 to 4 to 8 would be expected to continue to take place in 2011-12 and 2012-13 in more aggressive transitions, which is an important immediate consideration for planners.

Looking at the numbers, then, it was recognized even early in the HSD business that trying to guess the next big application was less likely to capture the growth requirements than a compounded growth rate. In other words, over 10 years, it is smarter to base projections on the smoothed curve of growth as opposed to the more realistic series of step functions underlying the average growth trajectory, which led to the following long range projection:

"The 10-year forecast, therefore, is for consumption to grow at 50% per year"

The assumption of historical growth rates continuing has generally come true, although the actual growth as calculated in terms of 4 DOCSIS carriers served over today's node sizes turns out to be closer to 40% compounded on average. While this reflects a pretty good prediction, over 10 years this means being off by a factor of two. Using this growth premise, the analysis in [1] concluded that 6-8 DOCSIS carriers would be required to satisfy demand. That is, in [1], it was anticipated to be a moving target as well, and in so doing anticipated 6 DOCSIS carriers for 2010 and 8 for 2011. This is quite an accurately painted picture, given that we are looking at 2-4 at the moment, and moving to 4-8 in the near term. The difference between 6 versus 4 can be traced to the difference noted above in the impact of being off by 10% in the compounding rate for 10 years. Nonetheless, this is a pretty perceptive projection.

With respect to peak rates – an issue coming to the forefront as we introduce DOCSIS 3.0 channel bonding – and applying traffic engineering parameters predicted, it was anticipated that the peak service rate to the consumer would be approximately 10 Mbps. This is indeed in the ballpark of where downstream speeds offered by large MSOs in competitive environments, where tiers in the 5-20 Mbps represent high end downstream services.

Overall, DOCSIS downstream growth was quite accurately predicted, in particular given the limited amount of cable legacy to draw upon for HSD services.

Upstream

"Average upstream consumption increases from 7 kbps to 700 kbps"

This represents about 59% if calculated as compounded growth. However, the figure calculated in [1] to represent growing traffic is actually an estimate based on *downstream* compounding (actually further broken into compounded volume@25% and compounded concurrency @20%) multiplied by a factor representing the traffic mix trending towards more symmetry. The compounded growth tied to symmetry is described as follows:

"Upstream bandwidth increases more than downstream due to the expectation that rate asymmetry (the ratio of downstream to upstream rates) will decrease from 6:1 to 3.5:1"

We would likely not take this approach today, having noticed that this symmetry trend has actually reversed itself with the introduction of video clips as core drivers of HSD bit volume. On average, however, and recognizing that the above statement is on a per-sub basis (penetration accounted for as in [1]), this growth rate to 700 kbps overstates average rates today. However, as with downstream, and perhaps more so than downstream, adding upstream immediately is a high priority for MSOs today. It is a recognized potential bottleneck.

For the 700 kbps to closely represent the situation would mean getting the 4 upstream carriers that some MSOs are looking to turn on going, all at 64-QAM. Alternatively, it comes close also by considering a next stage reduction of average service group size shrinking, creating new virtual bandwidth – i.e. more bps/sub by reducing subs. With neither of these two elements quite in place, and not likely to happen until next year at best, this estimate appears off by about a year or so, which is not bad. This is not surprising when recognizing that aggressive compounded growth rate used to generate it, and noticing that the symmetry trend in fact did not continue as anticipated. All in all, the estimate represents a reasonable prediction of where things headed for upstream. They are certainly not there yet, but a noted objective in many camps is to do exactly what it would take to get to this range in the near term. And, it is preferred in any case to slightly miss on the high side than the low side.

"As for upstream…..the peak rate is expected to increase from 200 kbps to about 3.2 Mbps in year 2011"

This is a very high-end upstream service tier range, but nonetheless in the field of play of today's offerings.
All in all, then, the upstream predictions have turned out to be quite solid.

Streaming High Quality IP Audio and Video

We covered much about the implications of streaming video in the VOD section when discussing IP video. The context of the 2002

paper recognized over-the-top video services as a bandwidth driver in calculating the HSD growth, and identified the means by which this occurred and would continue to occur:

"PC based multimedia decoders (Windows Media Player, RealPlayer, QuickTime, and ultimately MPEG4) are widely used to deliver low resolution, low rate VOD over "Best Effort" Internet access service"

Note that, because of the obvious value to the IP world for cable, MPEG-4 encoding was anticipated.

This over-the-top video was distinguished in [1] from the "high-quality IP video" which we would, today, probably think of as delivery of managed MSO content on the IP pipe. This was not necessarily broken down into these segments in 2002, but given the overall novelty of the concept at that time, it is difficult to assess that added detail critically. Nonetheless, this farsighted statement was made in 2002:

"Mass-market penetration of streaming will likely wait until solutions are in place to move the content into the entertainment center and other places within the home. Lacking solid QoS guarantees, entertainment quality video and audio cannot be delivered reliably enough to satisfy paying consumers."

Indeed, precisely at this time we are seeing intense activity with MSOs and at CableLabs around MoCA, DLNA, and UPnP to ensure PHY throughput and QoS delivery around IP-enabled homes for multi-media content distribution and delivery. MSOs have suggested, modified, re-visited, and updated various approaches to developing home "gateways" to ensure high quality delivery, with multiple IP avenues of distribution throughout the home to IP client devices.

What MSOs want to avoid is a newly rolled out video services architecture that can be tainted by improperly subscriber "engineered" home networks. An IP home architecture, potentially managed, provides some assurances that are not available today when homeowners are combined with STBs, cable modems, routers, splitters, and coaxial cable.

Hi QoS Audio Streaming

"A successful service might grow to a saturation penetration of 20% HP by 2005"

This prediction is simply a swing and a miss. This can be attributed to the ease of high quality audio delivery over the top (Internet Radio examples), in part because of the modest bit rates when compared against the increasing downstream tiers. Figure 1 is a good reference point – when video becomes supportable, audio becomes nearly insignificant.

In addition, the expectation bar has been lowered somewhat for audio through the years. Audio services on the web are often likely background or in concert with other multi-tasking functions – something not behaviorally similar in a video environment. And, analogous to voice services, music on the go in the form of IPODs and PDAs with mp3 players have lowered the bar on consumer accepted audio quality. Perhaps audiophiles are fewer and further between, but consumers, en masse at least, have chosen convenience over hi-fidelity.

Fortunately, because audio is so bandwidth non-intensive, this misfire does not significantly impact overall bandwidth results.

IP Telephony

"IP telephony is estimated to grow from a 2% penetration in year 2002 to about a 30% HP penetration by year 2011"

VoIP growth has been robust since its introduction, although perhaps not quite as robust as predicted 10 years ago across the board for Cable VoIP services. Intervening factors were the introduction of over-the-top voice (i.e. Vonage), and the shift, in particularly demographically, towards consumers choosing a single voice service, and choosing the most convenient one – their cell service. Nonetheless, voice traffic is a rounding error in traffic analysis. It requires proper treatment (highest priority) in the DOCSIS world, but in terms of bandwidth consumption, it not significant.

Node Segmentation

Figure 2 shows a figure directly from [1], suggesting the time frame and justification for node splitting throughout the past decade.

This figure turns out to be quite prophetic. Its essential conclusion is that node sizes would be cut from the range of 2000 hp to 500 hp, which would be sufficient for downstream into 2011:

"By choosing to leave the downstream node size at 500 HP, more carriers are required but equipment cost is saved. This configuration supports expected traffic requirements through year 2011."

Most operators are pondering, planning, or executing that next post-500 hp split. On average, numbers are beginning to drop below 500 hp. This prediction turned out to be quite accurate as far as macro bandwidth trends driving downstream node sizing. Note also that the DOCSIS downstream count shows 6 carriers moving to 8. Again, while a larger set of DOCSIS carriers than in use in general today, most MSOs see this number of DOCSIS downstreams in their relatively near future. The introduction of DOCSIS 3.0, in which bonding technology enables higher service tiers @N x 40 Mbps, has likely accelerated the addition of channels. Basically, it makes it more probable that chunks of 4 channels at a time will be added than single new DOCSIS downstream.

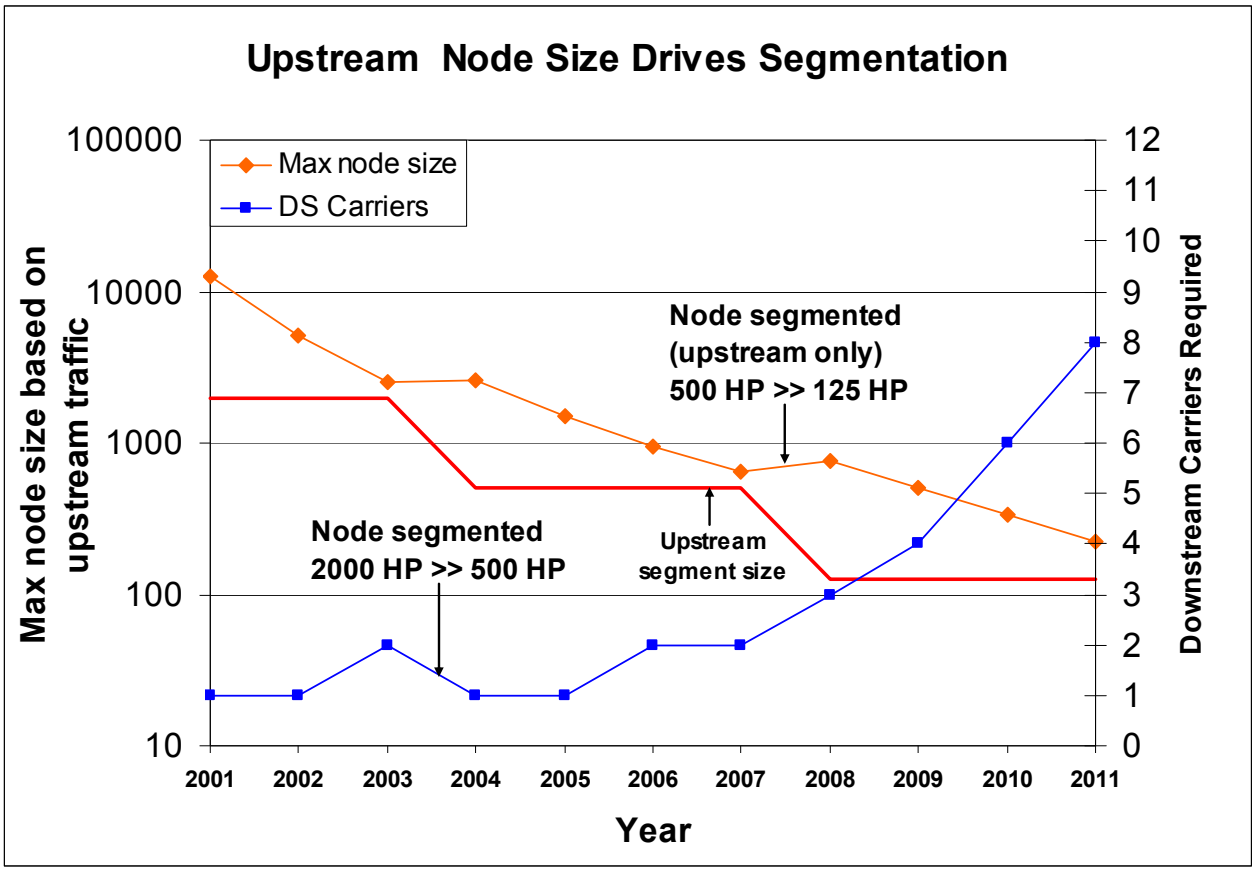


Figure 2 – Node Splitting Projections from [1]

COMPOSITE BANDWIDTH

Figure 3 shows a comparison of the cable spectrum, based on these components:

1) The historical basis from [1]
2) The predicted spectrum usage for 2011 from [1]
3) Three cases of "actual" meant to cover the range of "typicals"

Since the range of "typical" matters in a way that impacts available HFC bandwidth as defined by the upper band edge, it was felt that highlighting how service mix choices matter to this key parameter over a range would be valuable. This was also suggested in [4], where the ability to support new growth as a function of bandwidth available or created, and the broadcast/unicast mix, was quantified in years.

"It can be see there is spare capacity in HFC plants built out to at least 750 MHz. Downstream carriers are added to satisfy downstream demand through year 2011."

Notice that it was observed 10 years ago that the downstream spectrum supported the bandwidth growth needs of the anticipated service mix, given that serving groups sizes would be shrinking. As we know today, that has proven to be the case, generally, even with the underestimation of HD in the projection. Although, depending on the mix of other services, that the mix always fits comfortably in 750 MHz, which was the prediction, is not foolproof. However, actual growth has not exceeded the trends that already existed at the time that were resulting in the introduction of 870 MHz and 1 GHz plant equipment.

Consider now the upstream projection once again, and note that there was an expectation

that upstream growth would have forced a node split by this point. This is associated with the relatively accurate prediction that there would be more pressure on upstream bandwidth versus downstream bandwidth, and thus the upstream would drive the need for further segmentation. This has turned out to be the general scenario across MSOs offering HSD services, especially in competitive environments.

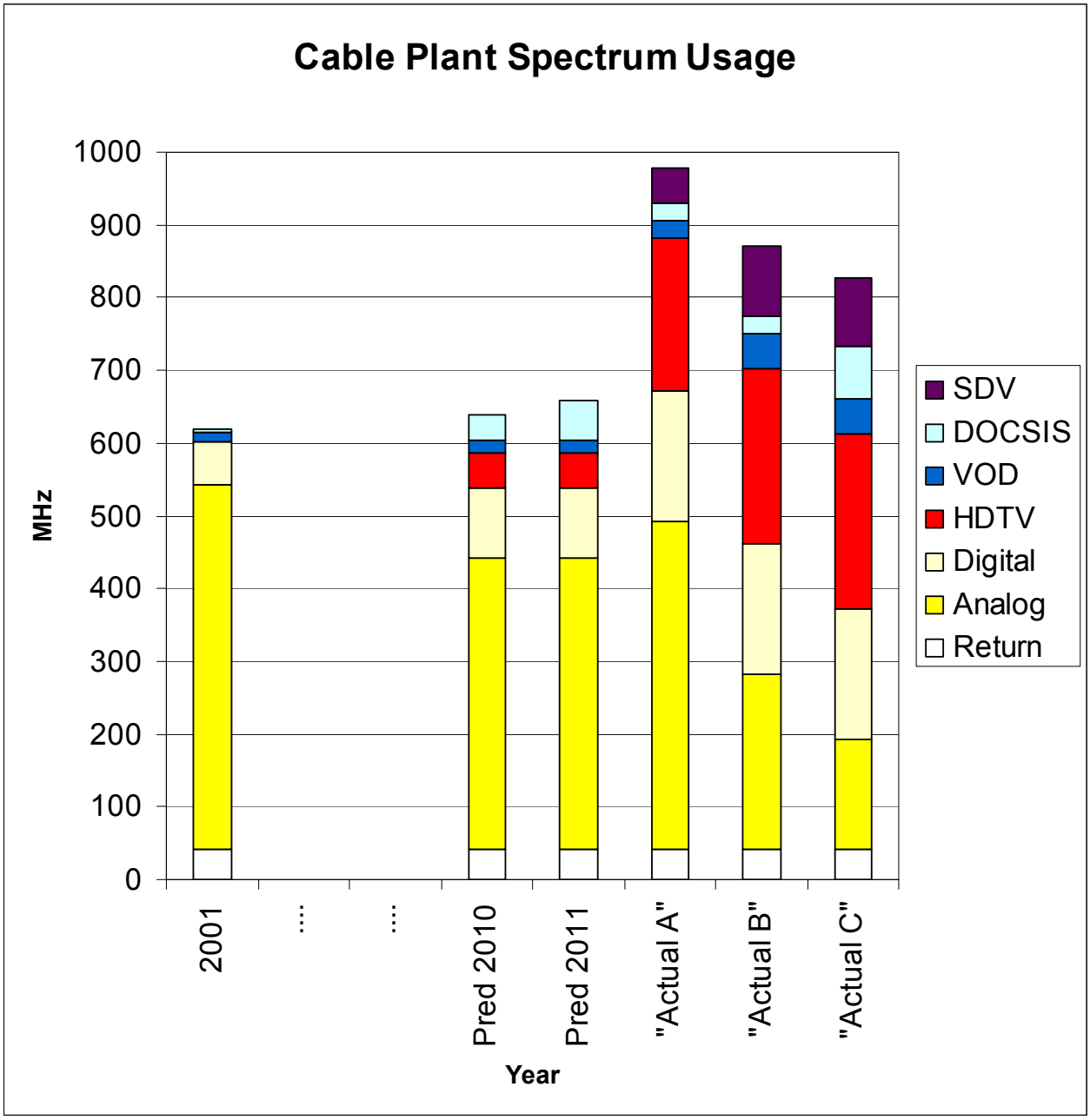


Figure 3 – Historical, Predicted, and Current Spectrum Usage

However, the prediction that an additional split to 125 hp would be necessary for the upstream was driven in part by an underestimation of what DOCSIS would evolve into. The original analysis assumed that the upstream would be capable of about 80 Mbps maximum, based on using 16-QAM

@ 3.2 MHz bandwidth, or about 10 Mbps per channel. It did not foresee the implementation of 64-QAM, nor S-CDMA to turn on the low end of the band. As such, the prediction is roughly one-half of what can be squeezed through a fully optimized and exploited upstream today. Thus, being off by almost one-half, there is nearly one additional "traffic doubling period" missing from [1].

Now, at 25% compounded growth, traffic doubles in roughly three years, while it doubles in roughly two years at 40% compounded growth. The original analysis from [1], based on approximately 59% compounded growth, is thus approximately 1.5 years off in time with respect to recommending a node split for upstream bandwidth. That is, traffic doubling at a 59% clip takes about 1.5 yrs. Applying this correction, the analysis would have concluded that a node split would be required at year "2008.5." And, indeed, dropping node sizes below 500 hp average lines up in time with this trend in competitive markets on an as-needed basis. The granularity of node-splitting was in fractions of one-fourth in [1], merely because the ability to segment most core node families in fourths. One segmenting visit per node was assumed.

Looking back, rather than compounding the concurrency year-on-year, we can recognize that increased speeds lead to lower concurrency for the same service rate on web-browsing type services, which has some logic to it – things get there faster to consume, but consumption time (human oriented) is about the same. This phenomenon is shown in Figure 4.

Of course, as web browsing continues to become multi-media oriented, this could reverse course in favor of supporting committed streaming requirements. Nonetheless, the point of this lesson learned was that, should compounded concurrency growth be removed, the upstream compounded growth rate predicted drops to approximately 30%, and thus roughly 2.5 yrs of growth time. In this case, the node split prediction would have recommended that 2009.5 would be go-time for the next split – again well within the ballpark on where upstream growth has taken us.

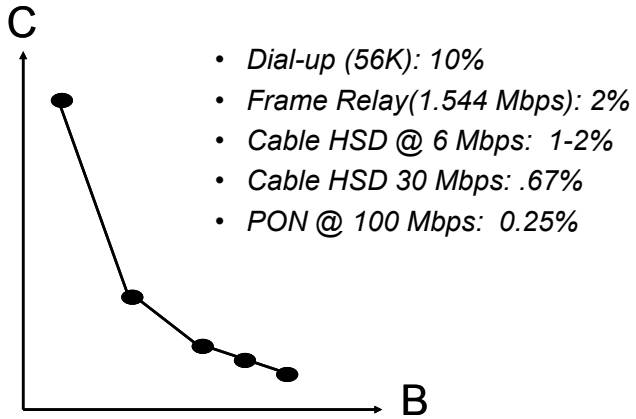


Figure 4 – Historical Concurrencies Observed

All things considered, although the mix had some miscalculations, both the upstream and downstream macro bandwidth requirements were reasonably well-aligned with where things have gone. The largest "miss" associated with HD penetration that underestimates its contribution to overall bandwidth is in part offset by the

introduction of SDV, which allows for additional HD programs to be added to the "broadcast" line-up without requiring an allocation of spectrum, at least not 1:1. Also, while the paper notes that analog reclamation will not be observed in a large way by 2011 – true in cases, not true in others – the consideration of this highly efficient means of exploiting HFC allows for the net spectrum predicted versus "Actual" (case C) to be similar.

While a key miss was the introduction of SDV into cable these last couple of years, there was certainly commentary suggesting the need for more and broader content, and the potential for the eventual role of HD content. This 10-yr old thought is the kind of logic that led to the march towards adopting SDV:

"….there are systems that would like to provide more broadcast channels. Needs include serving multi-lingual and ethnic populations with international programming and lots of HDTV, eventually"

FTTP ARCHITECTURES

"Beyond year 2011 MSOs will have to decide between pushing HFC capacity further or re-trenching to bring fiber to the home"

"Capital costs for FTTH are expected to become competitive for green fields deployments well with the 10 year forecast period"

Both of the above statements have turned out to be sound predictions. Many MSOs are looking at whether or when HFC naturally migrates to FTTP, and are developing transition technologies (i.e. RFoG and DOCSIS back-office for PON) to enable such a possibility. There is little argument that fiber and coax builds (parts & labor) cross in optimal choice as densities decrease, but the build out to the FTTP home still comes at a CPE premium under more typical HFC densities, although the gap is shrinking. Other factors, however, have driven "greenfield" environments to become based on FTTP, such as the real-estate development market, and the general perception that if FTTP is an expected endpoint, it does not make sense to be installing new coaxial cable.

"FTTH is considered the "end game" since it offers enormous bandwidth to the home"

It is hard to argue the point "more bandwidth is better." Of course, costs and legacy obstacles make this not a cut-and-dry question when considering time frames. It can be easily shown [4] that HFC architectures can be incrementally improved and exploited to deliver capacities that enable tremendous new growth that converts into years and years of life span. The question is whether that span exceeds the time frame for which practical business planning needs to take place, or if it is on the horizon in a way that plans need to be put into place now to enable this FTTP transition. Or, even if not an obvious "endgame," is "just in case" investment and planning warranted. Most operators fall into this latter category – its too big to ignore, and no one wants to be left without adequate bandwidth given the historical inability to recognize the onset of the new killer applications.

A DECADE AGO IT WAS CONCLUDED…..

"In 10 years the number of bits pouring into the home will be over 50 times the amount delivered today"

This was stating that straight line growth in bits *consumed* would likely continue at a compounded growth rate close to 50% (i.e. 1.48^10, or 48% growth for 10 years, equals 50x). It does *not* mean that there is 50x more QAM carriers, of course. This QAM count number is on the order of 10x. Of these 10x more carriers, the consumption patterns of consumers has shown a steady compounded growth along the lines of Moore's law for computing power. This simplistic model is generally associated with data services, but as video and data services begin to blur in the digital realm, the proper traffic growth model to use for consumption becomes less clear.

"Rich interactive multimedia video will be commonplace"

While interactivity has increased, a decade ago it was felt that by now it would have emerged from the shadows and be representative of the "typical" viewing experience. This has not occurred for a myriad of reasons, but initiatives such as EBIF, OCAP, and architectures in CE circles still exist with expectations to do so. Of course, the fact that multiple initiatives are still in the mix is part of the reason interactivity has not scaled as expected.

"HDTV will succeed as one of the many services"

Though not much of a reach to predict at the time, HD taking hold had become a major topic of discussion at the time of [1] because of how slow this seemed to be taking place. Nonetheless, this is a clearly true prediction, and, as indicated, in fact was underestimated in [1], albeit not by very much in years at this instant of time. While not foreseeing the trend trajectory as aggressively as it has played out, HD mass adoption has been closer to the back end of the 10-yr period analyzed.

"Telephony will become a rounding error in the traffic analysis"

Indeed, once HSD scaled with year-on-year compounded growth, voice bps became insignificant from a bandwidth perspective.

"This growth has been shown to be easily supported by continuous upgrades to the HFC infrastructure"

The relative ease of scalability of the HFC plant (yes, field folks, I know what is involved here!) has been proven out yet again as new services are added with incremental changes to access networks. Also, there is much room left in HFC in terms of capacity:

"Much more can be squeezed out of HFC, if and when needed…."

– so much so that the discussion on where to "end" hinges more on capex spending/opex maintenance questions:

"MSOs will continually be faced with capital investment trade-offs between infrastructure upgrade costs vs. how much excess capacity to install…."

– as well as hinging on developing strategies for retiring legacy infrastructure:

"Legacy equipment will tend to make upgrade trade-offs more complex and optimum timing will vary"

TEN MORE YEARS OF POSSIBILITIES

Moving forward, there are a bundle of new service opportunities that can keep the bandwidth growth line trajectory moving northbound, as well as some practical and

less sexy reasons bandwidth will be on the rise.

Simulcast Redundancy

Somewhat ironically, the fact that there are multiple opportunities aggravates the situation because of the need to support consumers on the network for existing services. This plays out primarily on the video side, where, because of the notion of "TV Everywhere" and the continuing enhancement of the video experience, it will be important to have a "simulcast" strategy. That is, today's network uses simulcast to support digital and analog carriage of a subset of available channels to support the different tiers of cable services. In addition, digital channels broadcast in SD are simulcast HD for some (increasing) subset of the channel offering. We can expect this same situation to play out for 3D services, since essentially no consumers today have 3D TVs, so the initial availability of content for it will require a 2D version. Finally, supporting multiple devices typically means smaller screen versions being available (VGA, CIF, etc). While these are expected to be delivered over the IP infrastructure, they do represent redundant streams of bandwidth. Fortunately, for both access networks and storage architectures, the "small screen" bandwidths are much smaller than their HD counter parts, so they carry less impact. Furthermore, introduced via the IP network, the opportunity exists to introduce the service as MPEG-4 encoded only, another roughly 50% average savings.

This bandwidth logic above is one reason that multi-screen edge transcoding has given way to multiple-stream storage models. At this stage, real time transcoding at the consumer edge is to costly to envision as a way to resolve the simulcast bandwidth issue, and, initially at least, the bandwidth premium for IP video streams may be minor. In addition, storage is relatively cheap, and the incremental additional storage, even for multiple formats, adds up to something palatable given that there is a growing library of HD content that must also be managed.

In time, of course, with the transition to IPTV becoming the video delivery endgame, and DOCSIS being the HFC vehicle for so doing at least in the foreseeable future, the bandwidth management bubble will involve the pace of the retiring of MPEG-TS delivery while increasing IPTV delivery. Clearly, introducing the latter (IPTV) must come before the acting on the former (MPEG-2 TS) if the transition is not to be an abrupt one. Equal or better video QoE in the IP domain will be necessary, and that means SD and HD delivery, creating a non-trivial problem. With the deployment of SDV, both technologies can take advantage of switched multicast statistically, so each has these built-in efficiencies. However, it is worth pointing out that the SDV QAM count for virtually "infinite" content for SD must be multiplied by roughly four for HD content.

Thus, a key element of an MSO strategy we expect to encounter moving ahead is an effective strategy for managing the "simulcast bump" that new services create. The bump is likely to encountered as a series of smaller bumps to hurdle at different times.

Video Services – Still on the Move

Let's move onto the content itself. The bandwidth hog is, and will continue to be, what is required to support consumer video expectations. Today, that expectation is satisfied by 1080i HD. However, 1080p HD is not far behind, adding to the per-stream average rate. Resolutions higher than HD are also on the drawing board (e.g. Ultra-HD – 4k x 2k) that can be 4x or more the total

pixels of HD, or 4x the raw bandwidth. It is likely to be accompanied by compression advances, but also likely not 1:1 with resolution increases, in particular given the time it takes to develop encoding technology.

The maturation of MPEG-4 is the bit rate quiver available to battle the bandwidth bulge, in addition to the continuing use of current HFC tools of SDV, reclamation of analog, and plant expansion to 1 GHz. To the extent that a system objective may be to engineer for everything-on-demand, full unicast delivery, fiber deep supporting service group splitting supports the necessary per-subscriber bandwidth to enable this. That is, in order to support the demands of a unicast architecture from a traffic engineering of spectrum perspective, node splitting and segmenting to smaller and smaller serving groups makes it quite reasonable to deliver full downstream unicast under some pretty aggressive consumption assumptions.

Consider the following scenario first described in [4]:

Five simultaneous (viewing + recording) Ultra-HD streams in 3D (first generation) over MPEG-4. This scenario contains a mix of bandwidth killers and helpers:

- Ultra-HD is hungry as described above, but not "Super-Hi Vision," which is hungrier still
- Use of 3D adds bandwidth to capture the left-right eye perspectives. We do NOT account for the reduction of bandwidth over time.
- H.264/MPEG-4, though not widely used today in cable, will be in the time frames of this projection
- Five streams is clearly aggressive, but of course U-Verse today advertises four simultaneous streams (not all HD)

This adds up to slightly below 135 Mbps of CBR video services. Consider a very aggressive penetration @75%, a very aggressive concurrency of use @ 50%, and a 1 Gbps data service @1% oversubscription, typical for data access. For services provided by only today's 256-QAM, 6 MHz QAM channels, we can show that HFC can ultimately support this, as serving group sizes are reduced, as shown in Table 1 (Red-to-Yellow-to-Green having the obvious implication). In [4], techniques to exploit more coaxial bandwidth are also described, such that we can go beyond conventional wisdom of 5-6 Gbps of RF access bandwidth.

The use of CBR-only delivery is a significant conservative factor – adding the increased efficiency for VBR delivery previously described moves the solid bar of comfort level (Green-to-Yellow) northbound. This also implies that, similar to the prior discussion of using the introduction of MPEG-4 as an opportunity to inject transformative changes to the infrastructure, introduction of advanced services such as 3D, at lease in scale, might best be implemented within the context of the IPTV transition.

HHP/Node	Req's GBPS	QAMs/Node	Spectrum (MHz)
250	14.53	384	2304
180	10.46	276	1656
125	7.27	192	1152
75	4.36	115	690

Table 1 – Supporting an Extreme Services Mix Over HFC

Data …… or Video?

As described above, we can scale downstream data services to 1 Gbps under reasonable traffic metrics and support that growth over HFC. Whether bonding of DOCSIS QAMs is the most effective way to do that longer term is questionable, or even whether such a service rate should be an RF solution, as opposed to a fiber solution. For commercial service to large enterprises, the Gbps rates make sense. However, for this customer set, FTTP solutions overlaid onto the HFC are an effective alternative to burning RF residential bandwidth. There is an expectation with residential services that rates will continue to climb on average at some 20-40% compounded rate. Perhaps more importantly, however, is that concurrency factors will shift as the HSD content shifts from web pages to video clips to OTT video. This would require allocating more bandwidth, linearly with the increase in concurrency. Fortunately, if allocated for a service of 1 Gbps at 1%, the same math holds true for 100 Mbps at 10%. Similar to how data growth was estimated in [1], a compounded growth rate assumption looks still to be a useful way to capture growth. If it can be reliably broken into component pieces of rate growth and concurrency, it perhaps becomes a better tool for understanding total bandwidth needs going forward.

On the upstream, while audio file sharing was all the rage driving bandwidth in [1], it is less so today because of, well, the law, and itunes. Nonetheless, rapid upstream traffic and service tier expansion has continued as a strong element of HFC planning. As in downstream, video services upstream could be bandwidth busters, and many possible applications that might accelerate this have been kicked around for many years. They may or may not take hold in a big enough way to matter, but over the next several years at least, there is a more important concern for upstream than postulating about new applications. That issue is simply staying ahead of normal growth in the available 5-42 MHz of very imperfect spectrum, and preferably working the service rate up to 100 Mbps – a logical market target to support broadly, and recently set as an FCC national objective. It is not a simple thing to accomplish 100 Mbps within today's spectrum. And, it is not difficult to show that under compounded growth assumptions, the upstream lifespan runs out of steam before the downstream, and within a period of time to be planning for next steps for upstream bandwidth strategy [5].

CONCLUSION

Ten years ago, an analysis was undertaken to project bandwidth requirements on HFC, with the intent to derive bandwidth needs for a decade's worth of growth. It was hoped that MSOs could use the information as a planning tool. Several of the individual services were predicted quite well in how they would scale (DTV, VOD, HSD), some were significantly underestimated (HD),

and some were missed completely (SDV). The analysis offered many unquantified trend projections that turned out to be quite prophetic. Finally, macro bandwidth projections due to the services growth described turned out to be reasonable, the prediction that RF plant had runway to support the growth was on target, and the projection of timing and logic for node splitting for service group reduction also turned out to be pretty accurate, all things considered.

The resulting assessment leads to a couple of important, confidence-building conclusions:

1) A sound understanding exists on the important, larger picture of the behaviors of consumer broadband consumption and growth, and the business implications

2) The flexibility and scalability of the HFC architecture has held to be as powerful as anticipated. This can be traced to HFC being built with just the right number of component parts, each of individual scalability and interoperability, enabling incremental investment that can be simply implemented, and ultimately rapidly paid for and profited from.

3) We have our work cut out for us for the next ten years to be so prophetic!

The job now revolves around item 3): quantifying the next set of projections based on the incoming mix of new possibilities described – video services, data speed trajectories and potential for "killer apps," and the introduction of commercial and wireless services support. At Motorola, we continually update our thoughts on where the bandwidth comes from and where the access network goes over time to support it. As we put behind us this last decade of growth, we similarly project the next decade by combining the lessons learned herein, the anticipated mix of new services, along with a dose of reality check of where on the hype curve these new services exist, and what these factors are likely to project to relative to mass adoption.

REFERENCES

[1] Randy Nash, "A 10-year Residential Bandwidth Demand Forecast and Implications for Delivery Networks," 2002 National Cable Television Association (NCTA) Show Proceedings Technical Papers, New Orleans, La., May 5-8, 2002.

[2] Dr. Robert Howald, "Web Surfing to Channel Surfing: Engineering the HSD Edge for Video," 2009 Cable-Tec Expo, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Denver, CO, Oct 28-30, 2009.

[3] John Ulm and Patrick Maurer, "IP Video Guide – Avoiding Potholes on the Cable IPTV Highway," 2009 Cable-Tec Expo, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Denver, CO, Oct 28-30, 2009.

[4] Dr. Robert Howald, "Fueling the Coaxial Last Mile,"2009 Society for Cable Telecommunications Engineers (SCTE) Emerging Technologies Conference, Washington, DC, April 3, 2009.

[5] Dr. Robert Howald, "The Broadband Horizon," Cable Next-Gen Broadband Strategies: Docsis 3.0, Wireless, Fiber & Beyond, Denver, Co., Feb 25, 2010.

ACKNOWLEDGMENTS

The author would like to thank the following individuals for their insights and support in gathering data critical to developing this paper: Curtiss Smith, Mohsen Manoochehri, Tony Stormer, Michael Brannan, Jack Moran, John Ulm, Fred Slowik, Chris Bastian (Comcast).

COMPARISON OF TECHNIQUES FOR HFC UPSTREAM CAPACITY INCREASE

David Urban
Comcast

Abstract

This paper compares three techniques for increasing the upstream capacity and peak upload speeds for a hybrid fiber coaxial cable network. The first technique is to increase the spectrum utilization and signal robustness in the 5-42 MHz band. The second technique is to allocate more upstream spectrum by changing the mid-split cross over point between the upstream and downstream bands. This paper shows some measurement results to help quantify the levels and extent of interference to televisions, set top boxes, and digital transport adapters to upstream transmission in the 42-85 MHz band with downstream signals in the 108-750/860/1002 MHz band. The last technique is the use of spectrum above 1 GHz. The advantage of this approach is that it can be built incrementally as needed without impacting the 5-1002 MHz HFC plant and services.

Considering these three alternatives for upstream capacity increase, begin by fully utilizing the 5-42 MHz spectrum. If in the future, it gets to the point where the 5-42 MHz is close to being fully utilized and node segmentation reaches its practical limits, then a change in mid-split cross over point between the upstream and downstream spectrum can be implemented with a mid-split RF protection circuit that transmits upstream signals in the 42-85 MHz band while protecting in home devices from interference. A mid-split RF protection circuit works with existing devices and standards, and adds significant upstream capacity with a small sacrifice in downstream capacity. Use of 1200-1800 MHz spectrum is an approach to incrementally add 1 Gbps symmetrical services while preserving current HFC services.

INTRODUCTION

DOCSIS 3.0 with an upstream spectrum allocation of 5-42 MHz as built in North American hybrid fiber coaxial network architectures has greatly increased the upstream capacity and peak speeds when compared to DOCSIS 2.0. With single carrier DOCSIS 2.0, customers enjoy upstream speeds in the 2-10 Mbps range and with DOCSIS 3.0 upstream channel bonding customers can look forward to even higher speeds. DOCSIS 3.0 includes features such S-CDMA with maximum scheduled codes and selectable active codes that provide plenty of headroom for these speeds to grow. Using 64-QAM upstream modulation, four upstream carriers with 6.4 MHz channel width, three 3.2 MHz channel width upstream carriers, and one 1.6 MHz channel width upstream carrier will completely fill up the 5-42 MHz spectrum and provide a total upstream capacity of about 155 Mbps. DOCSIS 3.0 cable modems can bond four upstream carriers for peak upload speeds of 100 Mbps. Getting the upstream capacity to 155 Mbps with DOCSIS 3.0 in the 5-42 MHz spectrum faces challenges. Impulse noises from sources such as motors and lighting fixtures have short time duration. S-CDMA uses a long symbol time so that the impulse noise only impacts a fraction of the symbol. Ingress noise from sources such as short wave radio signals are narrow in spectrum but have long time durations, good ingress cancellation techniques are essential for fully utilizing the upstream spectrum. Portions of the upstream spectrum are used for digital cable set top box return path signaling, converting these devices to DOCSIS is necessary to fill up the 5-42 MHz spectrum entirely with DOCSIS carriers. Televisions in the home tend to be behind more splitters than data cable modems and

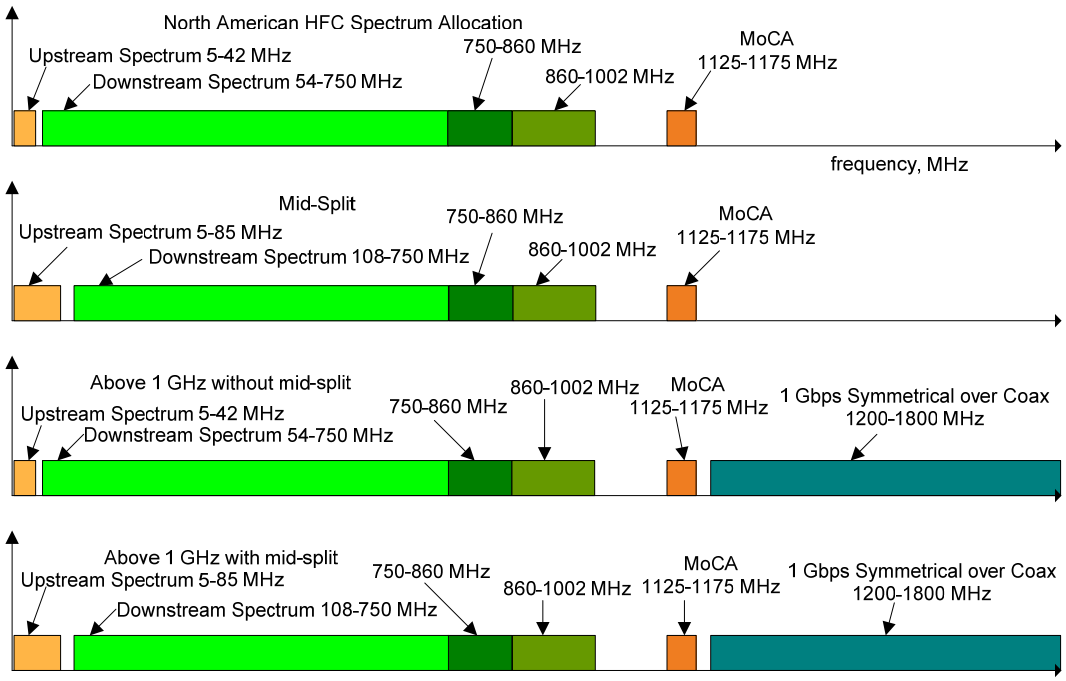


Figure 1. Spectrum Allocation Options for Upstream Capacity Increase.

voice cable modems; this makes it important to be able to operate with high attenuation in the upstream signal path. Regardless, 155 Mbps is the capacity using the highest order modulation for DOCSIS 3.0 and fully filling up the 5-42 MHz so something clearly will have to change to get any more than this.

This paper compares three techniques to surpass the 155 Mbps upstream capacity ceiling. The three techniques are 1) non-linear harmonic ranging and pre-distortion along with Hybrid ARQ (automatic repeat request with variable forward error correction), and adaptive modulation and coding for the 5-42 MHz upstream band, 2) upstream HFC mid-split to increase the spectrum allocated to upstream, and 3) use of HFC spectrum above 1 GHz. These options are illustrated in Figure 1 highlighting that a change in mid-split can be done alone, use of spectrum above 1 GHz can be done alone, or both mid-split and above 1 GHz can be done in conjunction. The downstream rolloff of fiber nodes and amplifiers is commonly 750 MHz or 860 MHz and can be as high as 1002 MHz. This paper addresses upstream capacity and all methods described are applicable to 750, 860, and 1002 MHz HFC plant.

The upstream capacity in HFC networks can be increased as needed by adding more upstream carriers and serving smaller groups of users. Capacity can be increased to match demand by first reducing the number of fiber nodes that share a common 5-42 MHz spectrum and adding multiple upstream carriers, then segmenting the fiber nodes into multiple upstream legs each with its own set of upstream demodulators and optical transmitter and receiver, and finally splitting nodes further in order to serve a small number of homes with the 5-42 MHz upstream spectrum. At some point, however, more practical methods to add capacity than further node splitting make sense. For example, rather than add an upstream carrier to a low portion of the upstream band that has interference issues, it may be better to work at a higher portion of the upstream spectrum that has less noise. Rather than adding more QPSK carriers, it may be more efficient to increase the modulation rate to 16

QAM, 32 QAM, 64 QAM, or even higher. Rather than continually serving a smaller and smaller number of customers with a given amount of bandwidth, it may make sense to keep the serving group size the same and increase the bandwidth.

The same situation occurred in early AMP (Advanced Mobile Phone system based on wideband FM) cellular deployments, in theory AMP cellular systems could be split into smaller and smaller base station sizes to provide whatever voice capacity was needed. However, the unattractive prospect of a base station outside everyone's home led to the development of more spectrally efficient solutions, TDMA then CDMA, then WCDMA, and finally OFDMA.

The upstream capacity and peak upload speed can be increased in three ways, 1) better spectral efficiency in the 5-42 MHz band, 2) a mid split to increase the upstream spectrum allocation, and 3) use of spectrum above 1 GHz.

A rough estimate of the upstream channel capacity for several of these techniques is shown in Table 1. The spectral efficiency used is estimated based upon the spectral efficiency of upstream carriers operating in cable plants today, 27 Mbps in a 6.4 MHz channel width is 4.2 bps/Hz. The 5.6 bps/Hz assumed for a more spectrally efficient use of the 5-42 MHz is the spectral efficiency of 256-QAM replacing 64-QAM. This is not to say that these spectral efficiencies are always possible under all channel conditions but hopefully the table provides an understanding of the relative potential of each approach.

Table 2 shows a rough comparison of the plant and CPE impacts for several options in upstream spectrum allocation. Sticking with the 5-42 MHz upstream allocation requires no changes to plant passives or actives, works with all set top boxes without the loss of any downstream spectrum and requires no new transceivers. Increasing the upstream spectrum to 65, 85, or 200 MHz requires changes to amplifiers and fiber nodes in the plant but no changes to plant passives, the forward data channel signaling to digital cable set top boxes is impacted by the choice of cross over frequency so some set top boxes may not be supported after a change in downstream spectrum allocation, a loss in HD streams is a tradeoff with more upstream spectrum allocation, mid-split approaches will work with DOCSIS 3.0 cable modems so that new transceivers are not required. Table 2 uses a loss of 3 HD streams for each loss of a 6 MHz wide spectrum slot and makes some rough estimates of diplexer separation requirements. The 5-200 MHz upstream band is not supported by DOCSIS devices, however, DOCSIS type devices could be extended in frequency or new transceivers could be developed to fully take advantage of the increased upstream spectrum. This is why the 5-200 MHz transceiver column is listed as a maybe. Use of spectrum in the 1200-1800 MHz bandwidth does not require changes to plant actives if fiber optic cable is run to the last active. Plant passives such as splitters, directional couplers, and taps need to be changed in order to pass 1200-1800 MHz signals. New coaxial transceivers are required for operation in the 1200-1800 MHz band. The column for "All STBs?" and "Most STBs?" is intended to indicate whether the solution will work with all digital cable set top boxes or most digital cable set top boxes and this is determined by the adjustments required for the out of band forward data channel center frequency.

The next section begins by taking a look at the first technique, getting more capacity out of the current 5-42 MHz upstream allocation.

Solution	Total Spectrum	Spectral efficiency	Upstream Capacity
Units	MHz	bps/Hz	Mbps
5-42 MHz DOCSIS 3.0	37	4.2	155
5-42 MHz 256-QAM	37	5.6	207
5-65 MHz mid-split	60	4.2	252
5-85 MHz mid-split	80	4.2	336
5-200 MHz mid-split	195	4.2	819
1200-1800 MHz	600	4.2	2520

Table 1. Comparison of spectrum allocation, spectral efficiency, and upstream capacity.

upstream band	Amp/Node changes?	Tap/DC changes?	STB Interference?	All STBs?	Most STBs?	HD loss?	New Transceivers?
5-42 MHz	NO	NO	NO	YES	YES	0	NO
5-65 MHz	YES	NO	YES	YES	YES	9 streams	NO
5-85 MHz	YES	NO	YES	NO	YES	27 streams	NO
5-200 MHz	YES	NO	YES	NO	NO	99 streams	Maybe
1200-1800 MHz	NO	YES	NO	YES	YES	0	YES

Table 2. Comparison of plant and CPE impact for several options.

Better Use of 5-42 MHz Upstream Spectrum

A common upstream carrier setting is 6.4 MHz channel width and 64-QAM modulation A-TDMA which has a TCP/IP data rate of about 27 Mbps after accounting for all the headers and error correction. A common upstream spectrum allocation is a 6.4 MHz 64-QAM A-TDMA upstream carrier with a 32.2 MHz center frequency and a 3.2 MHz 16-QAM TDMA upstream carrier with a center frequency of 37 MHz. The A-TDMA carrier has about a 27 Mbps data capacity and the TDMA carrier has about a 9 Mbps data capacity. The total upstream capacity is about 36 Mbps which is 23% of the upstream channel capacity of the 5-42 MHz spectrum entirely filled with 64-QAM carriers. Thus, it may be possible to increase the upstream capacity by a factor of four by adding upstream carriers and increasing the modulation rate while staying with DOCSIS 3.0 technology. By segmenting fiber nodes into four legs, another factor of four increase in upstream capacity can be obtained provided that four times the upstream demodulators are added. So the potential exists for as much as a 16 times increase in upstream capacity within the constraints of the 5-42 MHz upstream spectrum allocation and DOCSIS 3.0 technology.

This section of the paper on increasing upstream capacity within the 5-42 MHz band looks at two approaches. The first approach allows for a new OFDMA multiplexing technique while the second looks at how most of the benefits of the first approach can be applied using DOCSIS 3.0 cable modems by using S-CDMA. OFDMA stands for orthogonal frequency division multiple access and it is a common technique used in DSL, terrestrial broadcast, cellular, and wireless home networking which breaks spectrum up into very narrow tones so that symbol times are very long while the combined data rate is high. S-CDMA stands for synchronous code division multiple access and it is implemented in DOCSIS 2.0 and 3.0 cable modems using 128 orthogonal codes to again realize a long symbol time and still have a high aggregate data rate. Since there are so many cable modems deployed with S-CDMA capability, it makes sense to fully utilize this technology first. Nonetheless, we'll begin by discussing the potential of OFDMA and then take a look at how similar benefits can be realized with S-CDMA.

DTAB Discrete Tone Adaptive Bandwidth

OFDMA with pilot, ranging, and data tones, time and frequency coordination for compatibility with DOCSIS 1.0, 1.1, 2.0 and 3.0 cable modems, nonlinear harmonic ranging and pre-distortion, hybrid automatic repeat request, non-contention based scheduling should be looked at for their potential to increase the upstream capacity in the 5-42 MHz band.

The combination of these techniques to increase the upstream capacity is introduced in this paper as DTAB for Discrete Tone Adaptive Bandwidth. DTAB uses OFDMA to create sub-channels whereby each sub-channel consists of ranging tones, pilot tones, and data tones as shown in Figure 2. The ranging tones sweep through the 5-42 MHz spectrum and perform linear and non-linear harmonic ranging for each cable modem. Pilot tones use CMDA to support many cable modems at the same time and pilot tones also sweep through the entire 5-42 MHz spectrum in order to measure the channel quality for adaptive modulation and coding and adaptive amplitude control. Pilot tones are continuously sent by all cable modems and thus bandwidth requests for upstream transmission can always be made by every cable modem without delay or collisions using the pilot tones. Finally the data tones send upstream data transmissions using adaptive modulation and coding and adaptive amplitude level control, non-linear and harmonic pre-distortion, and hybrid automatic repeat request. DTAB is backwards compatible with DOCSIS 1.0, 1.1, 2.0, and 3.0 cable modems by allocating spectrum and time slots for these cable modems as shown in Figure 2. Figure 3 shows a break out of a sub-channel to illustrate the pilot, ranging and data tones that make up a sub-channel. Table 3 shows calculations of the data rate, sub-channels and pilots for DTAB.

OFDMA reduces spectral efficiency proportionately to the ratio of the cyclic prefix to the symbol rate. The cyclic prefix is needed to reduce inter-symbol interference due to multi-path reflections. A DOCSIS 3.0 cable modem has a pre-equalizer with 24 taps which has an equalization window of 4.7 microseconds. As an example, an unterminated 1,150 foot length of RG6 cable will have a reflection with 30 dB attenuation and a time delay of 2.75 microseconds at 32 MHz. With an OFDM symbol rate of 100 microseconds and the guard time of the cyclic prefix set to 5 microseconds to allow for a 5 microsecond reflection then 5% of the upstream channel capacity is lost.

How much guard time needs to be allocated for OFDM? Consider an amplifier spacing of 1,000 feet. The difference in travel time between a direct signal traveling from one amplifier to the other and a signal reflecting off the input of the second amplifier, traveling back to the first amplifier, reflecting off the output of the first amplifier and then back to the input of the second amplifier is 2.3 microseconds for 87% velocity factor cable. If the input and output return loss is 16 dB and the cable attenuation is 0.4 dB/100ft at 30 MHz for 8 dB cable loss then the channel impulse response will include a component with a 2.3 microsecond delay and a 40 dB down amplitude. An example of such a micro-reflection taken from plant data is shown in Figure 4. This reflection will need to be equalized for modulation schemes requiring high signal to noise ratio. Since 64-QAM is such a modulation scheme then the OFDM guard time should be a minimum of 4.6 microseconds to allow for a window of +/- 2.3 microseconds.

Looking at pre-equalization coefficients for DOCSIS 2.0 and 3.0 cable modems in operation, the tap energy is mostly concentrated in the three taps before and after the main tap as shown in Figure 5. Thus, there is considerable energy within a delay spread of 1.2 microseconds.

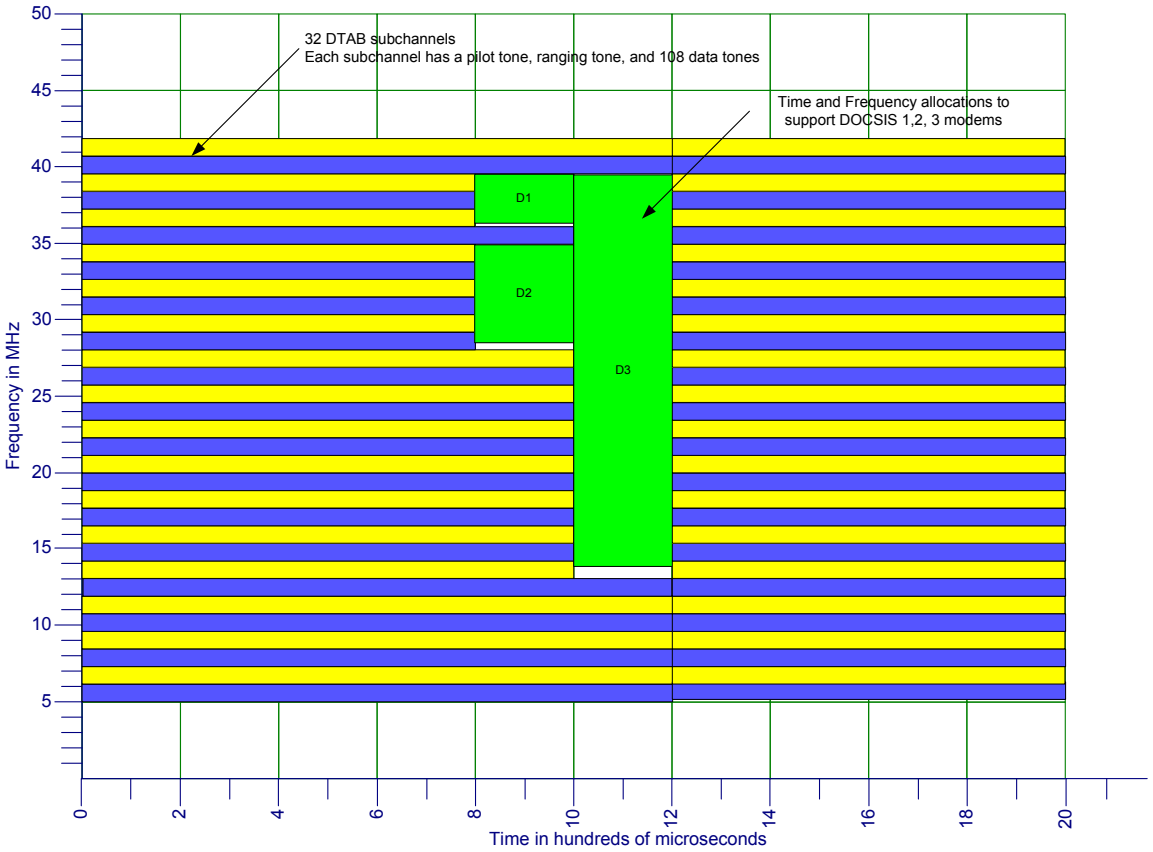


Figure 2. DTAB time and frequency domain showing backwards compatibility with DOCSIS 1.1, 2.0, and 3.0 cable modems.

Generally much further down in amplitude, energy is present in taps as far out as tap 24 as shown in Figure 6. This indicates that the full 4.8 microsecond equalization window is useful in correcting for reflections in the cable plant. Thus, metrics from the tens of millions of cable modems in operation support the minimum OFDM guard time of 4.8 microseconds.

Limiting the loss in data capacity to 5%, the minimum OFDM symbol time is 100 microseconds to allow for a 5 microsecond guard time.

Adaptive modulation with real time channel measurement can be used to increase the upstream capacity for a given availability. OFDMA with pilot tones that constantly scan the upstream spectrum can assign data tones 256-QAM modulation during times of good signal to noise ratio conditions and fall back to 16-QAM modulation during poor signal to noise ratio conditions. If the OFDMA pilot tones have narrow bandwidth and employ CDMA to support multiple modems on the same pilot tone then the upstream capacity sacrificed in order to make measurements of the real time channel conditions for each modem will be small. With a real time measurement of the upstream channel conditions, the optimal modulation for any given part of the spectrum at any given time frame can be chosen. This can result in a significant increase in upstream capacity with much higher availability. Surveys of upstream channel conditions reveal that a 33 dB SNR threshold allowing for 256-QAM is met for a significant portion of the upstream spectrum for significant time periods.

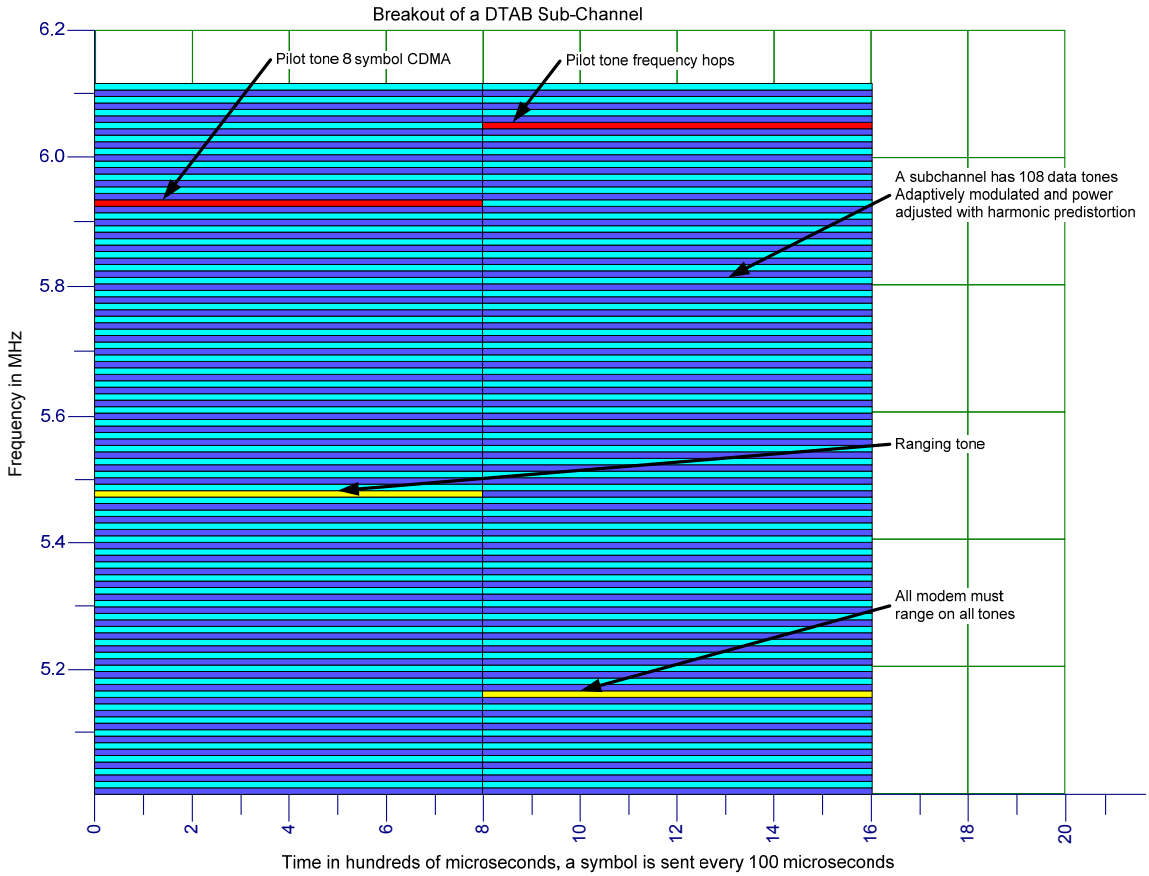


Figure 3. Breakout of DTAB Sub-Channel showing the Pilot and Ranging Tones Hopping and the length of the Pilot CDMA symbol time.

symbol period	100E-06	s
lowest frequency	5,000,000	Hz
highest frequency	42,000,000	Hz
BW	37,000,000	Hz
pilot tones	32	
data tones per pilot tone	108	
ranging tones per pilot tone	1	
number of tones	3,520	
carrier spacing	10,511.36	Hz
useful symbol period	95.135E-6	s
cyclic prefix (guard time)	4.865E-6	s
Spectral Efficiency	8	bps/Hz
total raw data rate	276E+06	bps
data tones	3456	
data rate per tone	80,000	bps

Table 3. Example of Sub-Channels, Pilots, and data tones for DTAB.

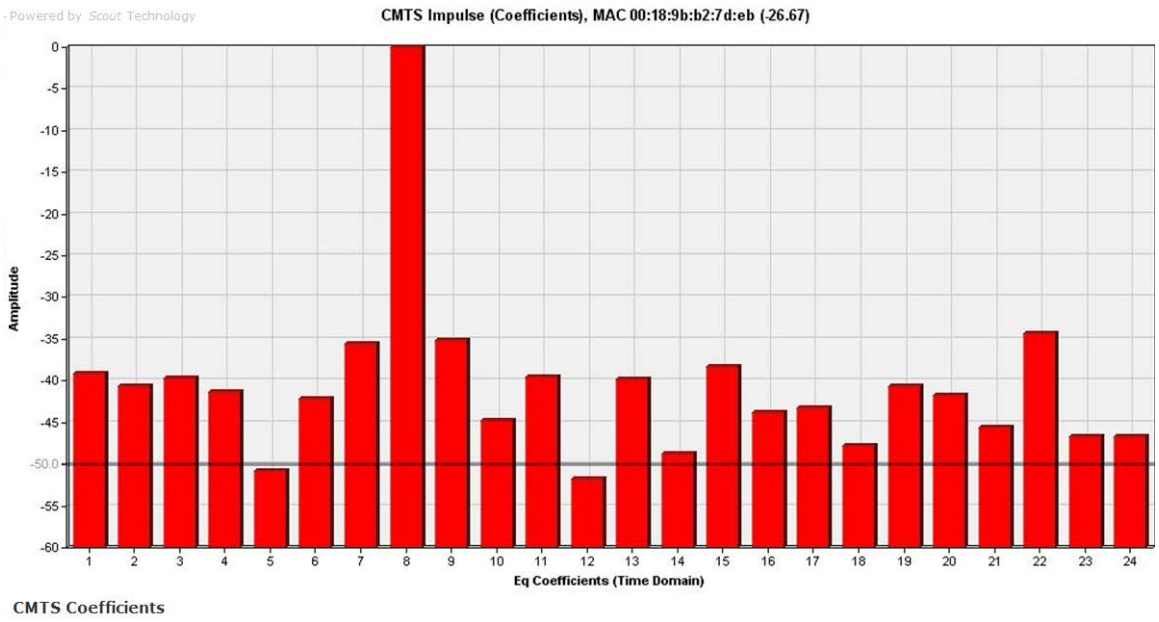


Figure 4. Example of reflection at 2.7 microsecond delay from main signal.

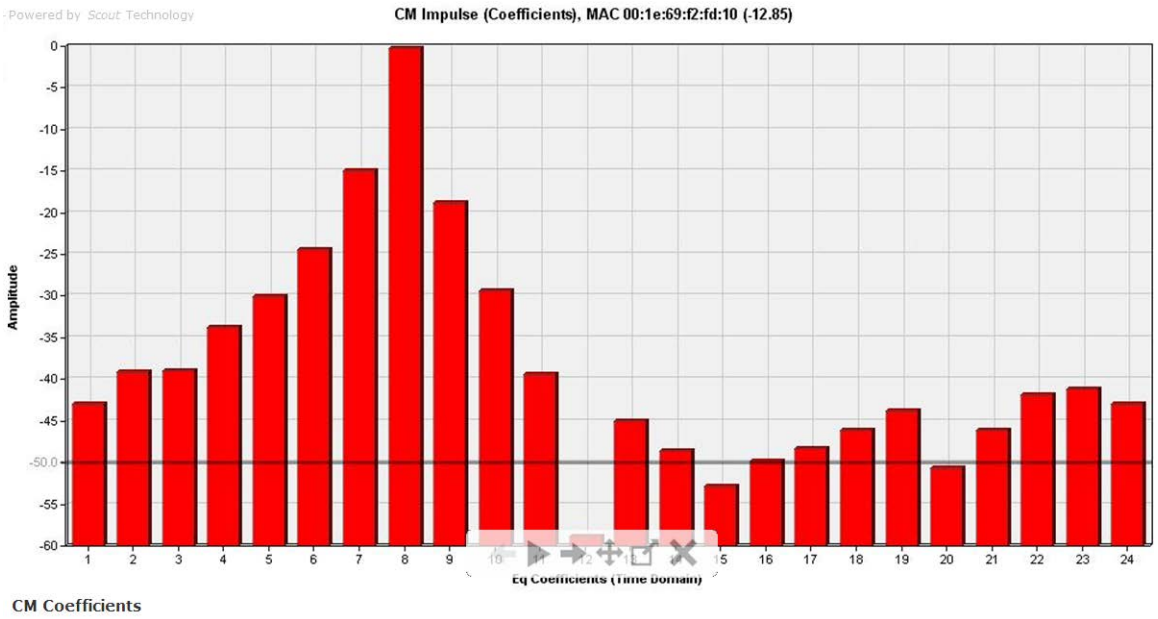


Figure 5. Tap energy is mostly focused within a 1.2 microsecond window.

A fundamental determinant of the upstream signal-to-noise ratio is the nonlinear and noise characteristics of the return path laser transmitter commonly illustrated by the noise power ratio (NPR) curve. If the input power to the laser is set to the optimum level, DFB laser transmitters can have a noise power ratio of 40 dB. However, it is wise to set the input level at a reasonable back-off below the optimal level to account for changes in temperature and measurement uncertainties. This is due to the fact that at input levels above the optimum NPR point, the noise is dominated by nonlinear third order intermodulation distortion whose degradation is much steeper (2 to 1 vs. 1 to 1). This allows for a significant opportunity to increase the total upstream capacity using nonlinear harmonic ranging and pre-distortion.

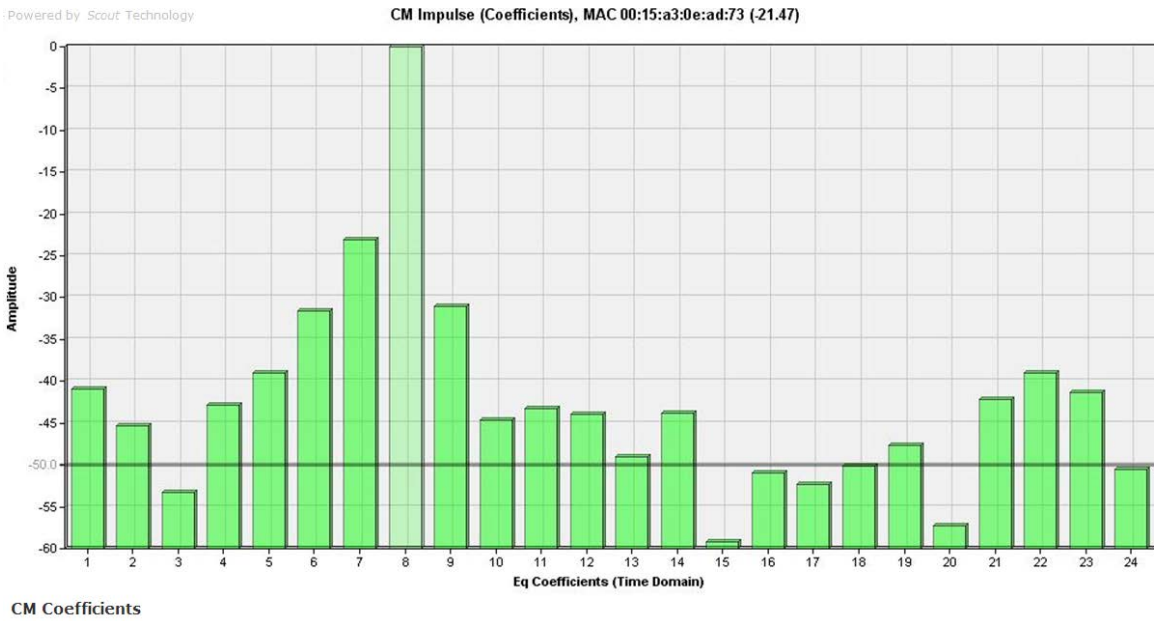


Figure 6. Pre-equalization tap with energy at taps 21, 22, and 23.

With OFDMA, tones can be assigned for ranging with the linear ranging for frequency, amplitude and time delay performed in the same fashion as DOCSIS 3.0. In addition to linear ranging, a harmonic nonlinear ranging can be performed, whereby a tone is adjusted in amplitude and the resulting 2^{nd}, 3^{rd}, 4^{th}, 5^{th} harmonics are measured at the headend receiver. By doing this the full nonlinear characteristics of the return path have been measured.

This nonlinear response is dominated by the return path laser transmitter but also includes the impact of the cascade of return amplifiers and even in home amplifiers. Since the ranging tones are used to fully characterize the return path nonlinear response in real time, the optimum set point of the NPR curve can always be used. Today, many upstream channels have signal to noise ratios below 30 dB so if nonlinear ranging results in a consistent 40 dB NPR, the net results could be an improvement in upstream signal to noise ratio of 10 dB. This would allow for the use of higher modulations and increased upstream capacity.

In addition, to using a control loop to find and set the level into the return path laser at the point of optimal NPR, since the OFDMA modems are capable of outputting tones throughout the entire 5-42 MHz spectrum and since the nonlinear response of the return path has been measured in the nonlinear harmonic ranging process, non-linear pre-distortion can be used to reduce harmonic and intermodulation distortion. This will further improve the modulation error rate of the upstream symbols and increase upstream capacity by allowing higher order modulation at low error rates.

Much of the interference levels in the 5-42 MHz band are due to shortwave radio signals, communications band radio signals, electrical lighting, electric motors. This type of interference is not consistently present and its occurrence is not predictable. For this type of interference, hybrid ARQ is more efficient than the RS-FEC and interleaving used in DOCSIS 3.0 upstream. RS-FEC adds parity bytes to every code word. This is wasted bandwidth when the interfering signals are not present. On the other hand, when the

interfering signals are present and reach a certain level, uncorrectable codewords result. Uncorrectable codewords can be seen in almost every cable modem if looked at over a long enough time period. With hybrid ARQ, very little FEC is applied during times when the interfering signals are not present. And when interfering signals show up and result in packet loss, those packets are re-transmitted when the interference goes away. Thus hybrid ARQ can increase the capacity and the availability at the same time for typical HFC upstream interference that occurs intermittently at very high levels.

OFMDA is power efficient for cable modem upstream transmitters since each modem can transmit a small portion of the total upstream spectrum. This helps with high attenuation cable modems such as those behind many splitters in the home. OFDMA is proven to be very robust against micro-reflections, narrowband ingress noise, and impulse noise due to the guard band of the cyclic prefix and long symbol times. Thus, the robustness of OFDMA will add capacity to the upstream in cases where parts of the spectrum are unusable with single carrier modulation.

S-CDMA with adaptive coding, channel measurement, ARQ, non-linear ranging

The cable modems in digital video set top boxes, data cable modems, and voice cable modems as well as next generation high data rate DOCSIS 3.0 chip sets are limited to S-CDMA upstream since they are not capable of OFDMA. While OFDMA does have some advantages over S-CDMA as indicated by the move from 3G cellular networks based upon CDMA to 4G networks based upon OFDMA, S-CDMA performance if fully developed can be very comparable to OFDMA, particularly for the static micro-reflections of a cable plant.

In fact, the 4G LTE (long term evolution standard for cellular networks) does not use OFMDA for upstream transmission but instead adopted variable bandwidth single carrier modulation to improve battery life of handheld devices. Several important techniques to increase upstream capacity in the 5-42 MHz band proposed in DTAB can also be implemented with S-CDMA and thus work with DOCSIS 3.0 cable modems.

Noise and interference conditions exist in portions of the 5-42 MHz return spectrum at certain periods of time that result in codeword errors with DOCSIS 3.0. As a result, the cable operator is faced with two unappealing options, to set the modulation rate overly conservative so that cable modems work most of the time or to operate at higher modulations and allow for longer periods of errors.

Periods of time can easily be observed in nearly all parts of the upstream spectrum with interference levels above the threshold for 64-QAM so that a fixed setting of 64-QAM will experience periods of time with codeword errors. Adaptive modulation with real time channel measurement provides a capacity increase by allowing higher orders of upstream modulation than one would dare with a fixed setting while at the same time improving reliability. Some versions of adaptive modulation and channel measurements have already been developed for DOCSIS. The developments to date show that even within the constraints of DOCSIS 3.0, adaptive modulation and channel measurement is possible. S-CDMA with maximum scheduled codes has been demonstrated to perform an equivalent function by reducing the number of active codes as the noise level rises. This is the fundamental technique used in 3G WCDMA cellular networks using a variable spreading factor. While a permanent reduction in the number of active codes would be inefficient, for example a 3 dB gain in noise immunity is

realized by using half the codes, the noise level in upstream receivers can be observed to drift by several dB over time. This allows for the use of most of the active codes under normal noise levels with slightly less codes being used during brief periods where the noise level rises.

Hybrid ARQ has been proposed for some systems such as PON (passive optical networks) for implementation at a higher layer in the OSI stack than MAC and PHY, using this approach it would be possible to implement hybrid ARQ with DOCSIS 3.0 S-CDMA upstream cable modems.

OFDM breaks the channel width into narrow frequency bands resulting in a long symbol time. Likewise, S-CDMA uses codes to spread narrowband input signals over a wide channel width resulting in a long symbol time. Since, OFDM consists of a number of narrowband carriers, or tones, tones that fall on top of narrowband interference sources can be turned off or reduced in modulation level. This gives OFDM the property of good ingress cancellation. With S-CDMA each code has a unique frequency response and a narrow bandwidth interference source will impact some codes more than others. Selectable active codes with S-CDMA can be used to effectively operate in the presence of narrow bandwidth ingress noise.

S-CDMA has many of the same properties of OFDM and it is proposed in this paper to investigate the idea of supporting adaptive coding, hybrid ARQ, and non-linear harmonic ranging and pre-distortion to S-CDMA systems. To further explore this idea, the next section shows some measurements of the non-linear harmonic distortion characteristics of a DFB (distributed feed back) return path laser with the notion of applying non-linear harmonic ranging and pre-distortion to DOCSIS 3.0 systems.

Using non-linear ranging and pre-distortion to increase upstream capacity.

There are many sources of interference in the HFC upstream 5-42 MHz and the observed interference varies with frequency, time, and amplitude. One trick to deal with the interference levels is to increase the upstream transmit level of the cable modem above the levels of the interfering sources. The upstream transmit level can be increased relative to the amplitude levels of the interference sources in order to operate error free at higher orders of modulation. However, this approach is limited by two hard ceilings; the cable modem maximum transmit level and the clipping threshold of the return path amplifiers and lasers. Figure 7 shows the distortion products from an overdriven DFB return path laser. The distortion products consist of spectral re-growth adjacent to the fundamental signal waveform and 2nd, 3rd, and 4th harmonics.

The following use case illustrates the conceptual implementation of non-linear pre-distortion to increase the upstream capacity. In this example, a high level interference source is present at low frequencies. Figure 8 shows a field measurement of this condition. In order to operate a carrier at 10 MHz center frequency with equal SNR to that of carriers centered at 20 MHz and 30 MHz, the 10 MHz center frequency carrier must transmit at a 20 dB higher level than the 20 MHz and 30 MHz carriers.

Figure 9 shows the output signal from the sample port of an optical receiver with a cable modem connected to a fiber node with 25 km fiber link between the upstream return path laser and optical receiver. The return path laser transmitter is a DFB. The cable modem signal is set to a 20 MHz center frequency with a 2.56 MHz symbol rate and 16 QAM in TDMA mode. The level of the cable modem transmit power was increased until the SNR was 33 dB.

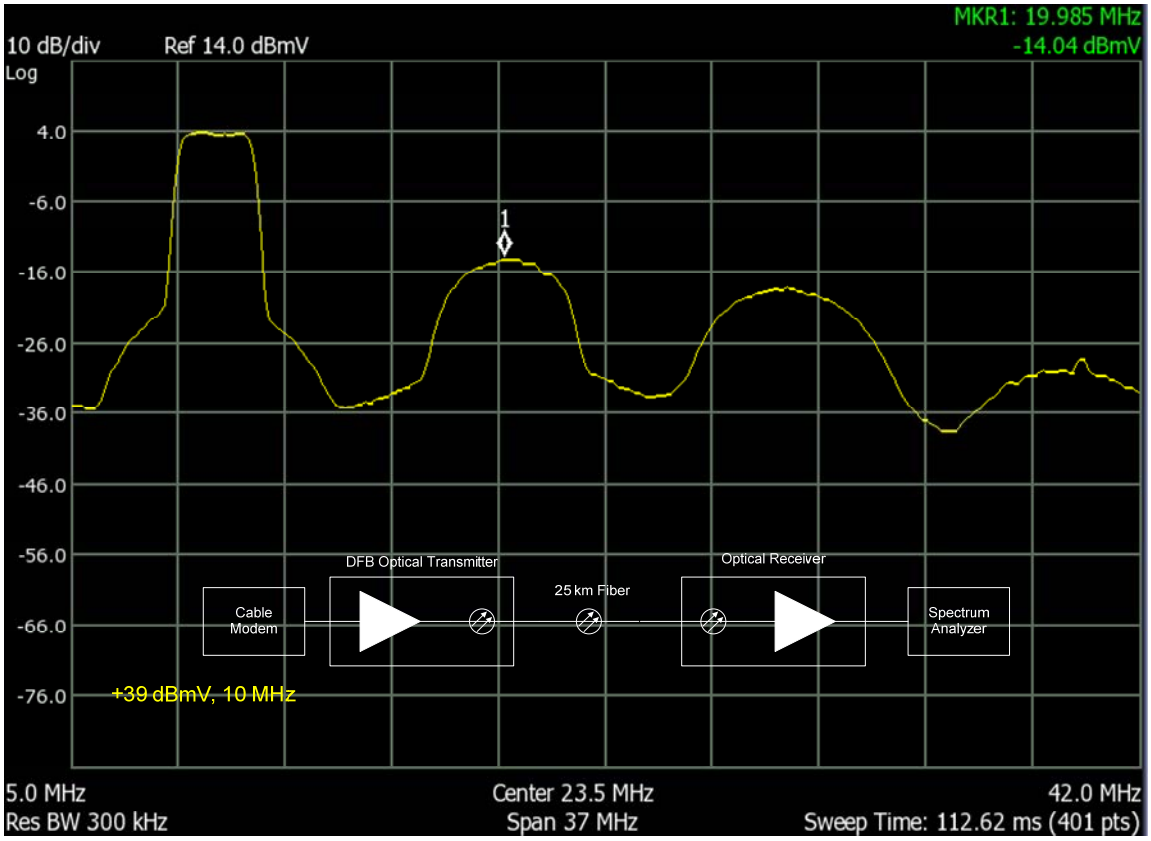


Figure 7. Overdriven Return Path Laser Showing Harmonic Distortion.

Figure 9 thus shows the lowest possible transmit power from a cable modem that will maintain at least a 33 dB SNR through the return path optical transmitter and receiver. Note that 33 dB SNR only includes the noise level from the return path laser and receiver, other noise and interference sources would further degrade the total SNR at the demodulator output.

Figure 10 shows a 10 MHz center frequency upstream signal with the cable modem transmit level set 20 dB higher than the level shown in Figure 9. The distortion produced at the 2nd, 3rd, and 4th harmonics would prevent the reception of signals at 20, 30, and 40 MHz. So in the case where external interference sources raise the noise level at 10 MHz 20 dB higher than the noise level at 20, 30, and 40 MHz, one cannot maintain a constant SNR across carriers by simply increasing the 10 MHz transmit power by 20 dB. The dynamic range of the return path does not allow this due to harmonic distortions.

What if the 10 MHz carrier could somehow transmit at high power without creating harmonic distortion? In fact it is possible using a very common technique in RF power amplifiers and downstream optical transmitters, non-linear pre-distortion. The cable modem could send up a test signal that varies in amplitude that is received by the CMTS. The CMTS could measure the spectral re-growth, harmonics, AM-AM, and AM-PM nonlinear distortion produced as the CM signal level is varied and precisely characterize the non-linear channel response of the return path. This non-linear channel response could then be communicated to the CM. The CM could then apply the inverse of this response so that the distortion produced by the return path laser will cancel.

Figure 11 shows conceptually how this circuit could be implemented.

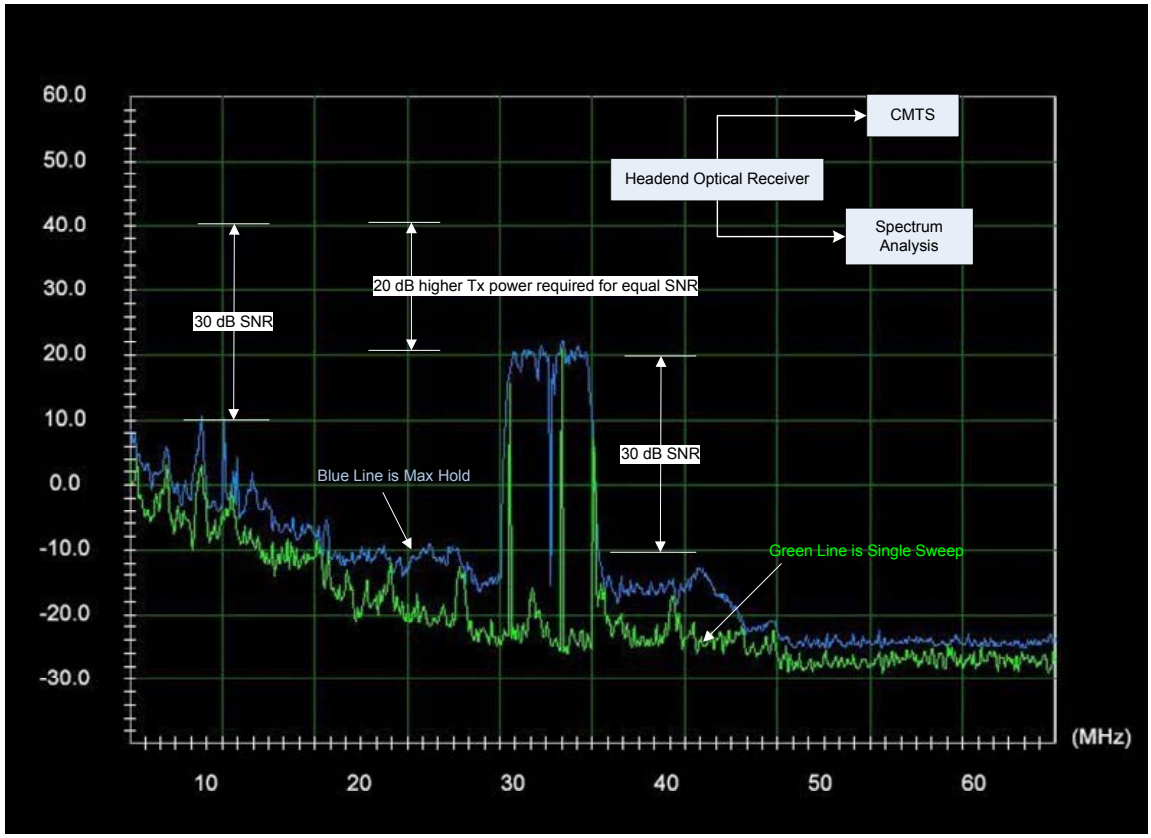


Figure 8. Field Measurement showing increased interference levels below 20 MHz

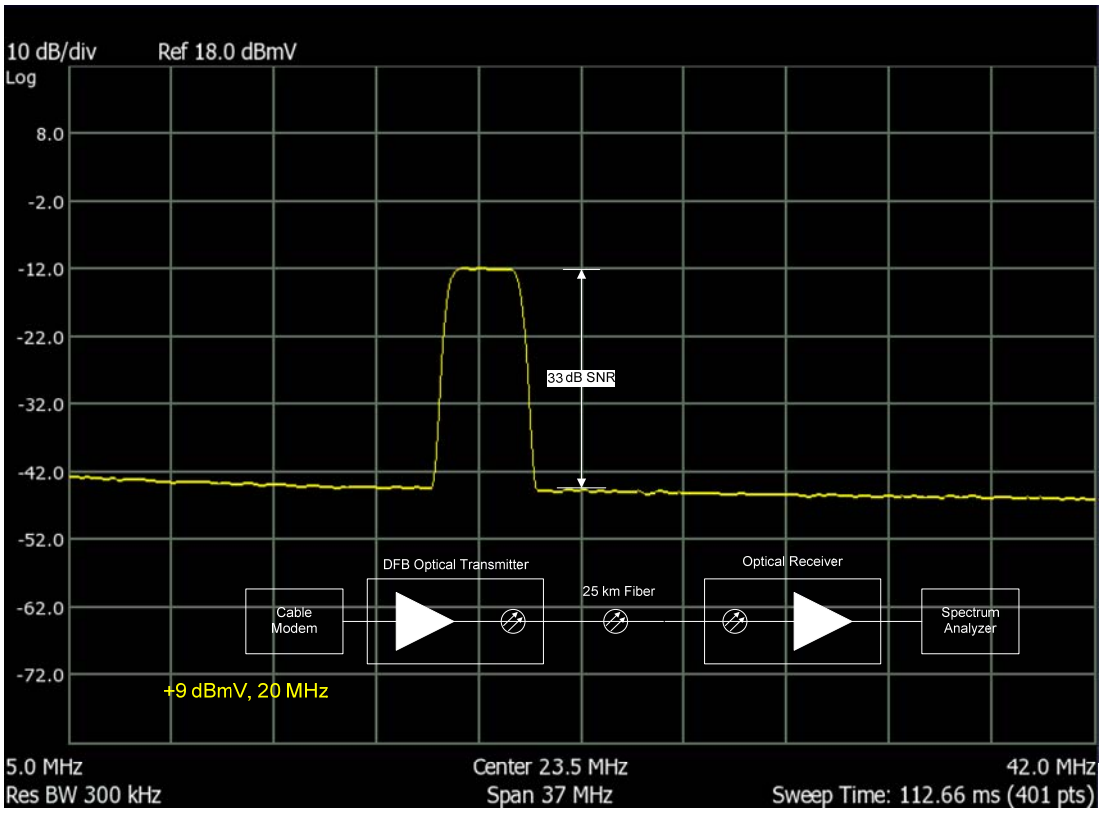


Figure 9. Lowest Amplitude Level Signal for 33 dB SNR over 25 km of fiber.

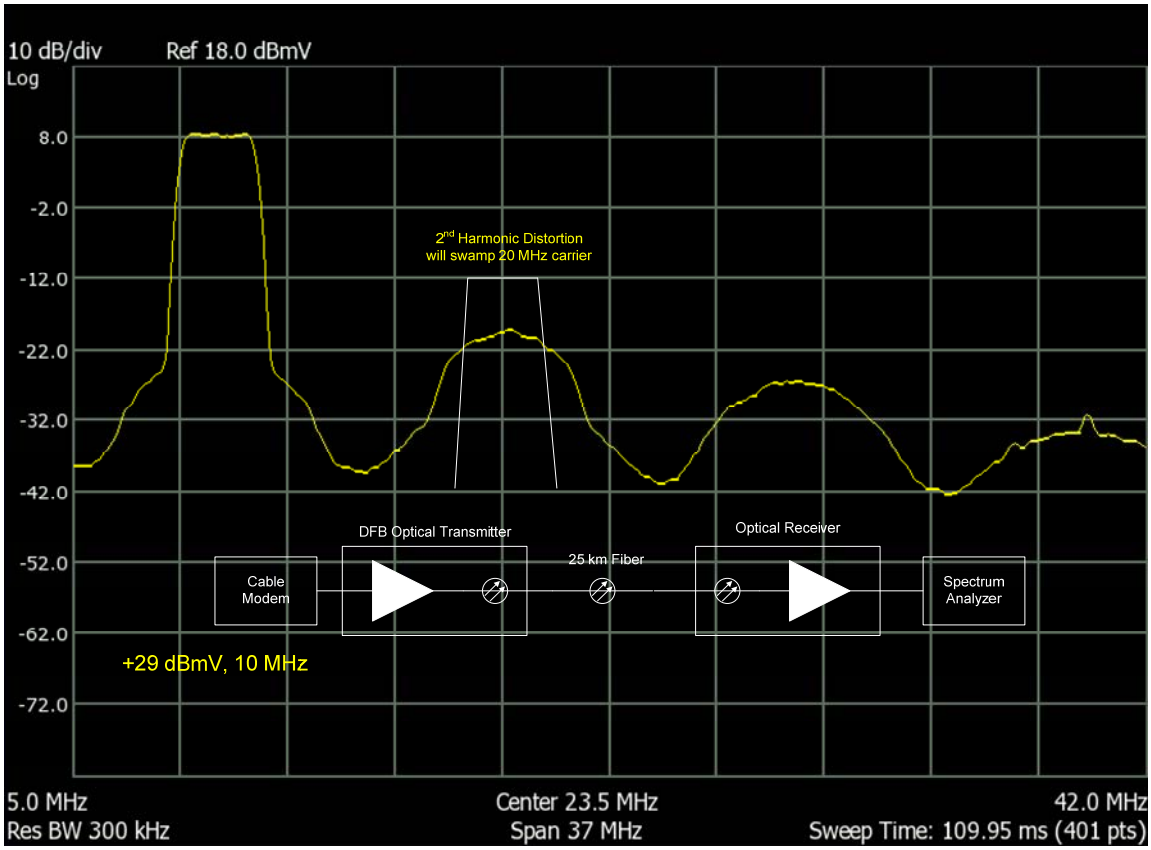


Figure 10. Increasing the level at 10 MHz by 20 dB adds significant distortion.

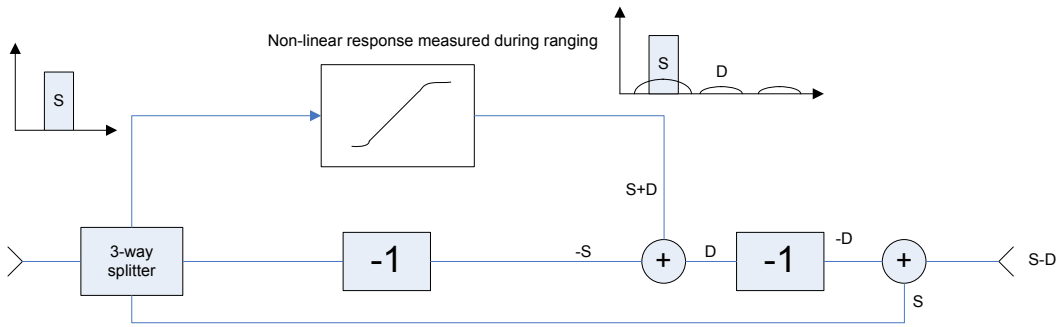


Figure 11. Conceptual Diagram of cable modem pre-distortion circuit.

This circuit could be built with common RF components, however, it is anticipated that practical implementations would perform the equivalent functions in the digital domain prior to the digital to analog converter. The signal is split with one side being subjected to the same non-linear characteristics measured by the CMTS during non-linear harmonic ranging. The output of the non-linear channel simulator would have signal, S, and distortion, D components, i.e. S+D. The signal sample from the splitter is shifted in phase by 180 degrees; equivalent to multiplying by -1, i.e. the output of the phase shifter is -S. These two signals are added, S+D-S=D, and the output is the distortion components since the signal has been cancelled. The distortion components are then shifted in phase by 180 degrees and added to another sample of the original signal. The CM output is thus S-D.

When this signal hits the return path laser, distortion components will be produced that are 180 degrees out of phase with the distortion components artificially produced in the cable modem, S-D+D=S. The distortion components cancel and only the signal is present at the output of the optical receiver.

This can be adjusted in an adaptive closed loop system so that the CMTS measures errors and adjusts pre-distortion parameters to minimize the mean squared error. By doing this the carriers at 10, 20, 30, and 40 MHz are all usable. Since a carrier can have as much as 25 Mbps data throughput, this method could potentially add as much as 25 Mbps of upstream capacity.

Simpler methods could also be used that have less benefit, but still provide substantial capacity increase. For example, by varying the CM transmit power and monitoring the harmonic distortion during a ranging process, the optimum SNR input level to the return path amplifiers and fiber node has been measured. Thus the CMTS could always direct the CM to transmit at the highest possible SNR. Since the return path characteristics change over time and temperature and there is measurement uncertainty in the manual set up of the return path input levels, the automatic optimization of the return path SNR would provide a significant improvement in SNR and thus upstream capacity.

These are just two examples of how non-linear harmonic ranging and pre-distortion can be implemented. The nonlinear response of the return path is determined by the composite signal of the superposition of all carrier waveforms from all cable modems. The non-linear pre-distortion could account for the composite waveform which would require precision control and timing from multiple cable modems and upstream carriers. The pre-distortion circuit could account for just one dominant carrier while restricting other carriers to low levels that do not impact the non-linear characteristics. The pre-distortion circuit could require that only one cable modem transmit at high power for specified time slots so that only the non-linear response from a single cable modem need be accounted for in the pre-distortion circuit. The simplest method is to just sum the average powers of all the cable modems and set the amplitude of the cable modems so that the average power into the return path laser is at the highest level and yet still in the linear operating range. More work needs to be done to make pre-distortion in the cable modem practical, hopefully these test results and ideas will serve to stimulate progress.

CHANGE IN MID-SPLIT CROSS-OVER POINT BETWEEN UPSTREAM AND DOWNSTREAM

The second technique for upstream capacity expansion is a mid-split whereby the 5-42 MHz upstream spectrum is expanded to 5-65 MHz, 5-85 MHz, or even a higher upstream to downstream crossover point. A change in the mid-split cross-over point between upstream and downstream is not compatible with the downstream transmission of analog NTSC low band VHF TV channels 2, 3, 4, 5, and 6. Analog viewers over the years have come to expect their local broadcast stations to be at a particular setting on the dial. As digital becomes the standard the downstream transmission of analog NTSC television in the low band VHF spectrum will no longer be a requirement so that the upstream spectrum can be expanded. For purposes of this paper, it is assumed that low band VHF spectrum can be reclaimed for upstream use in the future. The increase in upstream capacity is linearly proportional to the increase in allocated upstream spectrum and no changes need to be made to DOCSIS 3.0 in order to realize the capacity increase. The peak upstream speed of an individual

DOCSIS 3.0 cable modem is proportional to the number of upstream channels it can transmit. Most DOCSIS 3.0 cable modems can transmit the minimum of 4 upstream carriers and thus can reach upload speeds of 100 Mbps. More noise and interference is observed in the lower portion of the 5-42 MHz spectrum, below 20 MHz. DOCSIS upstream carriers are rarely if ever set to the lower portion of the 5-42 MHz spectrum. Narrower channel width QPSK signals for set top box return path data are often the only signals below 20 MHz, with 10.4 MHz being a commonly used frequency. In addition to adding more spectrum with a mid-split approach, the spectrum, since it is higher in frequency should have less noise and interference than that of spectrum below 20 MHz. An increase in the mid-split cross over point between upstream and downstream adds more upstream spectrum and also allows for higher order modulation due to lower levels of noise and interference.

It is important to remember that digital televisions connected to antennas, digital transport adaptors, digital cable set top boxes, and digital televisions connected directly to the cable plant at no time tune to NTSC analog television signals on channels 2-6. In Philadelphia, channel 3 is a familiar local TV station. There is no longer a broadcast over the air TV signal on channel 3 which is 60-66 MHz. When an antenna is connected to a television, the auto-program function will map the over the air broadcast ATSC signal centered at 573 MHz to TV channel 3.1. Likewise, with a digital transport adaptor or set top box, selecting channel 3 with the remote control will tune to the familiar local station because an SD stream carried on a 256 QAM signal centered at 545 MHz is mapped to channel 3. The digital transport adaptor will convert the SD stream to NTSC on EIA channel 3 or 4 for reception on an analog TV. With a digital HDTV, the customer will get a better picture if tuned to the HD stream, so as a convenience when a customer tunes to channel 3 with an HD STB, a reminder to "watch in HD" appears which if clicked will tune the TV to channel 803, the HD version of the local broadcast station which happens to be a stream in the 256 QAM RF carrier centered at 633 MHz. Only an analog NTSC television set connected directly to the cable plant utilizes the analog NTSC signal at 60-66 MHz. Thus, at a point in time whereby all customers are equipped with digital transport adaptors for NTSC analog televisions, digital televisions, and digital set top boxes then the NTSC 60-66 MHz signal will no longer be used and can be turned off without anyone being the wiser. Once this happens then a mid-split to allocate 42-85 MHz for upstream transmission will be possible.

The amplifiers in an HFC plant have a diplexer at the input and the output consisting of two filters that separate the upstream 5-42 MHz signals and the downstream 54-1002 MHz signals. Likewise, the fiber node has a diplexer at the coaxial output to separate the upstream and downstream signals. The optical receiver in the fiber node will input downstream signals into the 54-1002 MHz filter of the diplexer while the 5-42 MHz filter of the diplexer will direct upstream signals into the optical transmitter. A mid-split approach requires that all diplexers in fiber nodes and amplifiers be replaced with diplexers compatible with the chosen diplexer split, for example 5-85 MHz upstream and 108-1002 MHz downstream. Passive components in the HFC plant such as directional couplers, splitters, taps, trunk cable and drop cable as well as in home splitters and cable do not need to be changed in order to change the upstream and downstream spectrum split. Amplifiers in customer's homes will not work with a change in upstream and downstream spectrum split and thus will need to be eliminated or replaced.

A mid-split increases upstream capacity by increasing the amount of spectrum allocated to upstream transmission. Potential new upstream bands are 5-65 MHz, 5-85 MHz and 5-200 MHz. 5-85 MHz seems to be the most desirable mid-split since it provides significantly more upstream spectrum than 5-65 MHz and still allows most set top boxes to receive their forward data channel which is not the case with a 5-200 MHz split. 5-65 MHz has some advantages in that it matches the mid-split frequency used in many countries, in particular EuroDOCSIS, and it may still allow for the forward data signal of set top boxes to remain at 72-76 MHz, although this would not allow for much of a filter transition. 5-200 MHz has the advantage of providing a significant amount of upstream bandwidth; enough to potentially reach 1 Gbps upstream speeds. The disadvantage to 5-200 MHz is the significant amount of downstream spectrum lost and the loss of the forward data channel for set top boxes. Set top boxes could run in DSG mode with a 5-200 MHz mid-split. Still, all in all, the 5-85 MHz mid-split cross over point seems to be a good choice. Increasing the upstream high frequency cutoff from 42 MHz to 85 MHz provides an additional 43 MHz of upstream spectrum which is enough for six 6.4 MHz wide upstream carriers. This adds as much as 162 Mbps of upstream capacity. With four 27 Mbps upstream carriers in the 5-42 MHz band, the total upstream capacity of a 5-85 MHz mid-split is 270 Mbps.

The DOCSIS 3.0 specification includes an optional 5-85 MHz upstream mode. Adding upstream spectrum will necessarily reduce downstream spectrum. The stop band attenuation of a low pass filter is determined by the number of elements and the ratio of the stop band frequency to the pass band frequency. A 12 element Tchebyscheff low pass filter with a 0.1 dB ripple and a 42 MHz pass band will have 55 dB attenuation at 54 MHz. Likewise, a 12 element Tchebyscheff low pass filter with a 0.1 dB ripple and an 85 MHz pass band will have 55 dB attenuation at 108 MHz. Thus the downstream pass band must be changed from 54-1002 MHz to 108-1002 MHz if the upstream spectrum is changed from 5-42 MHz to 5-85 MHz in order to have comparable upstream diplexer filter requirements [1]. The downstream spectrum in the 54-108 MHz spectrum that would have to be sacrificed in order to mid-split 5-85 MHz is today used mostly for analog TV NTSC signals and set top box out of band signaling. TV channels 2, 3, 4, 5, 6, 95, 96, 97 fall in the 54-108 MHz band and 72-76 MHz is often used for SCTE-55 digital video set top box signaling out of band modulation (OOB) [3]. Analog NTSC TV signals in low band VHF likely will not be needed in the future due to digital set top boxes, digital TVs, digital transport adapters. SCTE 55-1 has a default carrier center frequency of 75.25 MHz for the Out-Of-Band channel with a footnote that includes 72.75 and 104.2 MHz OOB center frequency. So the OOB can be moved to 104.2 MHz and digital cable set top boxes can still operate after a 5-85 MHz mid-split. There may be temporary disruptions when set top boxes are first asked to change OOB frequency and there may be a small percentage of set top boxes that are fixed tuned and cannot change to a 104.2 MHz OOB center frequency.

Prior to the launch of a change in mid-split cross over point between upstream and downstream, care must be taken to protect television receiving devices that have been designed to receive signals that are now part of the upstream transmission spectrum. A change in mid-split must be executed without degradation or disruption to existing services. This requires filters or protection circuits to prevent interference. Since this is such an important issue and because it is so critical that a change in mid-split does not create interference issues, this paper will explain the mechanisms for potential interference,

provide measurement results of interference tests, and propose a specific circuit to prevent interference from upstream transmissions in the 42-85 MHz band to television receiving devices designed to receive in this band.

Cable modems and set top boxes must have diplexers at their coaxial input to allow them to receive signals in the 54-1002 MHz band and transmit signals in the 5-42 MHz band. There can be cases where a set top box can transmit a 6.4 MHz channel width signal at +54 dBmV with a 38.8 MHz center frequency while receiving a video signal at channel 2 with a 6 MHz channel width and a 57 MHz center frequency with a level of -15 dBmV. Within the same box the transmit level is 69 dB higher than the receive level. Typically, consumer electronics receivers can operate with adjacent channel levels about 10 dB higher than the desired signal, thus the diplexer must at least provide 59 dB of attenuation between the receiver input and the transmitter output for signals in the 54-1002 MHz band. The transition region between 42 and 54 MHz allows for a filter to be built that has a pass band of 54-1002 MHz and a rejection of 60 dB for 5-42 MHz signals. Since such a diplexer must be in every set top box, these devices will need to be exchanged for devices with a different diplexer split in order to benefit from the upstream capacity enhancement of a mid-split. However, it would instead be anticipated that these devices continue to use the 5-42 MHz for upstream transmission. Since in the past, the low VHF band was used for video signals and mostly NTSC analog signals, cable modems were often not designed to receive signals below 108 MHz. Thus cable modems have more relaxed diplex filter requirements. These cable modems may have filters to protect the receive chain in the 54-108 MHz range. Typical 2.0 cable modems have a downstream receive band 88-860 MHz while DOCSIS 3.0 cable modems typically have a downstream receive bandwidth of 108-1002 MHz. The upstream transmit band for both DOCSIS 2.0 and 3.0 in North America is 5-42 MHz.

Upstream transmission in the 42-85 MHz band can be shown to interfere with devices that are designed to tune to 54-108 MHz downstream signals. Such devices include televisions, digital cable set top boxes, digital transport adapters. The IF frequency of many TV tuners is 41 to 47 MHz which will be in the upstream transmit band after a change in mid-split frequency. To tune into channel 2, the voltage controlled oscillator, VCO, is set to 57+44=101 MHz and the mixer output will be SAW filtered to the lower sideband centered at 44 MHz. Thus transmitters operating in the 41-47 MHz band have to be well isolated from traditional single conversion TV tuners, if it gets into the IF chain it will result in co-channel interference. The single conversion TV RF tuner front end is illustrated in Figure 12.

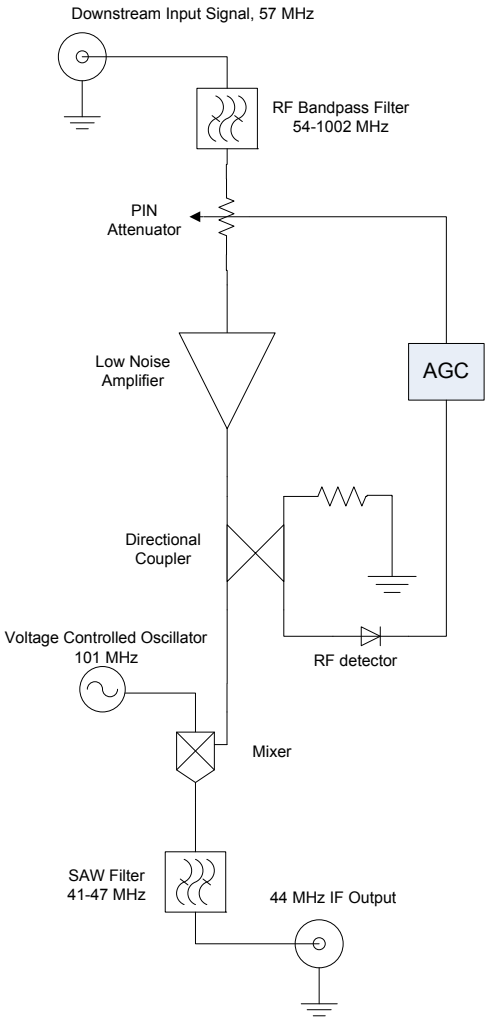


Figure 12. Block Diagram of Single Conversion TV Downconverter.

Figure 12 is helpful in understanding why mid-split upstream transmission as well as other sources of overload interference can cause picture degradation. The AGC (automatic gain control) adjusts the receiver's front end gain so that at low input levels the gain is cranked up while at high input levels the gain is set very low. This allows the front end to have high receiver sensitivity for low level signals since a dB of attenuation prior to the low noise amplifier adds a dB of noise figure to the receiver and thus degrades the signal to noise ratio by a dB. At the same time, the attenuation kicks in for high level signals so that the low noise amplifier does not get overloaded and create intermodulation distortion which also degrades the signal to noise ratio. Degradation in signal to noise ratio due to out of band high level interference arises when the diode detector circuit picks up interfering signals that are much higher than the desired signal. If the high level interfering signal is picked up by the diode detector then the AGC circuit will add attenuation. If the desired signal is very low then the attenuation prior to the low noise amplifier will degrade the signal to noise ratio of the amplifier output. This can lead to picture degradation and even to no reception at all.

A simple test was conducted whereby a cable modem was set to output at a center frequency of 55.5 MHz and a digital TV was tuned to receive a center frequency of 111 MHz, the lowest frequency downstream channel for a 108-1002 MHz downstream spectrum allocation. The mid-split frequency was chosen to be at half the desired digital video center frequency so that second harmonic components of the distortion would fall co-channel to the video signal. The mid-split signal was set to TDMA 16-QAM with a 5.12 MHz symbol rate. The digital TV showed visible macro-blocking artifacts when the upstream transmitter signal at the TV input was 27 dB higher than the desired signal. When the upstream signal was less than 26 dB higher than the desired video QAM signal then no visible distortion was observed. For interference 28 dB or more higher than the desired signal, no TV reception was possible. The spectrum analyzer plot from this test is shown in Figure 13.

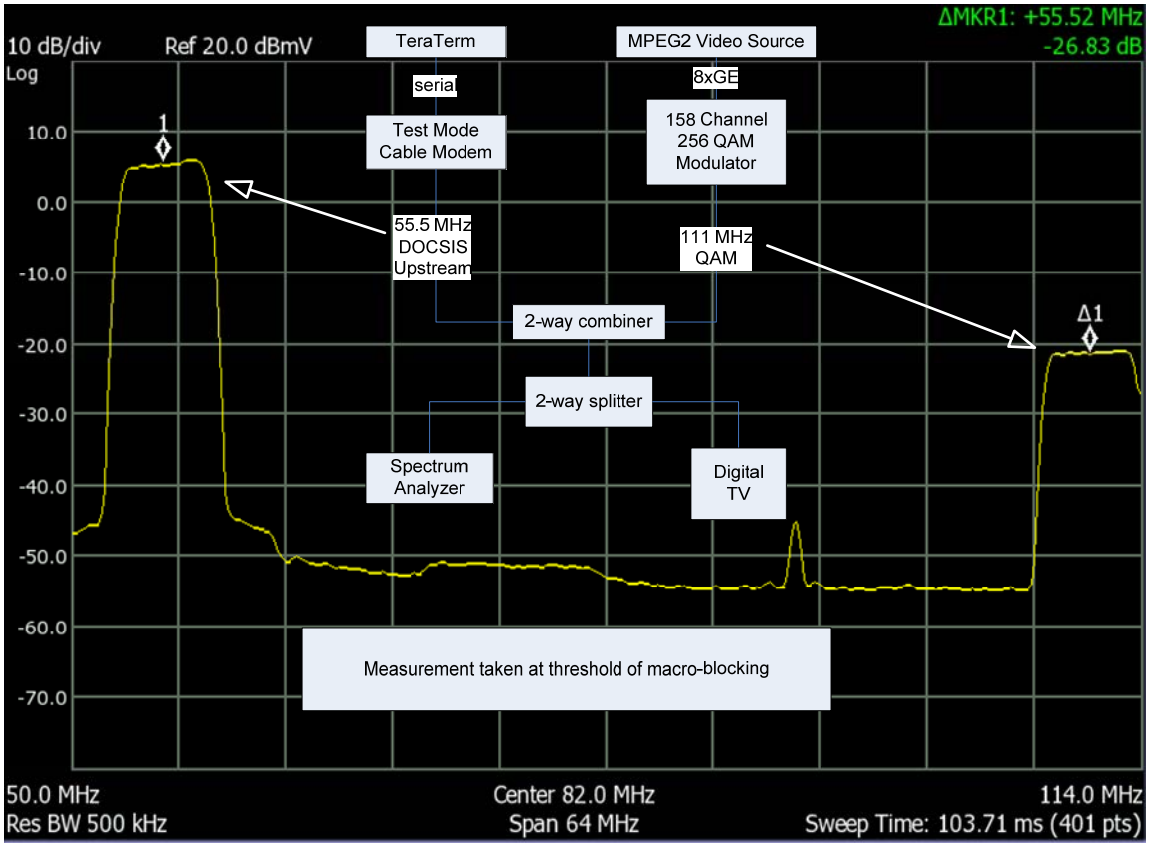


Figure 13. Interference from a mid-split cable modem to a digital television.

Another test was performed at the author's home. A set top box was tuned to a 256 QAM digital video carrier with a 97 MHz center frequency. A two way splitter was added to feed the set top box and a test mode cable modem. The test modem cable modem was set to transmit an upstream signal with 16 QAM TDMA using a 5.12 MHz symbol rate with a 48.5 MHz center frequency. It was found that a +55 dBmV upstream transmit level caused visible tiling and any level +56 dBmV or higher prevented any reception at all.

In another test a splitter was used to feed a digital transport adapter, DTA, and the test cable modem. The DTA was tuned to a 97 MHz center frequency, the input level of the 256 QAM carrier was reported in the DTA diagnostics as -18 dBmV with a 30 dB SNR. The cable modem was put in a test mode to transmit an upstream 16-QAM TDMA signal at a center frequency of 48.5 MHz and a 5.12 MHz symbol rate. An upstream transmit level of +47 dBmV was found to produce severe tiling in the picture. Transmit levels above +48 dBmV prevented any reception of the digital video signal. These tests show that DTA devices, like digital televisions and digital video set top boxes, can suffer interference from upstream transmission in the 42-85 MHz band.

Figure 14 shows the insertion loss sweep from one tap port to another for a 14 dB directional coupler and 4-way tap. The port to port tap isolation is 37 dB at 85 MHz. A 100 foot drop cable of RG-11 has a loss of about 1 dB at 55 MHz so about 2 dB of loss can be expected for a mid-split upstream signal traversing a 100 foot drop from home to tap and another 100 foot drop from tap to another home. If one home is transmitting in

the mid-split upstream band at +58 dBmV with 2 dB of cable loss and 37 dB port to port tap isolation, then the input level of the upstream transmission is +19 dBmV at the neighbor ground block. With downstream video levels of -8 dBmV the difference between the upstream transmission and the downstream receive level in the neighbor's home is 27 dB which has been show to impact video picture quality. Home to home isolation needs to be considered to protect television receiving devices with upstream transmission in the 42-85 MHz band.

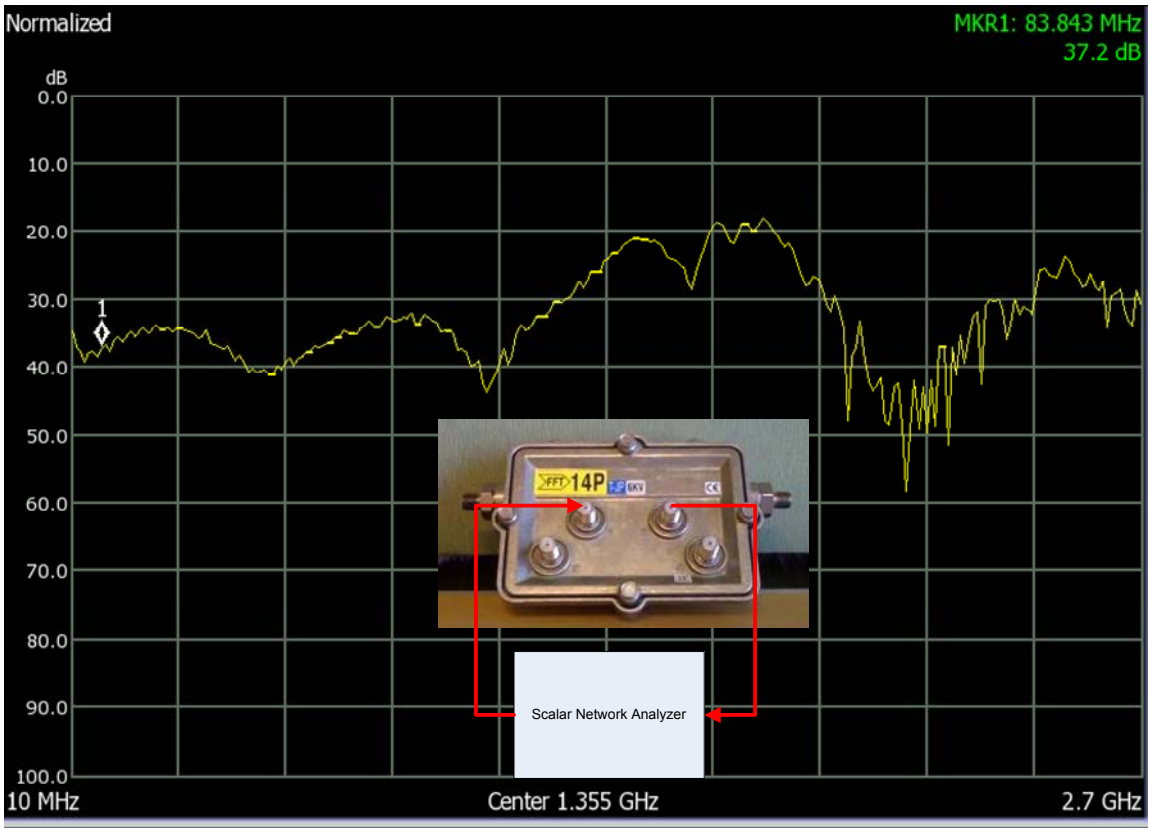


Figure 14. Tap to Tap Isolation measured for a 14 dB Coupler.

An illustrative block diagram of a mid-split protection circuit is shown in Figure 15. The upstream and downstream signals from the HFC plant are split with a diplex filter having an upstream band of 5-85 MHz and a downstream band of 108-1002 MHz. The mid-split RF protection circuit has a DOCSIS 3.0 Cable Modem, in this case with the capability of bonding 16 downstream channels and 4 upstream channels. The four upstream channels would most likely fall within the 42-85 MHz spectrum band so that upstream capacity for the mid-split protection circuit is additive to upstream capacity of cable modems within the home transmitting in the 5-42 MHz band

The mid-split RF protection circuit physical connector to the home coaxial cable is fed with a diplex filter separating the downstream 108-1002 MHz signals and the 5-42 MHz upstream signals from devices in the home such as SCTE-55 STBs, DSG STBs, CMs, and eMTAs. In this case an optional downstream amplifier with AGC is

added so that the home network is always fed with an optimal downstream signal level. Likewise, the diagram includes an optional upstream amplifier to help with high attenuation cable modems, particularly DSG STBs. The 5-42 MHz diplex filter and the upstream amplifier provide high isolation between the mid-split upstream signals in the 42-85 MHz range and the in home devices that could suffer interference from such signals.

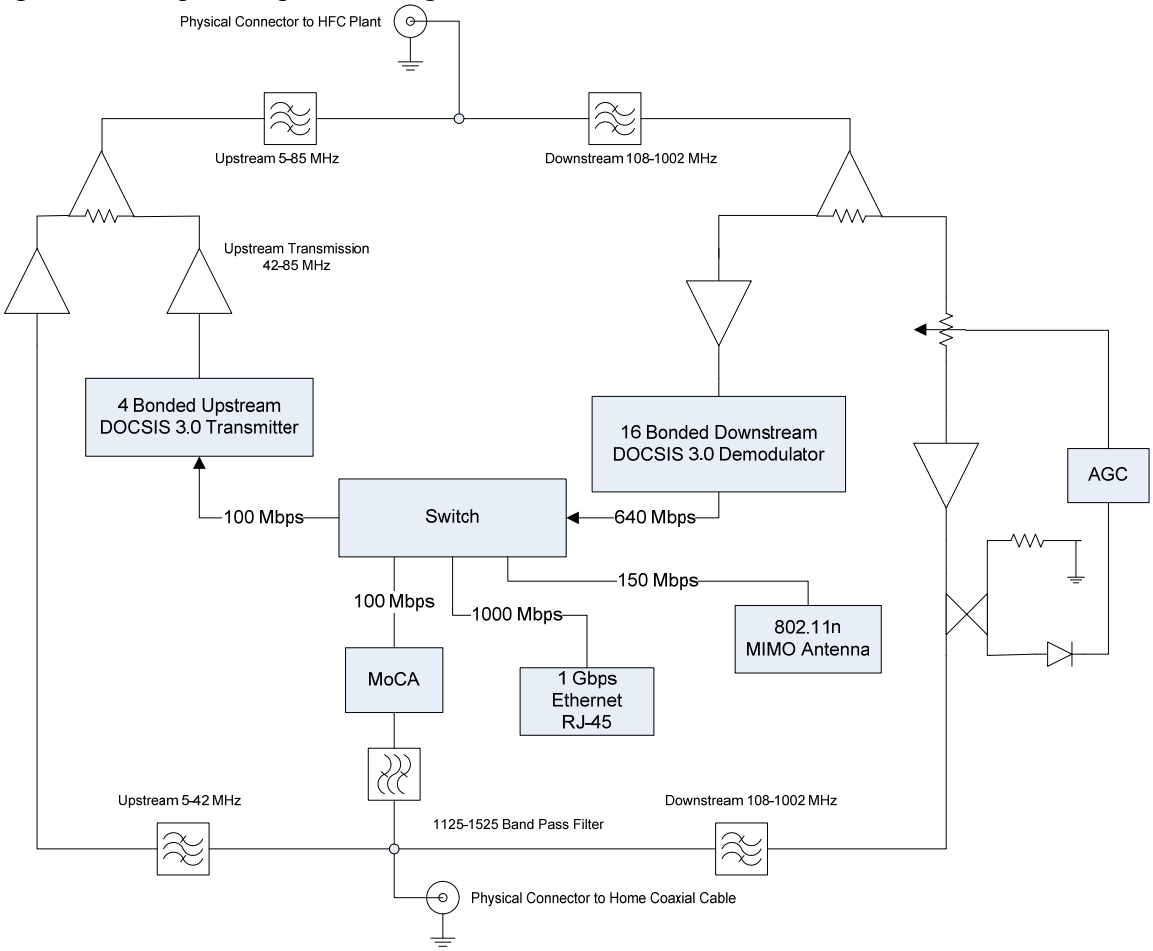


Figure 15. Mid-Split RF Protection Circuit between the HFC RF and home RF

INCREASING UPSTREAM CAPACITY BY USING SPECTRUM ABOVE 1 GHZ

Trunk cable and drop cable can support frequencies in the 1-3 GHz range at reasonable attenuations. Fiber nodes, amplifiers, directional couplers, splitters and taps in the plant will need to be modified to support the use of spectrum above 1 GHz signals. In home coaxial cables, splitters, and amplifiers add too much attenuation for practical use of signals in the 1-3 GHz range traveling all the way from the cable headend to the end user device within the home. Figure 16 shows a divided network architecture for using spectrum above 1 GHz with a fiber optic cable section, a 1200-1800 MHz over trunk and drop coaxial cable, and an in home coaxial MoCA section. The fiber node and amplifiers are bypassed with fiber optic cable to the last active. The directional couplers and taps between the coaxial 1200-1800 MHz transceivers are replaced with units supporting these frequencies. A coaxial transceiver is placed at the last active and at

the customer entrance. The 1200-1800 MHz signals are terminated at the customer entrance and transported over home networking technology such as 1 Gbps Ethernet, MoCA, or 802.11n so that in home wiring does not need to support the high spectrum signals.

Trunk cable 750 feet in length will have about 25 dB of loss at 1500 MHz with a variation from 1200 to 1800 MHz of about 5 dB. Drop cable of 214 feet length will have a loss of about 17 dB with a variation from 1200 to 1800 MHz of about 3 dB. In the case where the path from the last active to the home includes a 3 dB splitter and two directional couplers with 2 dB of through loss and an 11 dB tap, the total attenuation is 60 dB using the trunk and drop length above. If the transmit power is +60 dBmV then the receive level is 0 dBmV which is equal to -49 dBm. If the receiver noise figure is 4 dB then since the thermal noise at room temperature is -174 dBm/Hz, the SNR is 34 dB using a 500 MHz equivalent noise bandwidth. This is a high enough signal to noise ratio for 64-QAM which has a spectral efficiency of 6 bps/Hz. Allowing for forward error correction and overhead, a final spectral efficiency of 4 bps/Hz seems reasonable. With a 500 MHz bandwidth and 4 bps/Hz spectral efficiency, the total channel capacity is 2000 Mbps. Using time division duplexing, 1 Gbps can be allocated for downstream and 1 Gbps can be allocated for upstream. The 500 MHz equivalent noise bandwidth can fit in the 600 MHz channel width within the 1200-1800 MHz band with reasonable OFDM guard bands on either side.

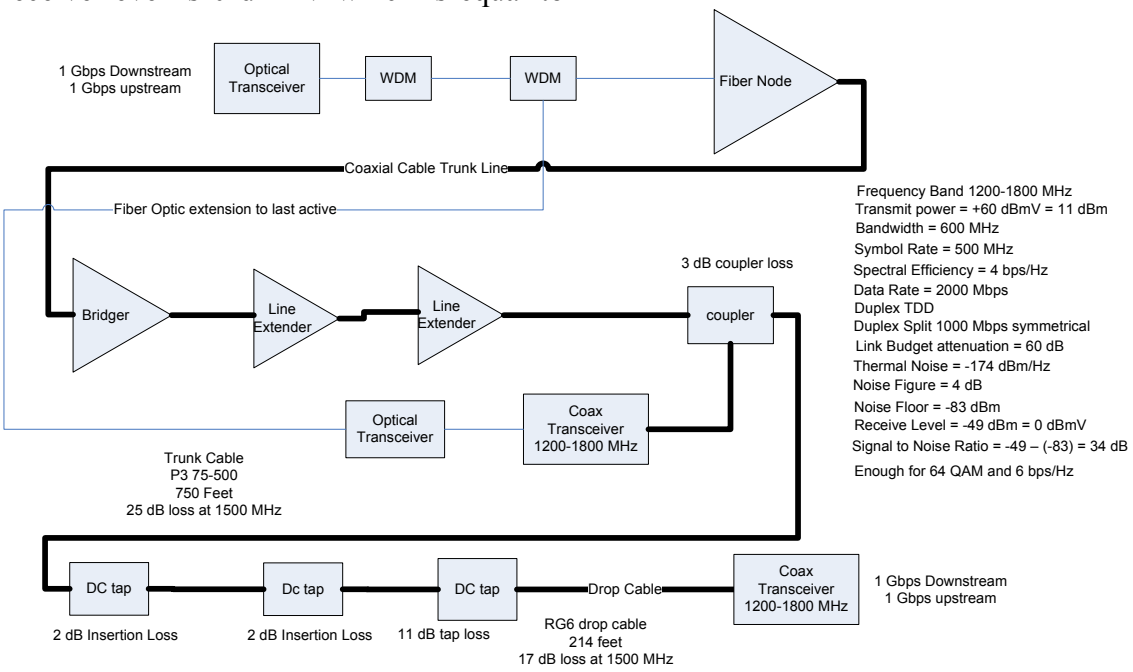


Figure 16. Block Diagram of 1 Gbps over coaxial cable network architecture.

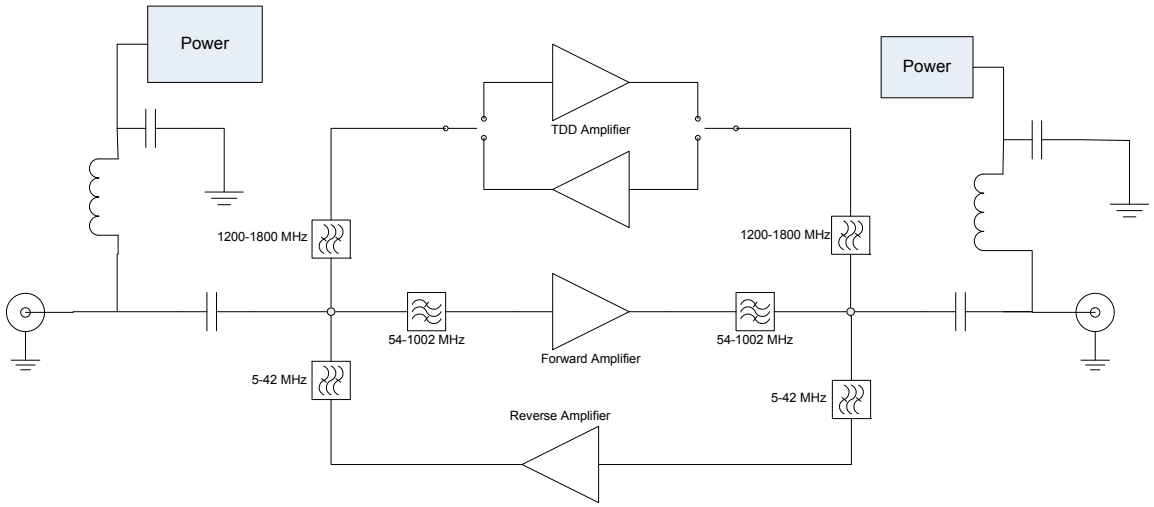


Figure 17. Block Diagram of TDD Triplexer to support HFC and high spectrum operation.

Further work needs to be done to better understand the actual channel characteristics above 1 GHz over the coaxial plant. Field measurement and lab measurements are underway to develop channel models to use in evaluating proposed system solutions. Data is needed on interference such as harmonics, MoCA, radar, and radio communication networks. Reflections and multi-path as well as frequency notches due to impedance imperfections at high frequencies need to be accounted for in evaluation.

The fiber optic cable does not always need to be run all the way to the last active. 2-3 GHz spectrum could be used for point to point backhaul to eliminate the need for running the fiber all the way to the last active. Another method is to use amplifiers as shown in Figure 17. In order to inject 1200-1800 MHz TDD prior to the last active, new amplifiers will have to replace the old amplifiers between the 1200-1800 MHz transceivers. The new amplifiers will require triplex filters at the input and output and an additional TDD amplifier for 1200-1800 MHz signals.

CONCLUSION

Today, a common upstream frequency plan includes a 6.4 MHz channel width 64-QAM modulation carrier centered at 34.2 MHz and another 3.2 MHz channel width 16-QAM carrier centered at 37 MHz. The total upstream capacity is about 36 Mbps which is about 23% of the potential capacity in the 5-42 MHz band when compared to the full spectrum being filled with 64-QAM carriers. By segmenting nodes with four upstream optical transmitters, the upstream capacity can be increased another four times. Thus, within the 5-42 MHz upstream band increased upstream capacity demands of 16-fold can be addressed. Beyond that, new spectrum for upstream carriers further increases capacity.

The logical first step to increase upstream capacity is to increase the spectral efficiency and utilization of the 5-42 MHz. Better utilization requires the use of more upstream carriers which in turn requires the use of noisier parts of the 5-42 MHz band. The objective is to increase the upstream capacity and at the same time improve the reliability. Placing upstream carriers in noisier parts of the spectrum risks reducing the reliability and availability of upstream signals, to deal with higher levels of noise and work with higher reliability and availability substantial

improvements are required in the upstream signal robustness. Three methods have been described in this paper that will work with OFDMA systems as well as S-CDMA systems. The first is adaptive modulation and coding with real time channel measurement, with S-CDMA this can be accomplished by allowing the number of active codes to be reduced during noise bursts. The second method is hybrid automatic repeat request which uses very little error correction while seeking acknowledgement of successful packet transmission, when packet transmission is unsuccessful then the upstream transmission in repeated. Hybrid ARQ works well for the intermittent and unpredictable noise characteristic of the HFC 5-42 MHz upstream path. Third, since a fundamental determinant of the upstream channel signal to noise ratio is the dynamic range limits and changes over time and temperature of the return path optical transmitter and upstream amplifiers, a non-linear harmonic ranging technique with pre-distortion could be investigated further as a means to increase the upstream capacity.

If at a point in the future the 5-42 MHz upstream appears to be headed towards full utilization, then a mid-split to 5-85 MHz will add significant upstream capacity with a small sacrifice in the downstream. The diplexers in fiber nodes and amplifiers will have to be exchanged for diplexers compatible with the new upstream to downstream spectrum split. So a mid-split will require extensive work to the HFC plant. The out of band modulator downstream signaling carrier for digital video set top boxes in most cases will need to change center frequency and televisions, set top boxes, digital transport adapters will have to be protected against upstream transmission in the 42-85 MHz band. A mid-split RF protection circuit that will be the only device to transmit in the 42-85 MHz upstream band with filters and amplifiers to protect devices in the home from interference is a good approach for mid-split upstream capacity enhancement.

Finally, for incremental addition of 1 Gbps symmetrical services, primarily for large businesses, office parks, and multi-dwelling units, a frequency overlay approach using spectrum above 1 GHz has been introduced. While this requires bypassing the fiber nodes and amplifiers with a fiber overlay or replacing the fiber nodes and amplifiers with devices having above 1 GHz capability, this can be done on an as needed basis without disruption to the existing HFC infrastructure and the services running over the HFC infrastructure. More study and most importantly field measurements are needed to better characterize the interference levels and the channel characteristics of the HFC plant modified to support frequencies above 1 GHz. This paper has tried to begin this process by including link budget, bandwidth and duplexing scheme based upon insertion loss sweeps above 1 GHz.

In conclusion, the HFC plant has a healthy upstream capacity with room to grow within the 5-42 MHz band, further room to grow in the 5-85 MHz band, and even more capacity growth is possible above 1 GHz.

Glossary of Terms and Abbreviations:
HFC Hybrid Fiber Coax a network architecture that transports upstream RF signals over fiber optic cable over long distances and then coaxial cable over short distances to homes and businesses.
CPE Customer premise equipment typically refers to digital cable set to boxes, televisions, cable modems, home routers, digital voice adapters.
CM cable modem
CMTS cable modem termination system
eMTA embedded media transport adapter for cable modem voice over IP
OFDMA orthogonal frequency division multiplexing a technique that divides an RF channel into many small frequency segments

to realize protection from multi-path interference
S-CDMA synchronous code division multiple access a DOCSIS 2.0 and 3.0 upstream technique that assigns 128 orthogonal codes which can be synchronously transmitted and detected over the same RF frequency without interference, multiple cable modems can share an upstream channel by the assignment of unique orthogonal codes.
Digital Transport Adapter, DTA, device that receives digital 256 QAM signals and converts them to channel 3 or 4 NTSC analog signals.
TDD, time division duplex, using a single frequency band shared for both upstream and downstream transmission separated in time.
FDD, frequency division duplex, using a separate upstream and downstream frequency band.
DC directional coupler, part of HFC tap or stand alone to couple upstream and downstream signals to the truck cable.
DTAB discrete tone adaptive bandwidth a method introduced in the paper for increasing HFC upstream capacity.
WDM wavelength division multiplexing allows multiple signals to be sent on the same fiber optic cable using different wavelengths
Hybrid ARQ automatic repeat request a technique to use acknowledgements to repeat upstream transmissions while adding error correction as needed.
NPR noise power ratio, the return path laser is tested with a noise source having a notch, the input level is varied and the ratio of the power spectral density of the input signal to the distortion components at the notch frequency is measured.
OOB Out-Of-Band downstream signaling for digital cable set top boxes SCTE-55 [3]
MoCA multimedia over coax a standard for home networking over coaxial cable.
802.11n the latest WiFi standard for wireless home networking.

Acknowledgements:
The author would like to thank Doug Jones, Sam Chernak, Jorge Salinger, Wayne Davis, Dave Feldman, Rob Howald, Brent Arnold, Ross Gilson, Rick Gasloli, Chris Albano for help with this paper.

References:
[1] Matthaei, G., Young, L., Jones, E.M.T., Microwave Filters, Impedance-Matching Networks, and Coupling Structures, McGraw-Hill, 1964, pp. 88.
[2] Chang, P., Howald, R.L., Thompson, R., Stoneback. D., Rathod, V., Moore, C., "Characterizing and Aligning the HFC Return Path for Successful DOCSIS 3.0 Rollouts", SCTE Cable-Tec Expo, Denver, Colorado, October28-30, 2009.
[3] Society of Cable Television Engineers, ANSI/SCTE 55-1 2009, Digital Broadband Delivery System: Out of Band Transport Part 1: Mode A
https://www.scte.org/documents/pdf/Standards/ANSI_SCTE-55-1-2009.pdf

CONSIDERATIONS FOR A REAL TIME BROADBAND DELIVERED INTERACTIVE GAMING SERVICE

Charles H. Jablonski
VP Operations, OnLive Inc

Abstract

Real Time graphics intensive applications, such has "first person shooter games" have presented challenges in being server/cloud based because of the inherent roundtrip latencies.

This paper will describe an approach that provides the ability not only to deliver real time games but also that is able to deliver any graphics intensive application over existing broadband infrastructure.

The approaches taken to provide the necessary low latency compression, server and data center architecture, integrating into exiting carrier and ISP data infrastructures as well as the impacts of "last mile" ISP (DSL and Cable) considerations as well as the required performance of the date delivery, packet loss and jitter specifications of the infrastructure.

An approach using existing PC/Mac client for delivery of the service to end consumers will be also be described as well as a small stand alone low cost (<$30) terminal adapter that connects a consumer television directly to a internet connection to enable game play.

The system allows a variety game content, media and enterprise application to also be provisioned and delivered over the same infrastructure.

.

Introduction

The march of enabling technology and the continued improvement in broadband connectivity and availability have enabled many services and products to be delivered to end consumers and enterprises as various compressed data and media files. The success of services such as video on demand and "ultra" high speed broadband are well known and their effects have caused consumer tastes and expectation to evolve accordingly.

In few short years linear packaged media has evolved from a shrink wrap retail point of sale product to a anywhere, anytime, almost any device availability. At times the business issues have been far more daunting that the technological, but one undisputed fact is that the consumer continues to be more demanding and the expectations of availability, selection, schedule and independence increases.

Further the same technologies that have enabled these distribution advances have lead to a far greater availability of consumer unilateral and multilateral consumptions platforms (PCs, Net books, Tablets, Portable Devices) and the ability to imbed the functionality into broader consumer platforms, and inexpensive stand alone devices.

One of the last frontiers remaining is true real time interactive consumer experiences in the home, using inexpensive and available end devices, that are provisions by centralized "cloud based" systems. In these applications, such as gaming, the consumer experience

cannot be fulfilled unless the challenges of latency, quality and availability using cloud based systems and broadband networks are met.

OnLive is one such system, and it is believed the first among many that will be taking advantage of the broadband infrastructure to widely deliver these interactive real time applications.

Business Issues and Consumer Expectations

As with any good technological innovation, there must be an underlying business reason to support the development and implementation and the implementation must be compelling enough for all parties to embrace it. Science experiments, although intellectually fulfilling, must have the support and resources of a real business problem to evolve into broad and accepted use.

The gaming industry, although a large mature business (>$50B annually worldwide), the distribution of the end product to consumers has been dominated by retail point of sale. Even though the economics of game development resemble other media (high development costs, intensive consumer promotion and advertising, short selling season and short initial release life) the actual end distribution is the last unfulfilled digital delivery for modern media.

Typically many scores of millions are spend developing a game, a few score spent in advertising its release, and then the selling season is typically a few weeks before the holidays. This is all delivered through shrink wrap product that is sold through retail distribution channels. These channels suffer from various impeded costs (duplication, distribution, platform fees, retailer margins) as well as theft, returns and "spoilage". Also since the end product is in the consumers hands piracy is a constant battle, and >30% of the games are resold through the used market where the initial developer and published receiving none of that incremental revenue. Some high end titles are resold and used 4-5 times (as evidenced by registrations to the publishers).

This "quantized" method of sale also makes incremental releases, episodic, upgrades extremely problematical in the games market. Although there is a strong online community and infrastructure its broad use is limited to small add-ons, consumer social features test and chat. What few online sales mechanisms in existence they either suffer from extremely long and large downloads, or they are very limited in their catalog of current first release titles.

Typically software publishers receive less than 50% of the MSRP at the retailer point of sale. Clearly there is room for improvement in the economics of the existing system. Now, when one examines the exiting consumer experience for high end interactive games, that picture has room for improvement as well.

Typically a consumer has to spend $500 (in early cycle) on a proprietary closed gaming console, that by definition and practice will require replacement in 3-5 years, purchase games at retail, sometimes having to wait in line or just wait for availability of a hot game, spend additional monies to participate in a social network for that console. The other choice is to purchase an extremely high end PC (>$5000) and suffer the same issues acquiring the games and software.

Naturally with retail purchase model the end consumer may not have immediate access to new "hot" games, and be in the position on not knowing the game experience until it has been purchased brought home, installed and used.

As with the publisher side of the equation, this systems is ripe for a disruptive change and consumers continue to expect to enjoy real time availability and experience of all other digital media, an online, cloud based gaming system that addresses these business and consumer needs can be a commercial success.

The Consumer Experience

In order for a server based, cloud served system to be successful; it must meet or exceed the existing experience expectations for current consoles and PCs. What does that incorporate?

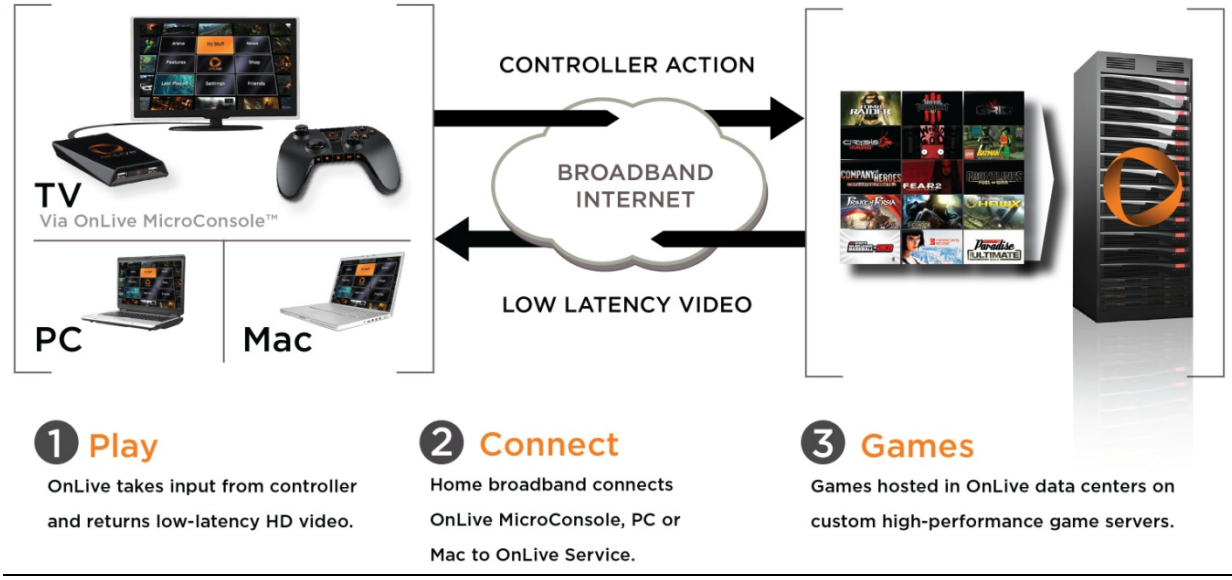
One the image quality must be the same. That requires the ability to render, in real time, high definition, high frame rate images. The response time for the system, using readily available controllers, must be fast enough not to limit the game play because of lag or latency, a large amount of amount devices must be available and supported, be deployable on existing PCs or Macs, it must be extendable to a large screen display device (i.e. TV), be enabled and installed simply over exiting broadband infrastructure, afford social infrastructure and features and have wide availability of new release games from various publishers.

Up until now, the key limitation in existing technologies has been the latency of real time video compression systems. These existing technologies (JPEG, MPEG et.al.) have either required bandwidth only available on large corporate LANs, or required many milliseconds to compress real time images. (Typically game play requires <100ms of response time to be effective in competitive game play)

The OnLive Platform

The following is a brief description of the OnLive platform:

When the user performs an action on a computer or TV connected to OnLive (e.g. presses a button on controller or moves a mouse) that action is sent up through the internet to an OnLive data center and routed to a server that is running the game the user is playing (or the application the user is using—since the interactive demands for video games are generally higher, remote video game operations will be primarily described in the following paragraphs, but these discussions are entirely applicable to remote application operations).



OnLive Platform

The game computes the next video frame based on that action, then a proprietary chip compresses the video from the server very quickly, and the user's PC, Mac or OnLive MicroConsole™ decompresses the video and displays the new frame of video on the user's computer display or TV set. The entire round trip, from the point the button is pressed to the point the display or TV is updated is so fast that, perceptually, it appears that the screen is updated instantly and that the game is actually running locally.

When the user performs an action on a computer or TV connected to OnLive (e.g. presses a button on controller or moves a mouse) that action is sent up through the internet to an OnLive data center and routed to a server that is running the game the user is playing (or the application the user is using—since the interactive demands for video games are generally higher, remote video game operations will be primarily described in the following paragraphs, but these discussions are entirely applicable to remote application operations). The game computes the next video frame based on that action, then a proprietary chip compresses the video from the server very quickly, and the user's PC, Mac or OnLive MicroConsole™ decompresses the video and displays the new frame of video on the user's computer display or TV set. The entire round trip, from the point the button is pressed to the point the display or TV is updated is so fast that, perceptually, it appears that the screen is updated instantly and that the game is actually running locally.

The key challenge in any cloud system is to minimize and mitigate the issue of perceived latency to the end user.

Latency Perception

Every interactive computer system that is used, whether it is a game console, a PC, a Mac, a cell phone, or a cable TV set-top box, introduces a certain amount of latency (i.e. lag) from the point you perform an action and you see the result of that action on the screen. Sometimes the lag is very noticeable (e.g. on some TV set-top boxes it takes over a second to move a selection box in a program guide). Sometimes it isn't noticeable (e.g. if you have a well-designed game running on fast hardware, and pressing the fire button results in what appears to an instantaneous display on your screen of the your gun firing).

But, it's important to note that, even when your brain perceives game response to be "instantaneous", there is always a certain amount of latency from the point you perform an action and your display shows the result of that action. There are several reasons for this. To start with, when you press a button, it takes a certain amount of time for that button press to be transmitted to the computer or game console (it may be less than a millisecond (ms) with a wired controller or as much as 10-20 ms when some wireless controllers are used, or if several are in use at once). Next, the game needs time to process the button press. Games typically run between 30 and 60 frames per second (fps), so that means they only generate a new frame every $1/30^{th}$ to $1/60^{th}$ of a second (33ms to 17ms). (Further, when games are generating complex scenes, sometimes they take longer.) So, even if the game responds right away to a button action, it may not generate a frame for 17-33ms or more that reflects the result of the action. And, then finally, there is a certain amount of time from the point the game completes generating the frame until the frame appears on your display. Depending on the game, the graphics hardware, and the particular monitor you are using, there may be almost no delay, to several frame times of delay. And, if your game is an online game, there typically will be some delay to send a message reflecting your action through the internet to other game players, and the game

may (or may not) delay the action occurring in your game so as to match your screen action to that of screen action of players who are playing the game remotely. So, in summary, even when you are running a game on a local machine there is always latency.

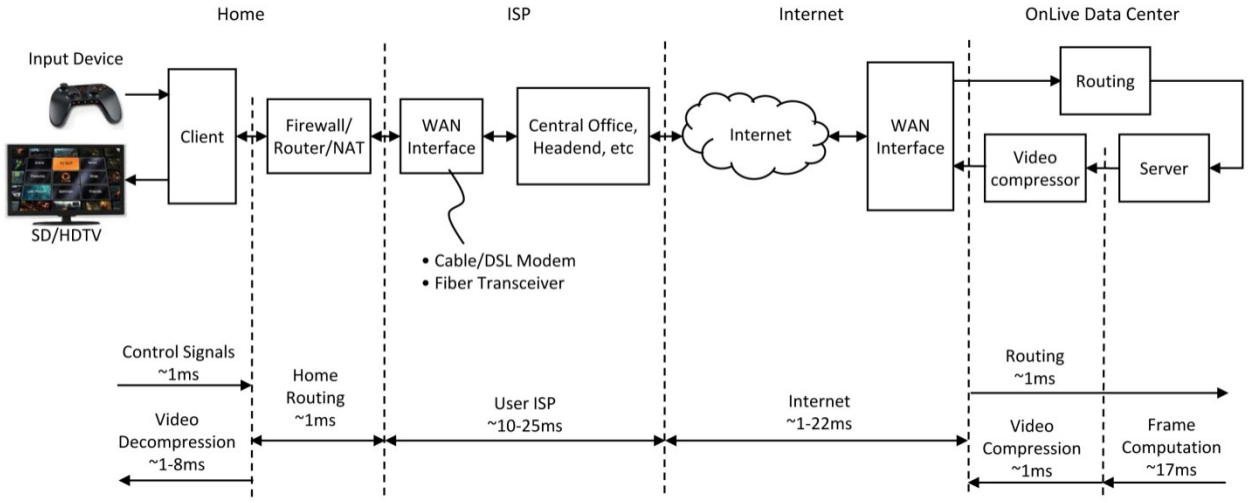
The question is simply how much latency. So, while there certainly are more subtleties to the perception of latency, as a general rule of thumb, if a player sees a fast-action game respond within 80ms of an action, not only will the player perceive the game as responding instantaneously, but the player's performance will just as good as if the latency was shorter.

And, as a result, 80ms is the "latency budget" needed to meet for the OnLive system to be practical. That is to say, OnLive has up to 80ms to: send a controller action from the player's home through the internet to an OnLive data center, route the message to the OnLive server running the game, have the game calculate the next frame and output the video, compress the video, route the compressed video out of the data center, send the compressed video to the player's home through the internet, decompress the video on the players computer or and output the video to the player's display. And, of course, OnLive has to do this at rate of 60fps with HDTV resolution video over a consumer internet connection, running through consumer internet gear in the home.

Over Cable and DSL connections, OnLive is able to achieve this if the user's home is within about 1000 miles of the OnLive data center. So, through OnLive, a user who is 1000 miles away from a data center can play a video game running on a server in the data center with the perception (and the game play score) as if the game is running locally.

OnLive's Latency Calculations

The simplified diagram below shows the latencies encountered after a user's action in the home makes it way to an OnLive data center, which then generates a new frame of the video game and sends it back to the user's home for display. Single-headed arrows show latencies measured in a single direction. Double-headed arrows show latencies measured roundtrip.



OnLive System Latency

There latency numbers shown here are numbers that OnLive has seen in practice, given the way the OnLive system was architected and optimized, and reflect what has been measured after using OnLive in various locations over the years. If you add up all of the worst-case numbers, it shows the latency can be as high as 80ms. That said, it is highly unlikely that every segment will be worst case so the total latency will likely be much less (and indeed, that is what we see in practice).

ISP latency

Potentially, the largest source of latency is the "last mile" latency through the user's Internet Service Provider (ISP). This latency can be mitigated (or exacerbated) by the design and implementation of an ISP's network. Typical wired consumer networks in the US incur 10-25ms of latency in the last mile, based on OnLive's measurements. Wireless cellular networks typically incur much higher last mile latency, potentially over 150-200ms, although certain planned 4G network technologies are expected to decrease latency.

Within the internet, assuming a relatively direct route can be obtained, latency is largely proportional to distance, and the roughly 22ms worst case round-trip latency is based on about 1000 miles of distance (taking into account the speed of light through fiber, plus the typical delays OnLive has seen due to switching and routing through the internet.

Consequently, OnLive will be locating its data centers such that the distance to most of the US population is less than 1000 miles
The compressed video, along with other data required by the OnLive client to keep it tightly sync'd with the OnLive service, is then sent through the internet back to the user's home. Notably, the data generated by the video compressor is carefully managed to not exceed the data rate of the user's internet connection because if it did, that might result in queuing of packets (incurring latency) or dropped packets. Since the user's home data rate is constantly changing, the OnLive service is constantly monitoring the available data rate, and constantly adapting the video compression (and if necessary, dropping the video resolution) to stay below the available data rate.

One common misconception about home broadband connections are that the latency is directly tied to data rate (i.e. the effective connection speed) and/or data throughput (i.e. the data rate available to a particular user). Latency is actually largely independent of data rate, so long as the data throughput demands are less than the capacity of the broadband connection.

OnLive video decompression latency

Once the compressed video data and other data is received by the OnLive client (i.e. the OnLive application running as a plug-in or standalone in your PC or Mac, or the OnLive MicroConsole attached to your TV), then it is decompressed. The time needed for decompression depends on the performance of your PC or Mac (CPU and frame buffer bandwidth…no GPU is needed), and may vary from about 1 to 8ms. If your computer's CPU and/or memory bus is tied up doing another processing-intensive task or if you have an extremely low performance computer, OnLive may find it is unable to decompress video at full screen resolution. If so, then it will scale down the video window accordingly. But, we have found most computers made in the last few years work fine up to their screen resolutions so long as they are not tied down running some other intense application at the same time. In any case, even if you are in a processing-constrained situation, OnLive will select a

video frame size which will maintain low latency.

OnLive round-trip latency

As mentioned before, while there is a certain amount of latency variability in each leg of the journey, it is rare that a given user will end up in a worst-case scenario with each leg. Consequently, what we typically see in practice are latencies on the order of 40 to 60ms. Sometimes we see latencies that are higher and sometimes we see latencies that are shorter. And, we expect latencies to continue to decline as "last mile" infrastructure is upgraded, both for wired and wireless networks.

Video Quality

As previously discussed video quality is paramount to successful consumer experience. Not only does the image quality have to be acceptable, the rendered frame rate must also be high to provide the gaming experience

During high action game segments, stereo audio

Resolution (p60)	Aspect Ratio	Peak Per Stream[1] Data Rate (Mbps)
480 x 270 (cell)	16:9	0.6
640 x 360 (SDTV[2])	16:9	1.2
1280 x 720 (HDTV)	16:9	4.0
1920 x 1080[3] (HDTV)	16:9	8.0

OnLive Data Rates

The 1280 x 720 resolution currently provided, provides adequate visual experience, and with a peak data rate at 5Mps allows the service to be widely available to consumers through existing broadband infrastructure.

Consumer Infrastructure

As described earlier the OnLive system has been engineered to be deployable on a wide variety of PCs and Macs using a downloadable software client. As for the large screen device, OnLive has developed its own consumer TV adapter, the MicroConsole™.

MicroConsole TV Adapter

The OnLive MicroConsole™, which will be generally available later this year, has been developed to allow simple integration into a home consumer display. The device has an standard NIC connector for access to the broadband service, an HDMI adapter to connect directly to the large screen device to provide high image quality as well as multichannel audio. (A high quality audio experience is another "must have" for a successful consumer gaming experience.)

There are USB connectors to allow integration of standard controllers and devices, wireless radios for headsets, and an optical audio connector. The cost of this device is minimal to allow wide deployment and availability to the consumer.

OnLive Game Service Portal

Consumer and Social Features

As with any current consumer offering, ease of use and simplicity of navigation is paramount. Social features must also be available to support competitive game play.

In addition to providing access to games, OnLive includes a multitude of social features such as social groups (friends) text and voice chat. One of the unique offerings in the OnLive service is the feature of brag clips and spectating. As all the games sessions are running in OnLive datacenters, the games sessions are available and can be seen by other OnLive users in real time. Using multicast distribution any of these streams can be distributed to any OnLive user so your friends can watch, comment and cheer your game play.

Further this system has a continuous recording buffer that allows the immediate capture of the preceding game play as a "brag clip" that can captured stores and downloaded for use in other social media and social networks.

As this system is completely cloud based, all user information is safely stored, include state of game play, so the user can pause, log out and resume from the existing place in the game and maintain all their statistics, rankings and data from their previous session. This information can be accessed on the OnLive system anywhere. This allows a user to log off and resume their play across town or across the country. This allows game playing and spectating by individuals throughout the system.

Product (Games) Availability

The OnLive system and its servers utilize standard PC games with minimal modifications. A SDK is freely available for any game publisher or author to provision their games for the OL platform. As virtually all games are developed on PC development systems virtually all the currently available and developed games can be easily provisioned and provided for the platform.

OnLive has secured long term arrangements with virtually every game publisher that allows their products to be provided to OnLive at launch in June of 2010. These agreements provide for the same day and date release of titles that will be available at retail stores.

Because of the nature of the OnLive systems publishers can now produce and distribute not only primary releases, but provide episodes, mini release and features for their games after the initial release. The OnLive system supports purchase, rental, subscription and demo distribution of product. Try before you buy now becomes a reality for the consumer.

OnLive Status

OnLive has been in development for several years. It was publicly announced in March 2009 at the Games Developer Conference, and beta began shortly after that. The Beta trials have been extensive and provided valuable insight and detailed information on the OnLive Technology, service and network performance. All the Beta participants have been helpful and the ISP and carriers have been extremely supportive in the efforts to bring this service to market. This information has enabled OnLive to improve its underlying technology and systems.

It also enabled the improved of the real time adaptive aspects of the system and provide valuable feedback to the carriers, games providers and ISPs.

In March of this year OnLive announced it will launch it service in June 2010.

Other Applications and Enterprise Solutions

As the OnLive platform has been initially targeted for the consumer gaming experience, other applications such as media distribution and playback are easily incorporated and blend well with the cost and performance targets of the consumer platforms.

Further this technology (low latency, high quality, cloud based) that enable high end gaming applications to be provisioned over broadband networks to inexpensive widely available PC and Mac platforms also enables cloud based "SAS" type implementation of high end, expensive, piracy prone design, graphics, engineering and architectural applications. These applications being cloud based can leverage off the OnLive infrastructure, and reduce the end users cost of high end PCs and support. Work continues by OnLive in this area.

Summary

OnLive has developed and implemented the technology that enables cloud based real time interactive gaming. This service can be widely available via exiting broadband connections, and has been optimized for minimal latency, high quality over these networks.

Further the system is easily integrated into the consumer environment and provides unique features and availability of games, products and social interaction. The business issues, games availability and features all mesh well and represent an marked improvement for both the consumer and the games providers.

OnLive is among the first of such services that will become widely available taking advantage of broadband and will help drive the uptake and demand of the consumer for improved bandwidth and quality of broadband. The system allows has unique advantages for business and enterprise applications.

CONVERGED MULTISERVICE ACCESS PLATFORM

Jorge D. Salinger
VP, Access Architecture
Comcast Cable

Abstract

A new headend equipment architecture option for implementation of the traditional CMTS and Edge QAM functions is presented. The equipment resulting from implementing this new architecture, called Converged Multiservice Access Platform (CMAP), incorporates all the CMTS and Edge QAM functions. Each CMAP downstream RF port implements all QAMs for digital narrowcast and broadcast services for a single service group. Similarly, CMAP upstream RF line cards implement multiple demodulators per port. This architecture, which can be implemented in a single, integrated chassis, or by separating the packet processing from the PHY and MAC in separate modules, enables unprecedented density in MSOs' headend facilities.

Starting by outlining the key goals and objectives of the architecture, this paper describes the various CMAP components, their key features, the density achieved by the architecture, multiple operational simplicity and efficiency improvements, and the transport agnosticism achieved.

A description of the specifications spelling out the full details of the CMAP product requirements and the timeline for their development is provided. The relationship between the CMAP product specifications and the various CableLabs® specifications, such as DOCSIS®, M-CMTS™, DRFI, MHA, PacketCable™, etc., is also explained.

Examples are presented to show how CMAP could be deployed in typical cable systems, including its deployment in MSO networks of varying sizes, capacities and composition of services. Space and power savings, the key benefits of CMAP, are depicted in comparative analysis.

NOTE: All examples presented in this paper are only for illustrative purposes, and do not reflect the actual deployment plans of Comcast or any other cable operator.

BACKGROUND AND RATIONALE

For a few years now, MSOs have been increasing the number of QAM channels used for narrowcast services. Most MSOs are deploying more and more QAM channels to support growth from the success of Video on Demand, especially as a result of the availability of more High-Definition TV (HDTV) content. Additionally, the use of Switched Digital Video (SDV) for an increasing number of multicast content offerings is driving the deployment of QAM channels even further. And, with the availability of channel bonding in DOCSIS 3.0, MSOs are deploying additional QAM channels for their CMTS equipment to support the newer, higher bandwidth data services.

At the same time, MSOs continue to reduce the size of service groups to make more efficient use of their networks. The drivers, for many years now, have been operational streamlining (smaller service groups result in improved service quality) and efficient use of spectrum (support narrowcast service growth by reusing spectrum).

These two trends result in a continuous increase in the number of QAM channels per service group. Moreover, the expectation from

the current service projections is that such growth will continue and even expand, especially as MSOs reduce the number of analog channels available in the network.

However, the deployment of additional QAM channels in Edge QAM or CMTS equipment cannot easily be supported in the space available within existing typical headend and/or hub/OTN sites.

As a result, the cable industry needs ever denser QAM-channel-per-RF-port Edge QAMs to reduce both the resulting environmental requirements and the capital and operational costs of the equipment itself.

TECHNOLOGY EVOLUTION

Interestingly enough, cable industry suppliers identified the above trends quite some time ago. Because of such foresight, Edge QAM vendors have been developing denser QAM-channel-per-RF-port implementations for several years now, even approaching densities that allow implementations of unique QAM channels for every channel in every RF port.

However, this technology evolution has been difficult to incorporate in equipment available for purchase. It is not a simple operational and financial matter for MSOs to take the leap towards such higher densities for any one service, and consequently it is difficult for vendors to justify the investment to implement this technology. This is not only true for Edge QAM vendors, but particularly difficult for CMTS suppliers implementing integrated architectures.

With Modular CMTS and the Modular Headend Architecture, as defined by CableLabs, it should be possible to achieve such densities as Edge QAM development evolves towards higher densities and CMTS equipment is developed for these network architectures. However, most CMTS development has focused on an integrated architecture for a variety of technical reasons.

ENTER CMAP

A new equipment architecture option is emerging that enables the implementation of denser network architectures in yet another way, providing both MSOs and vendors an alternative approach for achieving the original goals of the Modular Headend Architecture – denser QAM-per-RF port implementations.

Such equipment architecture is described in work underway at Comcast, which is developing product specifications for a new class of equipment called Converged Multiservice Access Platform, or CMAP.

CMAP leverages existing technologies such as DOCSIS 3.0 and current HFC architectures, incorporates newer ones such as dense Edge QAM architectures and Ethernet optics (EPON, in particular). It also leverages the experience acquired over the many decades of technology evolution for cable networks.

The key goals of CMAP include:

- Enabling implementation of denser headend equipment targeting a much higher number of narrowcast services, reducing costs and environmental requirements in headend/hubs/OTNs.

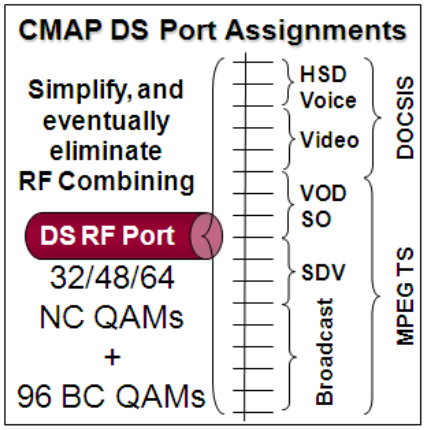
- Developing an access technology-agnostic architecture, making it possible to deploy newer access technologies with the same services architecture.

- Leveraging new and/or broadly deployed technologies to unleash further capacity in the cable industry's HFC network, using overlay architectures to simplify deployment.

CMAP OBJECTIVES

The Converged Multiservice Access Platform is intended to provide a new equipment architecture approach for manufacturers to achieve the Edge QAM and CMTS densities that MSOs require to address the costs and environmental constraints resulting from the success of narrowcast services. In addition to the architecture described in the Modular Headend Architecture Technical Report from CableLabs (i.e., Modular CMTS with Universal Edge QAM), the CMAP provides an alternate approach to the implementation of headend equipment that delivers QAM channels for different services.

To achieve the functionality described above, a CMAP device implements the various Edge QAM and CMTS functions in a consolidated platform. The result, as shown in the figure below, is that a single CMAP downstream port will include all the QAM channels for all



digital services. For example, a typical downstream RF port may be licensed to include 32 QAM channels for narrowcast and 96 QAM channels for broadcast services. If deployed in a 750 MHz system that maintains 30 analog channels, the CMAP RF port will provide 32 QAM channels for narrowcast video, data and voice services and approximately 50 additional QAM channels for broadcast services.

As with the existing CMTS architectures, a CMAP device can be implemented in an integrated or modular manner.

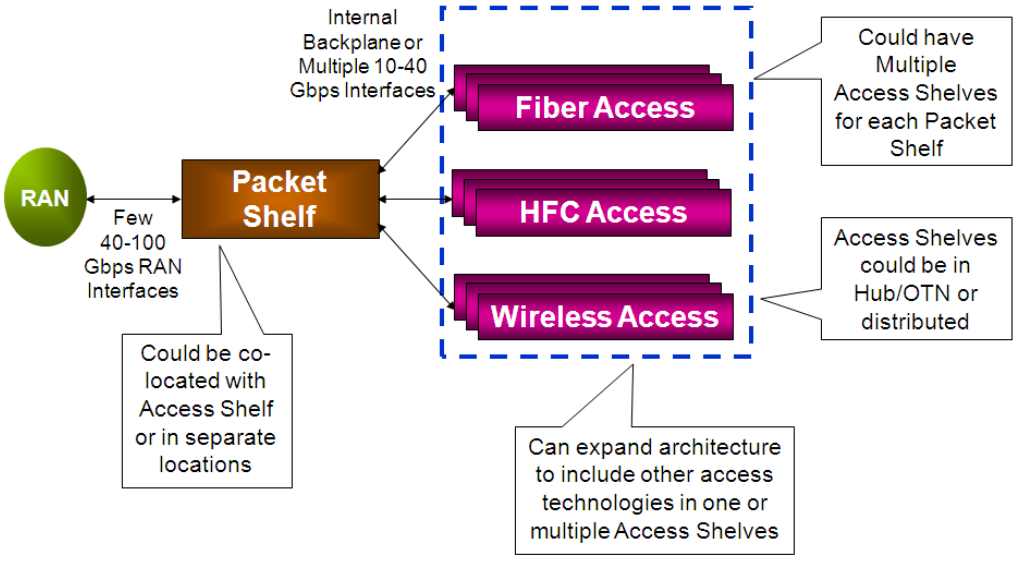
In the first case, all functions are implemented in a single chassis.

In the second, CMAP functions are divided between a Packet Shelf (PS) and an Access Shelf (AS), as follows:

- The PS implements the packet processing functions, such as subscriber management, service flow management, video program stream edge manipulation (e.g., multi-program transport stream creation, PCR restamping, etc.), layer-3 routing and higher layer protocol manipulation, and other such functions.

- The AS implements all the upstream and downstream PHY functions normally associated with the CMTS and the Edge QAM, and as much of the MAC as needed to support both upstream and downstream flows. A documented interface between the AS and the PS is defined to enable interoperability between AS and PS vendors.

The figure in the next page outlines one possible modular implementation of CMAP, where multiple types of Access Shelves are available for different access architectures. However, other implementations are also possible. For example, multiple access technologies could be incorporated into a



single AS, so that one centralized PS would interface with multiple AS devices possibly distributed across various sites.

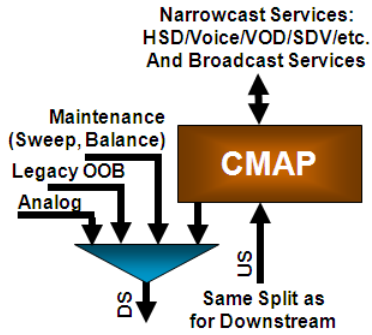
The key functional goals for CMAP include:

- Flexible use of QAM channels for the various services offered by MSOs, enabling the configuration of the CMAP to provide a changing number of MPEG transport stream-based services (e.g., for VOD, SDV, etc.) versus DOCSIS-based services (e.g., HSD, voice, etc.).

- Individually configurable assignment of QAM channels to the various service groups, so that it would be possible to have service groups for HSD/voice, VOD, and/or SDV overlap in different ways without requiring the various service groups to provide homogeneous coverage.

- Efficient implementation of Edge QAM blocks by implementing separate sets of QAM channels for narrowcast and broadcast applications, such that QAM channels for narrowcast services are individually implemented for each RF port but QAM channels used for broadcast services are shared amongst all the RF ports in each downstream line card (DLC).

- Simplification of the RF combiner by providing all QAM channels for all digital services from a single RF port, leaving only certain legacy functions for the RF combiner network. The list



includes any remaining analog channels, the legacy out-of-band control channel, and any maintenance equipment not incorporated into the CMAP, as shown in the figure above.

- Implementation of sophisticated proprietary encryption systems (e.g., PowerKEY, DigiCipher, etc.) without requiring special-purpose hardware, so that a CMAP from any vendor can implement either encryption mechanism,

or both mechanisms, with the same platform.

- A transport-agnostic network architecture, including implementation of PON and other access network technologies natively within the CMAP.

- Significant operational improvements, including significant environmental efficiencies (e.g., much less space, power consumption, and heat dissipation), implementation of functions such as upstream spectrum surveillance, continuous wave carriers for plant amplifier biasing, and many other operational enhancements.

SCOPE OF CMAP SPECIFICATIONS

The requirements included in the CMAP specifications currently under development at Comcast outline product requirements, including capacity, performance, network implementation functions, and other such targets and objectives. In doing so, the CMAP specifications reference industry standards, such as CableLabs specifications (e.g., DRFI, DOCSIS, VSI, PMI, PacketCable, etc.) and SCTE standards (SCTE-02, etc.), without duplicating the requirements detailed in those standards.

Please note that the CMAP specifications do not contradict or redefine any industry standards. Where necessary, changes are made in the industry standards, and not in the CMAP specifications. For example, certain changes are being made to the CableLabs DRFI specification and the SCTE-02 standard, which the CMAP specifications take into account or anticipate with appropriate descriptive language.

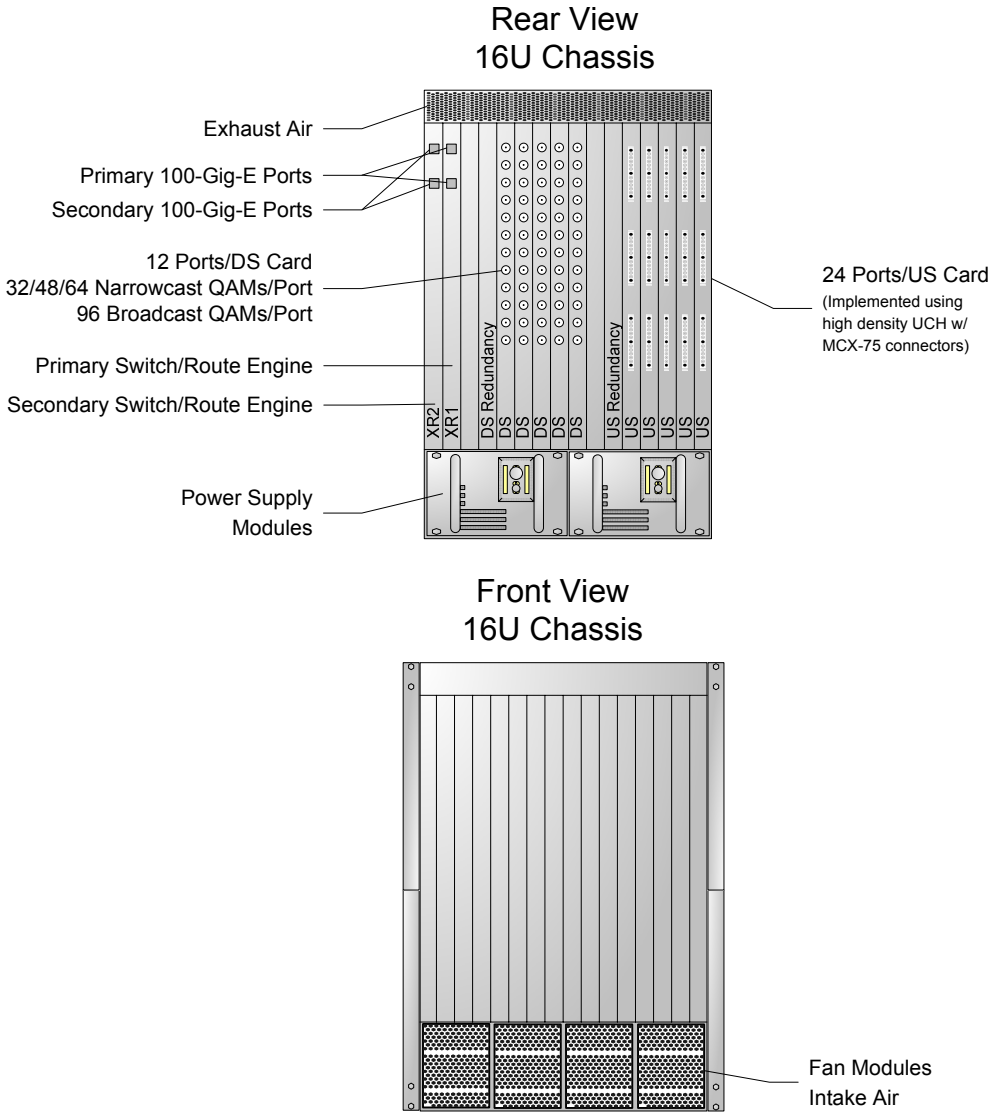
In some cases, the CMAP specifications detail requirements that exceed those of other industry standards. For example, an industry standard may indicate a preference with a SHOULD requirement while the CMAP specification might label it an absolute requirement with a MUST.

To develop the CMAP product specifications, Comcast is working with a broad group of industry leaders and technology experts from companies interested in the development of a CMAP, all of whom have volunteered to help Comcast develop these requirements. Staff members from CableLabs, Cable Europe Labs, and other advisers are assisting Comcast in this effort. Most importantly, a number of contributors from various North American and European MSOs are participating in the effort as well.

As detailed in the table above, the CMAP Team plans to complete three product requirements specifications in the next few months. The first of these product specifications, called the CMAP Hardware and Functions Specification, has been completed. The other two, Configuration and Management and PASI, are currently under development and should be completed by the middle of 2010. Additionally, following the completion of the product requirements specifications, the team plans to develop recommended test procedure specifications.

Specification	Objective
Hardware and Functions	Hardware components and requirements, and the various features and functions implemented by the CMAP.
Configuration and Management	Interfaces and requirements for configuring and managing the CMAP
Access Shelf-to-Packet Shelf Interface (PASI)	Functions performed by the PS vs. the AS and the characteristics of the interface between the two components.

In addition to the CMAP product requirements specifications, the team has contributed to additions and/or changes to existing or new industry specifications. The main body of work in that regard has been related to the CableLabs DRFI Specification, which has undergone several Engineering Change Requests (ECRs) to accommodate the



design and operation of implementations of large numbers of QAM channels-per-RF-port, which are applicable to the CMAP as well as other dense QAM channel-per-RF-port equipment implementations. Another area of industry specification work relates to SCTE-02 through the SCTE IPS Working Group, regarding enhancements to the "F" connector requirements and the addition of a 75 Ohm version of the MCX connector. The 75-Ohm MCX connector is commonly implemented in a gang holder known as UCH, which consists of a row of 10 connectors typically used with mini coaxial cable. Other industry specifications may be updated as deemed appropriate.

Unlike other similar efforts to date, such as RFIs, RFQs and/or prior requirements documents for CMTS and Edge QAM equipment, the CMAP specifications outline very specific chassis requirements. These requirements include comprehensive line card implementation details, such as the number of supported QAM channels for each function in terms of density and redundancy characteristics. They also spell out such physical interface objectives as type of connectors, detail preferences for power supply locations and orientations, set airflow direction and entry/exit requirements, and detail many other such implementation requirements. These requirements should limit

vendor innovation because vendors will still be left with many opportunities for differentiation. At the same time, they recognize the need to simplify operations by creating standards for key operationally beneficial parameters.

The following figure shows a possible front and rear view of the CMAP chassis that would be compliant with the CMAP specifications. In the figure, the following details are depicted: rear-facing connectivity for all components; downstream line cards (DLCs); upstream line cards (ULCs) with twice as many upstream ports as downstream ports for 2:1 upstream-to-downstream ratio; redundant switch-route engines with primary and secondary 100 GigE ports; redundant power supplies; and vents for air flow. The diagram does not depict PON line cards for business services, which are not mandatory but are strongly preferred.

Given the scope of each RF port providing all services for a given service group, it is important that the operation be highly reliable. Therefore, the chassis is required to implement N+1 redundancy for upstream and downstream line cards and 1+1 redundancy of all common equipment. This line card redundancy is achieved by way of a mid-plane near-passive RF switch and the use of physical interface cards (PICs), which provide the separation between the active components with critical mean time between failure (MTBF) and the RF interfaces to the minimum remaining external combining and downstream/upstream lasers.

CAPACITY ESTIMATES

To help guide equipment and network design, the table included below depicts three deployment scenarios for a CMAP, as follows:

- Minimum, albeit unlikely, deployment in the left column,
- Maximum, also unlikely, deployment in the right column, and
- Estimated probable initial deployment in the middle column.

These scenarios show that a downstream NSI capacity of 15 Gbps is easily necessary, while an absolute maximum of 155 Gbps could possibly be required for the line cards envisioned in the specification. But a capacity of approximately 30 Gbps is most likely.

Similarly, for the upstream NSI direction, a capacity of about 7 Gbps might be required upon initial deployment, as depicted in the table.

Please note that the scenario considered probable in this example depicts 5 downstream line cards, with 32 active narrowcast QAM channels, for which 50% of the content is unique (e.g., 50 % of the content is replicated via multicast). It also includes 1 broadcast line-up for the entire chassis with 75 active QAM channels.

Approx Calculations	Minimum Chassis Capacity (Mbps)	Probable Capacity (Mbps) 5 DLC, 32 NC/RF (50% unique), and 1 BC/Chassis (75 QAMs)	Maximum Chassis Capacity (Mbps)
DS NC QAMs per RF Port	a1 = 22 QAMs * 36 Mbps	b1 = 32 QAMs * 36 Mbps	c1 = 64 QAMs * 36 Mbps
DS BC Lineups per LC	a2 = 40 QAMs * 36 Mbps	b2 = 75 QAMs * 36 Mbps	c2 = 96 QAMs * 36 Mbps
DS NC+BC Chassis	((a1 * 8 ports) * 50% unique) * 4 LCs + a2	((b1 * 8 ports) * 50% unique) * 5 LCs + (b2 * 2 LCs)	(c1 * 12 ports * 100%) * 5 LCs) + (c2 * 5 LCs)
US Port	a3 = (2 US * 26 Mbps) + (1 US * 8.8 Mbps)	b3 = (3 US * 26 Mbps) + (1 US * 8.8 Mbps)	c3 = (4 US * 26 Mbps) + (2 US * 8.8 Mbps)
US Chassis	a3 * 8 ports * 4 LCs	b3 * 16 ports * 5 LCs	c3 * 24 ports * 5 LCs
DS NC QAMs per RF Port	792	1,152	2,304
DS BC Lineups per LC	1,440	2,700	3,456
DS NC+BC Chassis	14,112	**28,440**	155,520
US Port	61	87	122
US Chassis	1946	**6,944**	14,592

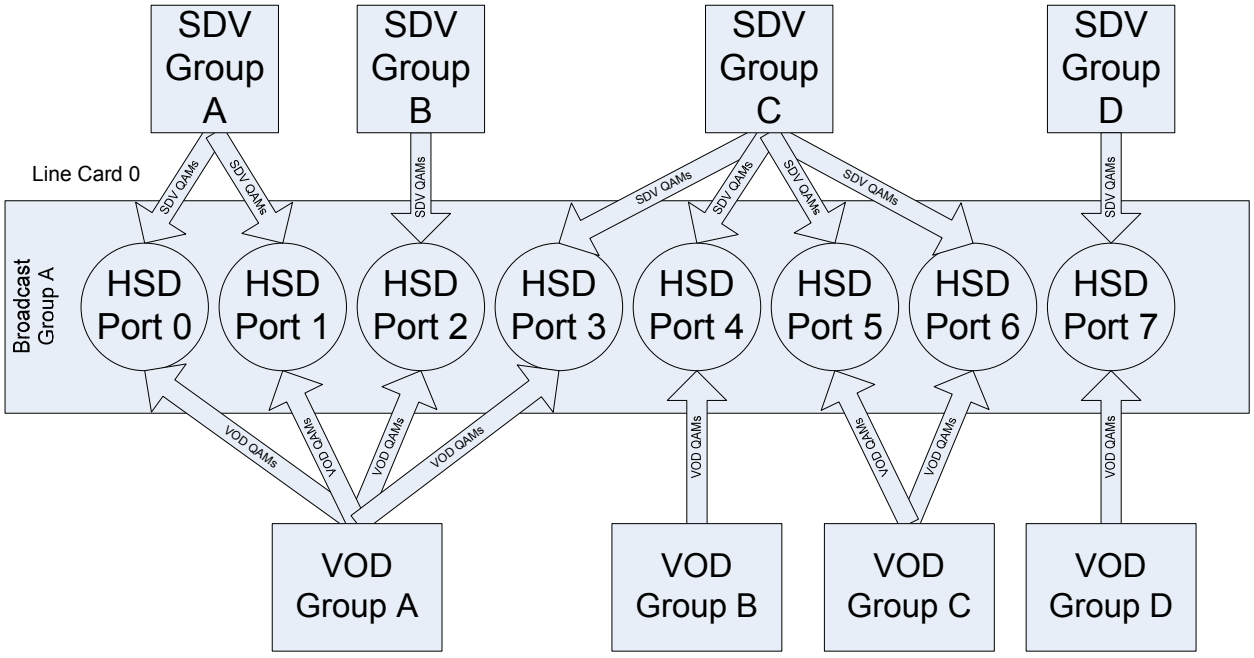
For clarity, the top portion of the table shows the calculations and the bottom portion of the table shows the calculated results in Mbps.

QAM REPLICATION

In order to simplify integration of the CMAP into existing systems, the CMAP requires a QAM channel replication feature. With this feature, the CMAP can "copy" the contents of one QAM channel onto the same QAM channel in one or more other RF ports, thereby implementing electronically an analogous function as RF splitting an Edge QAM port and combining the splitter outputs into multiple service groups.

The figure below illustrates the QAM Replication feature. Note that each RF port has a unique HSD service group depicted with a circle.

The purpose of QAM Replication is to allow



the creation of service groups on a decoupled (service-by-service) basis. With this feature, an MSO can replicate SDV and/or VOD QAM channels across multiple ports on a given line card.

The genesis for this feature can be found in the current deployment scenarios, where VOD service groups implemented by a separate set of Edge QAMs may span multiple HSD service groups. In other cases, a SDV service group may span a different number of HSD service groups, and perhaps more than one VOD service group.

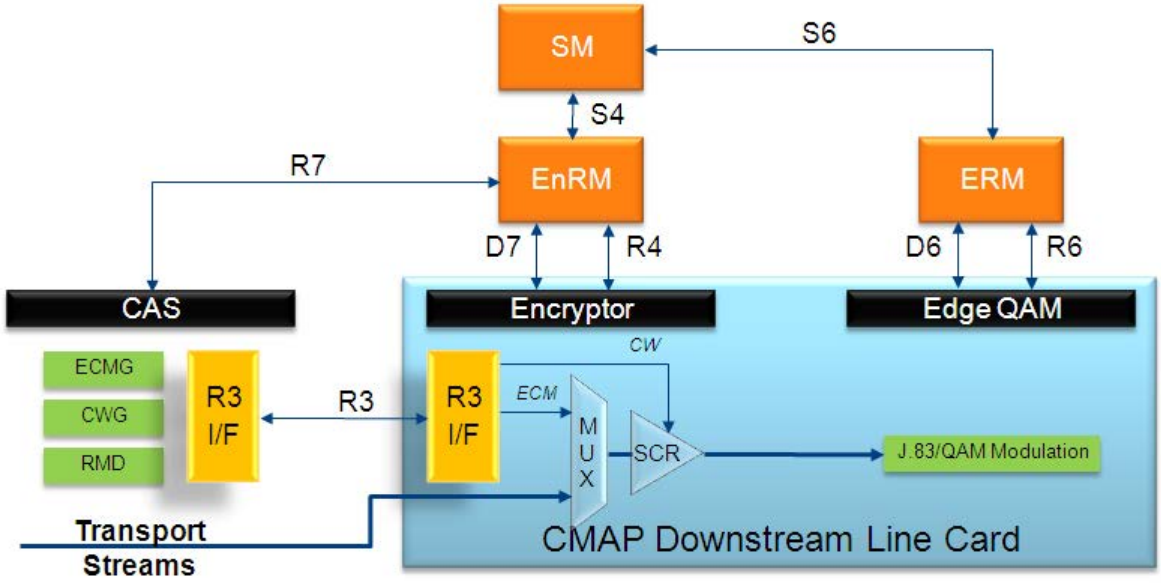
Because HSD, VOD and SDV service groups are currently implemented by separate Edge QAM devices, today's combining is done at the RF level on an as-needed basis. But, as MSOs move to the CMAP, where each RF port has its own QAMs, it is not possible to combine service groups at the RF level any longer.

One way to deal with the new scenario is to align service groups, where a VOD service group would geographically coincide with an SDV service group and with an HSD service group. Another approach is to use this QAM Replication feature outlined herewith.
Neither approach offers a perfect solution. Service group alignment requires additional work, as well as extra equipment in many cases, and results in a loss of economies of scale such as those that benefit multicast for SDV. QAM channel replication, on the other hand, saves on the work and equipment required to align service groups and maintains the current deployed scenario, including its economies of scale. On the flip side, however, it does not save on QAM channel cost because the hardware for each individual QAM channel has to be in place for each port anyway.

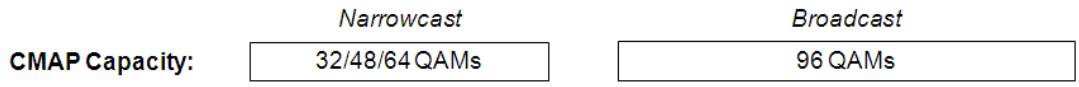
ENCRYPTION CAPABILITY

Given the broad video services supported by the CMAP, it is imperative that the CMAP implement extensive encryption capabilities. Therefore, the CMAP Team invested a significant amount of time in developing a strategy for encryption within the CMAP so that a CMAP from any manufacturer, given the appropriate licensing, will support encryption to the fullest extent from any Conditional Access System (CAS).

To accomplish this, the CMAP implements a very clever scheme, previously envisioned for a cousin technology called Next Generation



on Demand, or NGOD. As depicted in the figure below, the CMAP Downstream Line Card implements an Encryptor. This functional entity within the CMAP implements a superset of standard algorithms for scrambling content that work for most, if not all, CAS vendor systems. Additionally, the CMAP implements a set of interfaces, which are specific and defined by the CAS vendor for interfacing to an Encryption Control Message Generator (ECMG) and Control Word Generator (CWG). In turn, the CAS vendor would implement the ECMG and CWG according to the intricacies of their own CAS.

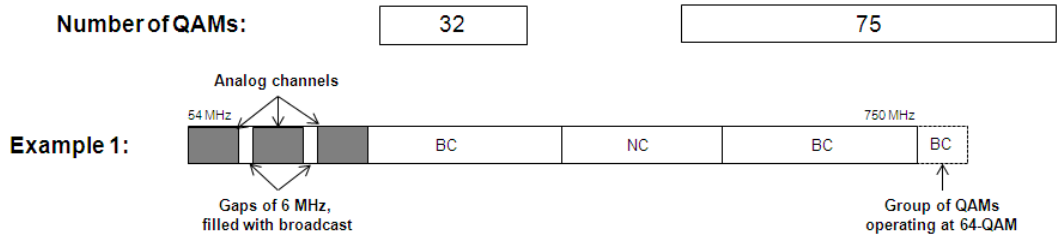


With this approach, the CMAP from any vendor that has entered the appropriate agreements with the CAS developer can implement full encryption system capabilities, including session-based scrambling. Not only is this not possible today, but today's technology allows only very minimal encryption functions by third party vendors, and requires very complex agreements to implement. The interfaces proposed in this approach are far more straightforward, revealing close to nothing regarding the CAS methods and procedures. Consequently, the agreements should be simpler to establish.

DEPLOYMENT EXAMPLES

Included in this section are examples of the possible deployment of a typical CMAP configuration in two types of systems, one implementing an HFC network with 750 MHz of spectrum capacity and another with 860 MHz of capacity. Both use cases are for typical systems, including a normal number of homes passed per node and per hub.

The two examples are for the deployment of a CMAP chassis consisting of the same configuration, as detailed in the following diagram.



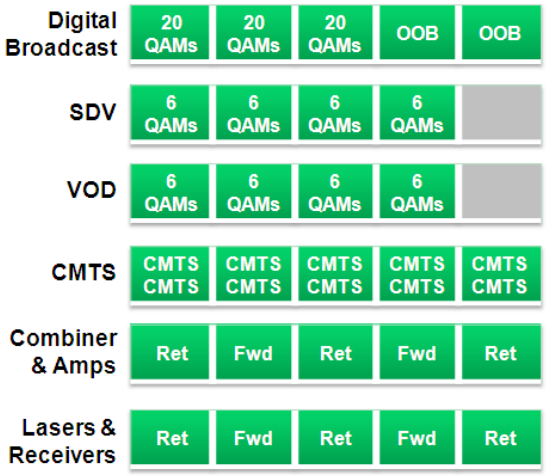
The CMAP chassis is capable of supporting a capacity of up to 64 QAM channels for narrowcast services and up to 96 QAM channels for broadcast services.
In the first example, detailed in the above figure, there is a group of analog channels (approximately 30) in the lower portion of the spectrum with a small number of gaps (2 as depicted in the example) consisting of a few 6 MHz channel slots. These gaps between analog channels are occupied by digital programs from the group of broadcast QAMs.

Additionally, there is a group of narrowcast QAM channels located towards the center of the spectrum. The remainder of the spectrum is occupied by broadcast QAMs -- a few of which are configured to operate in the roll-off portion of the spectrum and set to 64-QAM

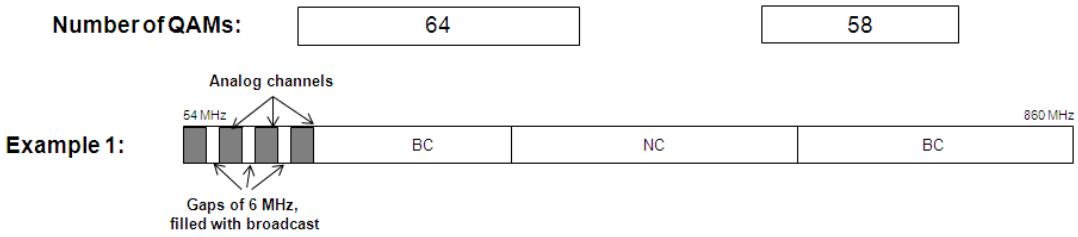
modulation, as opposed to all other QAM channels which will operate at 256-QAM modulation. Consequently, while all 32 narrowcast QAM channels would be used, approximately 75 of the broadcast QAM channels would be used in this example.

The second example, detailed in the figure at the bottom of this page, depicts the use of the same chassis.

In this example, the cable system is capable of supporting 860 MHz of spectrum. Similar assumptions for analog channels are made for this example, but additional narrowcast QAM channels are used instead and fewer broadcast QAM channels are needed to fill the available spectrum.



Considering typical CMTS and Edge QAM equipment as available today, the figure above depicts about 10 CMTS chassis, and about 4 racks for VOD and SDV, each containing 6 Edge QAM chassis configured for 64 QAM

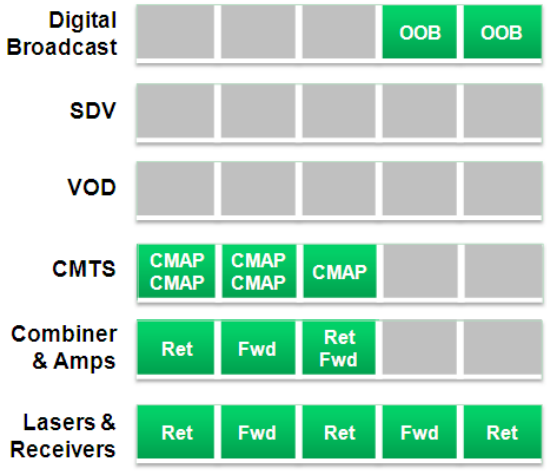


channels each at a density of 4 QAM channels per RF port. The digital broadcast lineup is composed of 60 individual QAM channels, plus the corresponding out-of-band equipment.

ENVIRONMENTAL EFFICIENCY

One of the key objectives of CMAP is to achieve significant environmental efficiencies. To that end, the following is an example of the space and power savings achieved by deployment of the CMAP in a typical system. The following figure depicts a typical installation in a headend consisting of the various digital services, including broadcast, SDV, VOD and HSD equipment, plus the corresponding combiner and lasers/receivers.

The example shown above is intended to serve a typical population of 200 nodes, combined in such a way as to result in 160 HSD service groups, and 120 VOD and matching SDV service groups.



The following figure depicts the analogous installation when considering the deployment of equivalent CMAP equipment.

The above figure shows the following:

- Given that the CMAP chassis would have twice the density of a typical CMTS, only ½ the number of CMAP chassis are required, resulting in equivalent space savings.

- However, the CMAP chassis in its basic implementation includes all the necessary QAM channels for supporting the VOD and SDV services. Therefore, no additional equipment is needed to support these functions, resulting in significant additional space savings.

- Given that the CMAP also supports sufficient broadcast QAM channels, the space previously allocated to the broadcast QAMs channels is no longer needed, further contributing to space savings.

- Finally, it is estimated that ½ of the space allocated to the combiner network would be saved, resulting in even further space savings.

With all this taken into account, it is easy to see how as much as ½ of the space previously required is needed for deploying the CMAP. But, moreover, given that the CMAP can serve twice as many narrowcast QAM channels as the previous architecture could, the depicted CMAP scenario actually results in even greater space savings, providing twice as much capacity in ½ the space.

Clearly, the space savings are staggering!

In addition, it is worth considering the power savings.

A cursory analysis of the difference in power consumption, assuming typical power draw for existing equipment and the expected power consumption for the CMAP, yields an estimated power savings >50%. And, this is taking into account the use of 32 QAM channels in the CMAP, or 2x the capacity indicated in the original typical deployment. And, this does not even include the cooling savings from the great reduction of equipment and power consumption.

Without a doubt, the power savings are also very significant!

SILICON DEVELOPMENT

One important consideration is the evolution and availability of silicon components.

The functionality described by the CMAP specifications does not require of any new silicon. This is the case for both the upstream and downstream. CMTS vendors are already using and/or planning on making available line cards with existing and available high density burst demodulator silicon and corresponding MAC chips for upstream. For the downstream, vendors can utilize existing technology for Direct Digital Synthesis (DDS) consisting of readily available FPGAs and Digital to Analog Converters (DACs) from multiple vendors.
However, multiple silicon suppliers are in the process of implementing chips that provide very large QAM channel counts for downstream implementations. Some of these implementations are able to support QAM modulators for the entire RF spectrum from a single chip!

Even though these new silicon implementations are not required to develop a CMAP, they will certainly simplify designs, help reduce printed circuit board space and power/heat dissipation requirements, help reduce costs further, and accelerate development once the new silicon is available.

ANTICIPATED TIMELINE

As with any other technology evolution such as this one, things take longer than desired. On the flip side, their acceptance usually has farther reach than expected.

Clearly, from the many discussions with other MSOs, both within North America and throughout Europe and South America, interest for the deployment of this platform is very high. Almost without exception, MSOs at large are interested on the operational simplifications that the CMAP offers and the new functions it enables.

From preliminary discussions with vendors, and without revealing confidential information and plans, initial availability of equipment for laboratory and field testing is planned for late in 2011, early deployments are planned for 2012, and broad availability from multiple vendors by 2013.

CONCLUSIONS

The CMAP Platform is a viable alternative to the existing Modular Headend Architecture, and in fact may represent the Next Generation CMTS and Edge QAM.

CMAP implements all the QAM channels for each RF Port, supporting all digital services, including VOD, SDV, broadcast, HSD, and others in the future.

Given the environmental savings alone, field operations could benefit from CMAP immediately. Without CMAP, considering the expected growth in narrowcast services in the years to come, MSOs would likely have to resort to expensive headend/hub expansions.

Moreover, CMAP ports will be much more cost effective than current CMTS and Edge QAM ports, so the cost of expansion will be greatly reduced.
Finally, while CMAP can be implemented with existing technologies, it can greatly benefit from the natural technology evolution in chip development.

The challenge for MSOs is to know how and when to begin deploying CMAP. For some MSOs, the answer is as soon as the equipment can be manufactured.

ACKNOWLEDGEMENTS

Thanks to my colleagues John Bevilacqua, Peter Hutnick, Doug Jones, Saif Rahman, Arun Rajagopalan, Esteban Sandino, Joe Solomon, and Dave Urban from Comcast for their contributions to the CMAP program and specifications.

Also, many thanks to my colleagues from CableLabs®, Cable Europe Labs, the many vendors and advisors of the Comcast CMAP VE Team, and especially fellow MSOs for their many contributions to the development of the CMAP specifications.

DELIVERING PIXEL PERFECT

Dr. Robert L Howald, Dr. Sebnem Zorlu-Ozer, Dr. Nagesh Nandiraju
Motorola Home & Networks

Abstract

Operators are continuing to enhance their service mix with more personalized content, solutions that deliver the content to multiple screens, and doing so at accelerated deployment speeds. These objectives, among others, have driven plans to evolve the cable infrastructure towards an end-to-end IP architecture. Cable's IP pipe on the access network is, of course, the DOCSIS platform. However, the origins of DOCSIS were not developed with video services in mind. That has changed with DOCSIS 3.0. Nonetheless, supporting video requires the revisiting of traffic engineering principles used on today's DOCSIS access links.

Video over DOCSIS is expected to use H.264 encoding and variable bit rate (VBR) delivery, compared to legacy CBR MPEG-2 TS-based delivery and MPEG-2 encoding. In addition, novel adaptive streaming technologies offer intelligent alternatives to streaming models. Using a proven CMTS simulation tool, performance of video over standard DOCSIS links has been evaluated [2]. We extend these results for high and low action content, including the effects of peak capping, buffering, and CMTS configuration parameters on network performance. Quantifiable insight into the relationship between transmission losses and video performance will be examined. Finally, we will introduce adaptive bit rate technology into the model. These results will help operators understand the variables involved to traffic engineer their DOCSIS network for video services.

INTRODUCTION

The video service mix has gradually grown over the years in terms of technology, complexity, and consumer offerings – VOD, PPV, SDV, MPEG-4, OCAP, HDTV. The momentum of this march to video services paradise was jolted when a key crossroads occurred, as shown in Figure 1.

Suddenly, "HSD" went from meaning "High-Speed Data" to standing for "Heckuva Streaming Demand." Of course, this stage of data speed evolution represents an essential "must have" for the cable IP pipe to be considered as a means for delivery of video content. Figure 1 puts the inevitable into pictures, identifying that crossroads in time when high quality video rates became low enough that the increasingly fat data pipe could effectively deliver it to residential subscribers. An important point to make on the topic of cable IP video is that, in the context of this paper, we are referring usually to MSO-owned video assets, as opposed to over-the-top providers.

Why the fuss over IP delivery given the cost-effective infrastructure in place? There is no single answer, but instead a list that, when taken as a whole, makes a compelling case for migrating from purpose-built video system architectures to an all-IP architecture. Operators have routinely described these perspectives in many conference sessions and industry events, where key technologists espouse their views on when, why, and how.

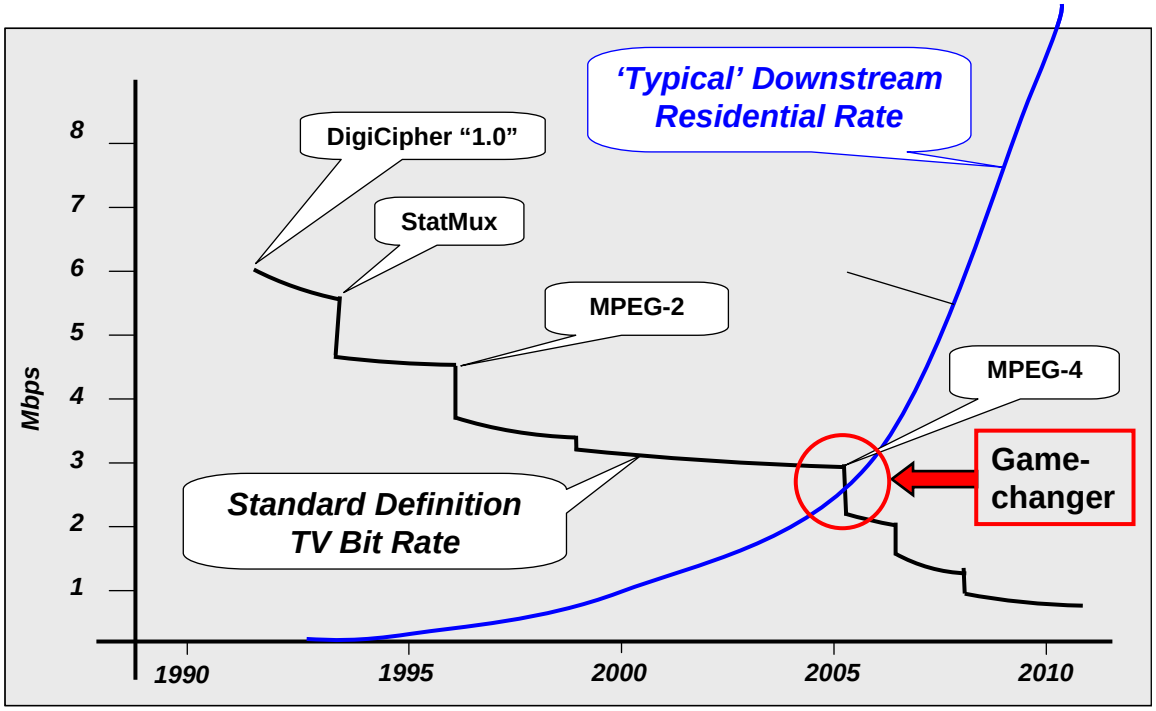


Figure 1 – Downstream Internet Speeds vs Digital Video Requirement

Typically, the reasons involved include readily enabling the multi-screen experience, compatibility with mature IP home networking technologies and initiatives, lower cost CPE, software-based security, enabling future alternative access networks, and closing the last loophole in E2E IP delivery, which is video delivery over HFC. This is expected to lead to improving the velocity of new services delivery and associated OPEX savings in the long run.

There are other obstacles besides the substantial legacy investment to achieving a full migration, one of which is new bandwidth. However, in general, there is quite a bit of underutilized downstream capacity. And, there are many techniques, traditional and not, to go about extracting it that can be deployed as traffic demands continue to increase [1].

While "how to" discussions take place and bandwidth expansion activities continue, a final important "how to" remains: how to system engineer the access edge for IP video. The significance with which video service affects traffic parameters and ultimately bandwidth occurs primarily in two ways:

1) The pure volume of bits-per-second required for video streams
2) The concurrency of use factor for video services vs browsing-based services for HSD

It is simple to show that video concurrency rates of 5-10% (VOD-like parameters) has significant impact on HSD bandwidth requirements, when considering that 1% or less is a typical data oversubscription rate.

PREVIOUS MODELING - SUMMARY [2]

Model Description

OPNET™ CMTS Model

In [2], a key modeling tool was developed for analyzing video over DOCSIS

performance. It is the basis for the results presented there, and is leveraged and extended in this paper to further develop and refine video over DOCSIS performance. The model is based on a DOCSIS model using OPNET™, version 14.5. A simple reference scenario is shown in Figure 2, where a CMTS serves a set of homes with cable modems connected to subscriber equipment.

Modeling Input Stimuli

A large bulk of the modeling research is based on volumes of traces captured and made publicly available, and which can be easily be imported to the model as stimulus. Traces from a video clip library at http://trace.eas.asu.edu and http://trace.kom.aau.dk were used [4]. A brief description of what is encompassed in these online libraries is discussed in [2]. Generally, there are volumes of CIF (352x288) and HD traces across a range of PSNR and quantization settings. As pointed out, lower resolution formats such as CIF and VGA – common for smaller screens – tend towards a higher peak-to-average and thus represent conservative examples from a modeling perspective. We choose from these clips only the high video quality samples (PSNR of 40 dB or greater). The associated quantization parameters have the effects of creating higher rate CIF streams, representing values close to cable SD rates for H.264 (MPEG-4 AVC) encoding.

In addition to the streams above, some clips captured by Motorola were mixed in, as will be seen in the tables that follow which list the streams. Finally, in some cases, H.264 Scalable Video Coding (SVC) clips were used where it helped fill a wideband channel to exercise it at high utilization. Like CIF, SVC also has the property that it tends to aggravate peak-to-average variation, or coefficient of variation (CoV).

Summary of Key Results

This paper builds on the results of [2], so we will briefly summarize some of the key findings from those simulation examples.

A simple "static" gain model was created to point out the potentially large variation of bandwidth efficiency over CBR delivery based on content mix. Table 1 shows the range of efficiency "gain" of VBR – or, more accurately, adjustable CBR – under a very simple, illustrative, assumption of two video classes and MPEG-2 encoding. Assuming a 3.75 Mbps CBR system of 40 programs (four bonded channels of DOCSIS 3.0), and gain made available by allocating 2.50 Mbps to the "easy" programs, there are bits freed up to add more channels. Thus, "easy" programs offer 33% savings to spend elsewhere. For a mix of easy and hard channels that exist, and a desire to add new channels, also of each type, Table 1 shows the effective gains of this scheme, pointing out the dependency on the content type.

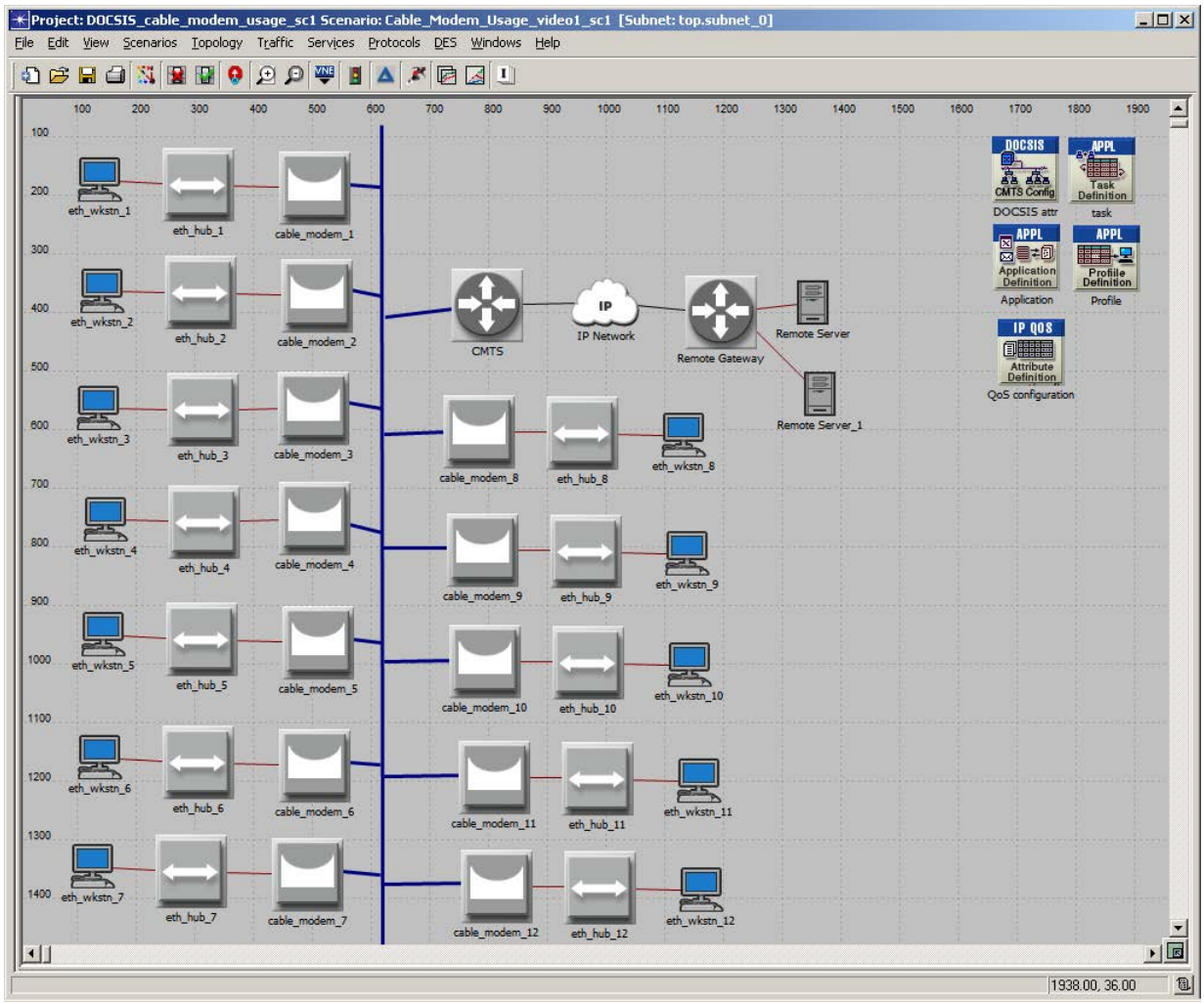


Figure 2 - Sample of a Simulation Scenario Using *OPNET*™

Table 1 – Efficiency Gain – Two Classes Example (Easy / Hard)

Added Programming Mix	Existing Programming Mix				
	70/30	***60/40***	***50/50***	***40/60***	***30/70***
70/30	30.4%	26.1%	21.7%	17.4%	13.0%
60/40	29.2%	25.0%	20.8%	16.7%	12.5%
50/50	28.0%	24.0%	20.0%	16.0%	12.0%
40/60	26.9%	23.1%	19.2%	15.4%	11.5%
30/70	25.9%	22.2%	18.5%	14.8%	11.1%

"Easy" = 2.50 Mbps
"Hard" = 3.75 Mbps

We see that the range of gain varies by nearly three times (11.1% to 30.4%) based on this reasonable range of content mix.

Channel Utilization and VBR Efficiency

Table 2 summarizes the comparison of existing research characterizing H.264 with the simulations described above in terms of percent channel utilization. This analysis, drawn from the same content pool, was described in detail in [2]. There is close agreement between analysis and simulation results, for both single channel and bonded, wideband channel models.

Table 2 – Simulated Utilization vs. Calculated [2]

	Supported Load	Overloaded
Analysis Min	55%	
Analysis Max	71%	
Analysis Avg	62%	
Simulated 1 QAM	63%	72%
Simulated 4 QAM	74%	76%

The simulation results were translated to VBR gains based on DOCSIS scheduling under a particular, typical configuration, a reasonable user buffering limitation, a packet loss threshold, and a factor for the grooming and multiplexing imposed on the streams prior to reaching the edge device due to standard video processing operations. Resulting estimates of VBR gain showed a range of 9-37% increased efficiency, depending on content mix and channel size (single vs four bonded channels).

The above summary of the analysis in serves as a useful baseline to further examples.

NEW SIMULATION RESULTS

DOCSIS 3 Channel Bonded SD + HD Mix

Additional simulations were performed on the mixed-resolution scenarios to quantify further the conclusions about the effects on bandwidth efficiency, and to further exercise system variables under typical configurations. The results of Table 2 indicated that for four bonded channels, 74% capacity utilization was achieved, while 76% utilization caused packet drops at a rate greater than the 1e-6 threshold chosen. In that model, the CPE buffer was fixed at 100 msec, putting a larger burden on the CMTS scheduler to process and deliver the video payloads efficiently without any statistical information to support network admission or congestion management.

This same HD + CIF content line-up was used as a starting point and modeled while making adjustments to network variables. If, for example, we allow the CPE buffer to increase to up to 500 msec – about the maximum that can be considered before other issues come into play – the model shows that additional streams (or higher rate streams) can be added, increasing the utilization efficiency. Measuring utilization efficiency as the overall mean rate of streams to throughput capacity, the channel utilization can be taken up to 78%, or about 6% additional gain over what was derived in [2]. This 78% efficiency can be held as well using a 400 msec CPE buffer, but in this case only if a maximum rate cap is enforced at 10.5 Mbps per stream. This would impact about half of the HD streams, and a few by a percent reduction that would be anticipated to be noticeable, particularly if sustained. With respect to VBR efficiency, the four channel bonded case relative to [2] was improved, resulting in efficiency gains varying from 24%-40%.

Figure 3 shows this truncated received traffic and the injected traffic when the CPE buffer was limited to 200 msec, an increase in buffer size, but still inadequate to accommodate the added traffic without a further increase.

Figure 4 and Figure 5 show the queuing and capping behavior that lead to mitigation of the congestion issue. The queuing delay can be outlasted as shown, and described above, with the appropriate buffer size. The impact of capping at different rates (10.5 Mbps, 11.0 Mbps, and 12.0 Mbps) is shown in Figure 5, where 10.5 Mbps draws the traffic level beneath the aggregate needed to avoid lost packets.

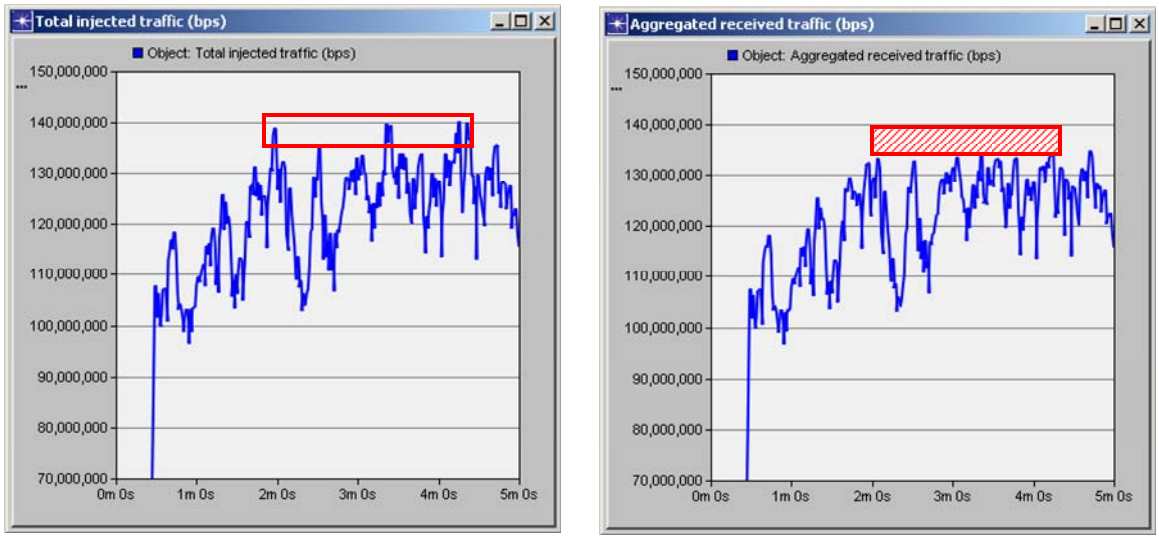


Figure 3 – Injected vs Received: Overloaded & Clipped Video Traffic

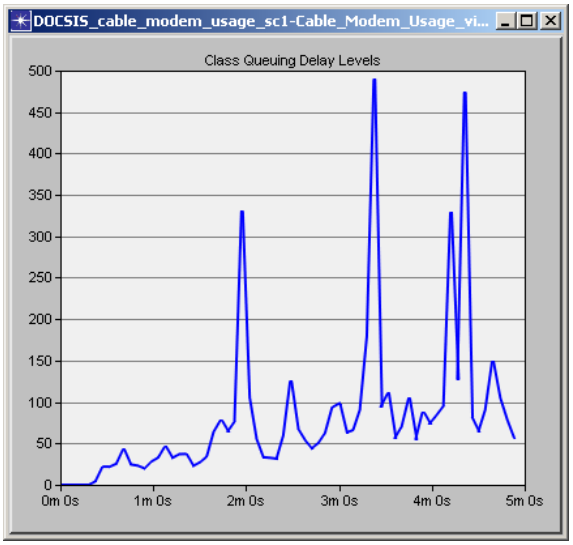


Figure 4 – Delay – Buffer Size Impacts

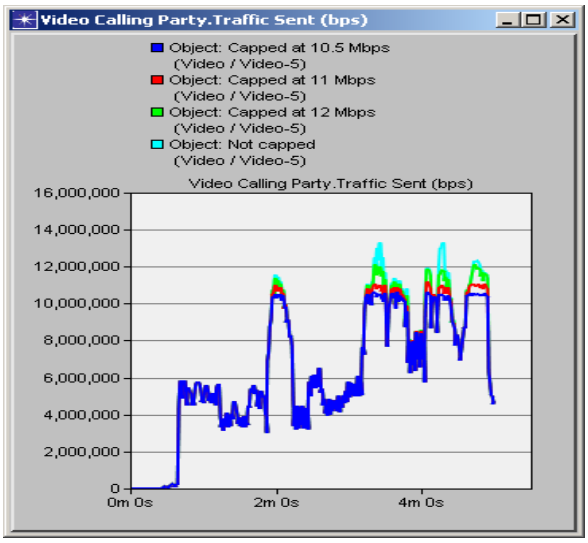


Figure 5 – Peaking-Capping Effects

Channel Efficiency: Low vs. High Action

To gain insight into the video content type dependency to efficiency, cases were run comparing sets of low-action and high-action content, in this case comparing single QAM carriage for HD – the least effective use case from a statistical multiplex perspective. While not a column in the Table 1, note that the efficiency gain of 100% "hard" content would be zero in the context of how Table 1 was derived. There would be no streams upon which an adjustment downward would be considered acceptable. Such is the case with the all high-action HD content simulation (Mean = 5.8 Mbps, pk-avg, sum basis = 2.34). Under this stimulus, drop-free transmissions occur when supporting four HD programs, which is essentially the expected CBR equivalent of HD/QAM for high action content using MPEG-4, and assuming an average 50% encoding gain. The utilization efficiency for this 4-HD program case was about 60%. This is in line with single-QAM efficiencies from prior simulations and shown in Table 2, despite the smaller statistical basis due to low stream count. This is likely due to the relatively well-behaved peak-to-average of the content mix (2.34) noted above.

Mixing in low and high action HD content, at what can be considered simplistically as 50/50 "hard" vs "easy," six streams of HD were fit within the single QAM. This represents a 50% "gain" in video programs compared to the "hard only" case, but roughly the same utilization efficiency. In this case, the slightly larger statistical basis leads to no better utilization efficiency than the high action case above. This is, again, likely because of the peak-avg behavior, which for the six-stream HD multiplex is higher than in the all high-action case.

The additional stream gain does compare favorably to Table 1, although not at first glance. Table 1 is based on a specifically chosen standard definition (SD) ratio that states that the CBR rate for "hard" content is established at 50% higher than the rate for "easy" content. For HD, at four programs/QAM, we would consider a CBR (considering overhead) of about 9.5 Mbps at MPEG-4 as a reasonable over-provisioned rate. Now, consider the six streams in the example shown in Table 3, and note the "easy" content – traces 14, 11, 1. The average of this set is about 2.2 Mbps, and the peak-to-average is nearly the same as the prior 4-stream example. Using the same relationship of average, peak, and allocation, this would establish "easy" HD at about a 3.70 Mbps (i.e. also about 60% utilization).

Table 3 – Mixed HD Content on a Single QAM

	Mean Rate (Mbps)	Peak Rate (Mbps)
13hd (Motorola trace) / (trace 1)	3.53	4.68
sony720_G12B2FxT22 / (trace 5)	6.50	13.87
Mars –segment 1 / (trace 6)	3.46	12.72
Mars – segment 3 / (trace 8)	6.25	16.20
Horizon – segment 1 / (trace 11)	1.64	5.63
Horizon – segment 4 / (trace 14)	1.50	6.52

To compare to Table 1, we need to begin, for example, with a 50/50 "Existing Program Mix" of "easy" and "hard." As such, assume the initially configured 50/50 CBR HD being composed of traces 1, 5, 8, 11. Remaining from the six are now one "hard" and one "easy," which means 50/50 also for the columns labeled "Added Program Mix." The analogous Table 1 column says that a 50/50 existing mix and a 50/50 added mix translates to a 20% stream count gain. However, seeing that our CBR hard-to-easy ratio is closer to 2.6:1, instead of 1.5:1, Table 1 instead becomes Table 4 below.

Going from 50/50 existing, to adding 50/50 with the savings from CBR, we in fact would expect 44% gain, for a total of 5.8 streams – nearly 6 streams. Of course 6 streams means adding two more, as this simulation has shown to be accurate. Looked at another way, the precisely 50/50 new stream case occurs when the existing ratio is somewhere between a 50/50 and 60/40 hard-to-easy ratio.

Finally, consider the case of only *low* action HD. The model uses the relevant subset of the HD traces used in [2], plus some not previously used traces to build up enough low-complexity content to fill a channel. Table 5 shows the line-up used for this example.

Table 4 – Efficiency Gain, HD – Two Video Classes Example

Added Programming Mix	Existing Programming Mix				
	70/30	***60/40***	***50/50***	***40/60***	***30/70***
70/30	74.6%	64.0%	53.3%	42.6%	32.0%
60/40	67.4%	57.8%	48.2%	38.5%	28.9%
50/50	61.5%	52.7%	43.9%	35.2%	26.4%
40/60	56.5%	48.5%	40.4%	32.3%	24.2%
30/70	52.3%	44.8%	37.4%	29.9%	22.4%

"Easy" = 3.7 Mbps
"Hard" = 9.5 Mbps

Table 5 – Low Action HD Line-up

	Mean Rate (Mbps)	Peak Rate (Mbps)
13hd (Motorola trace) / trace 1	3.53	4.68
06hd (Motorola trace) / trace 3	4.20	4.44
Mars –segment 1 / trace 6	3.46	12.72
Horizon – segment 1 / trace 11	1.64	5.63
Horizon – segment 2 / trace 12	1.55	2.93
Horizon – segment 3 / trace 13	1.57	2.77
Horizon – segment 4 / trace 14	1.50	6.52
Blueplanet – segment 1 / trace 51	1.7	12.2
Blueplanet – segment 2 / trace 52	1.9	8.01
Blueplanet – segment 3 / trace 53	2.03	9.64
Blueplanet – segment 4 / trace 54	2.18	7.18

In this case, 10 HD programs were able to be multiplexed in a single QAM channel, and an 11th channel is nearly able to be added. Drop-free transmission was possible for 11 channels, but only if the CPE buffer was allowed to exceed the maximum allowed by our definition (500 msec) at 750 msec. However, with bit rate capping at 7 Mbps (4 traces impacted, 2 with reservations), the buffer was able to be held within 500 msec. Mitigation of the peak excursions is shown in Figure 6.

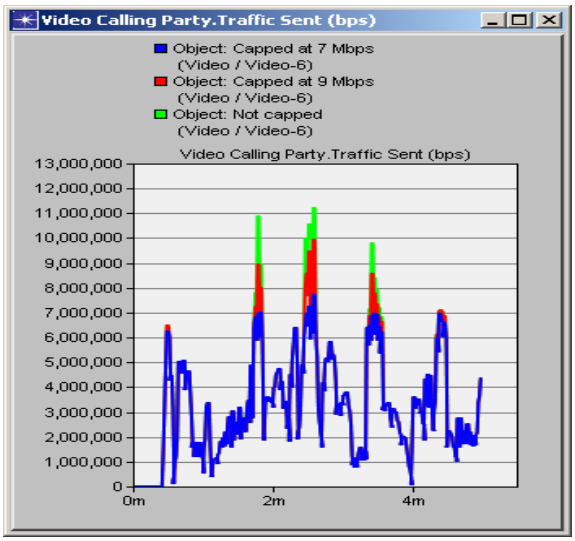


Figure 6 – Impact of Capping on Peak Excursions

A simple comparison to Table 1 and Table 4 can be made without creating a new table of possibilities. For 100% hard content, as discussed, we have four programs. If "easy" programs are 2.6 times as efficient, then we should have 2.6 times as many programs, or 4 x 2.6 = 10.4 programs, when all HD is low action. Indeed, we have shown that 10 streams are obtained, and almost 11.

On a broader basis, the ability to stream anywhere from 4 to 10 HD channels on a single QAM, depending on content type, again points out the high dependency of bandwidth efficiency to content type. The observed gains vary from -33% to +67% stream count efficiency if we consider as the baseline the mix of 6 HD streams, and consider that 4 streams fit when content is all high action, and 10 streams fit when the content is all low action.

CMTS Configuration for Video Traffic

It has been discussed often how video traffic characteristics differ in important ways than web browsing traffic. In general, video traffic is characterized by longer packet sizes and a more consistent rate of arrival. As such, the way a CMTS is configured for a voice + data mix is sub-optimal for how it might be configured in video-only mode. However, video frame size statistics are very complex, and video is much less tolerant of any issues in delivery. A mixture of video, voice, and data would be more complex still.

For modeling purposes, we will again consider the simple case at this point – assume that a four channel-bonded downstream is supporting video traffic only, a likely scenario initially for an MSO rolling out a managed IPTV service. The two primary mechanisms of packet drop are overflowing the transmit buffer, and excess delay that does not support buffer margin allocated on the receive side. Delay in the transmit buffer that is nearing the limit of time-to-live (TTL), and is determined unlikely to make it to the CPE in time, can also be dropped so other packets can be serviced, creating a secondary transmit-side packet loss scenario.

We choose the HD-only program line-up shown in Table 6, consisting again of segments of the trace library in [4] and Motorola-created segments. Some of the streams encodings are SVC, in order that the downstream channel would be filled at or close to its expected utilization for comparison and statistical purposes.

Table 6 – HD Streams on DOCSIS 3.0 Downstream

	Mean Rate (Mbps)	Peak Rate (Mbps)
13hd (Motorola trace)	3.53	4.68
05hd (Motorola trace)	9.64	15.25
06hd (Motorola trace)	4.20	4.44
sony720_G12B2FxT22	6.50	13.87
Mars –segment 1	3.46	12.72
Mars – segment 2	5.02	20.94
Mars – segment 3	6.25	16.20
Mars – segment 4	5.11	12.45
T2720_G12B2FxT22	5.43	12.04
Horizon – segment 1	1.64	5.63
Horizon – segment 2	1.55	2.93
Horizon – segment 3	1.57	2.77
Horizon – segment 4	1.50	6.52
Blueplanet1080_G16B3c – segment 1	2.98	6.83
Blueplanet1080_G16B3c – segment 4	4.68	17.7
Blueplanet1080_G16B3c– segment 3	5.26	13.3
Blueplanet1080_G16B3c – segment 2	5.28	11.98
Transporter2_1080_G16B3c – segment 1	10.24	29.84

This mix fits comfortably in the DOCSIS 3.0 channel with no packet loss issues, using a CPE buffer size of 300 msec. A buffer size of 200 msec works if rates are capped at 22 Mbps, which impacts only the extremely dynamic *Transporter 2* clip. While the vast majority of the time there is acceptable delay to the end user (we do not account for phy layer delay in our simulations), a portion of the aggregate traffic experiences a spike that dominates network performance during our 5-minute segment. The average transport packet delay experienced at this peak of the aggregate transmission burst is about 160 msec, and thus the reason the buffer size moving from 200 msec to 300 msec can make a difference. The rest of the sequence generally stays below 40 msec.

Rate Limiter Adjustments & Peak Bursts

In order not to vary CPE buffer size, which may exist in the field or be otherwise fixed due to memory limitations (such as to 200 msec), we can alternatively modify configuration parameters of the CMTS to accommodate the expected increase in burst size of video frames and avoid packet loss. The model also allows us to see what happens as streams align themselves unfavorably – peak bursts aligned – such that even this modest traffic load from a utilization standpoint (about 54%) can encounter congestions. We can then compare with tools available to mitigate this scenario. Let's examine these two scenarios.

Figure 7 shows two cases of traffic injection. On the left is aggregate injected traffic aligned through just the random time selection of segments from the library, versus on the right where they are slid around to create the most stressful network condition at roughly the 4.5 minute mark. The right-hand side of Figure 7, where the peaking is deliberately aligned, also shows the received aggregate traffic, pointing out the region of a lost burst of packets.

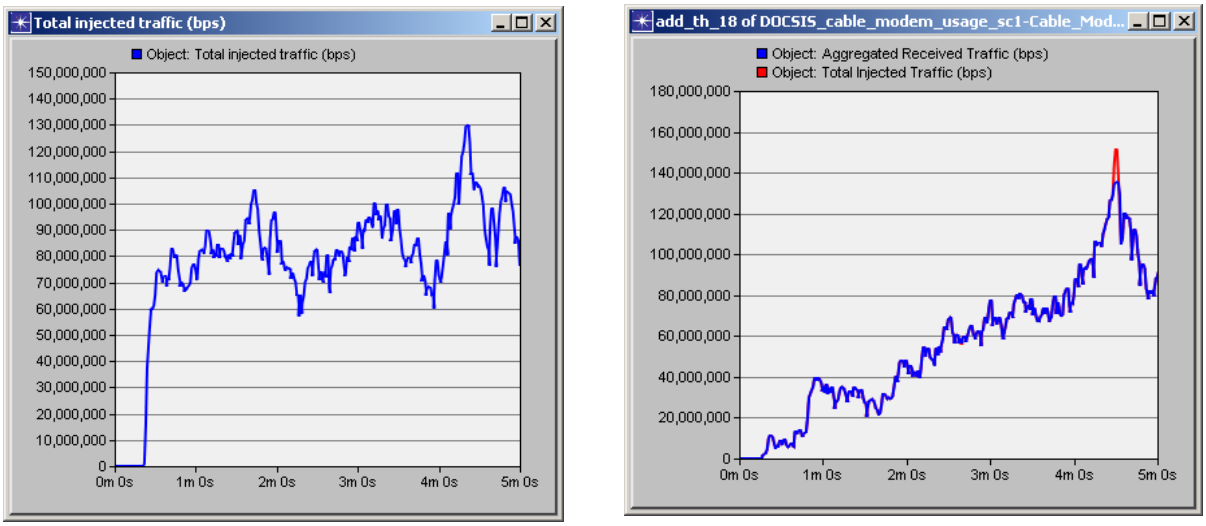


Figure 7 – Injected Traffic – Random (L) and Misaligned (R)

Now consider Figure 8, which shows the worst case packet delay observed between the two cases of stream alignment. In the top figure, we have increased an internal rate limiting function to provide a higher peak, so as to not allow a large burst to hit a stop sign on the way to the scheduler, meaning less opportunity for a large frame to be truncated. The result is that the *maximum* delay is dropped to about 65 msec (from the 160 msec avg in the section introduction).

In the lower figure of Figure 8, it is immediately obvious why this scenario could cause packet loss. We see delay exceeding at least 400 msec at the peak burst, even though the bulk of the time the network delay performance is quite sufficient. As would be expected given the perfect misalignment, a spike of traffic at the 4.5 minute mark is the cause of congestion and loss. While this example was deliberate misalignment, it was done to replicate a potentially realistic scenario for unmanaged streams, given that these are only five minute segments. Such a scenario becomes statistically more likely when the five minute span is scaled over by long periods of time and content mixes.

Quantum & A-Priori Knowledge

Another scheduler parameter that can be used to take advantage of the more predictable range of input traffic from video is the round robin quantum. In addition, some knowledge about the stream, either as a stored asset or gathered in near-real time, can be used to manage congestion and performance. We examine these cases here.

Figure 9 shows a comparison of increasing the quantum from several maximum Ethernet frames to the order of 100 Kbytes. The latter reduces the maximum observed delay, the important parameter in order that we do not drain buffers and ultimately starve decoders, by about 33%. This quantum size is enabled by the single class of service, which is also supporting only a single service type. Thus, only rounds of service are lost at the benefit of a high probability of fully servicing. But, as a single class and service type, there is decreased concern for the unfairness this can cause to smaller bursts. The former (the top half of Figure 9) results in a maximum observed delay of almost 100 msec, compared to about 65 msec when anticipating video-only traffic.

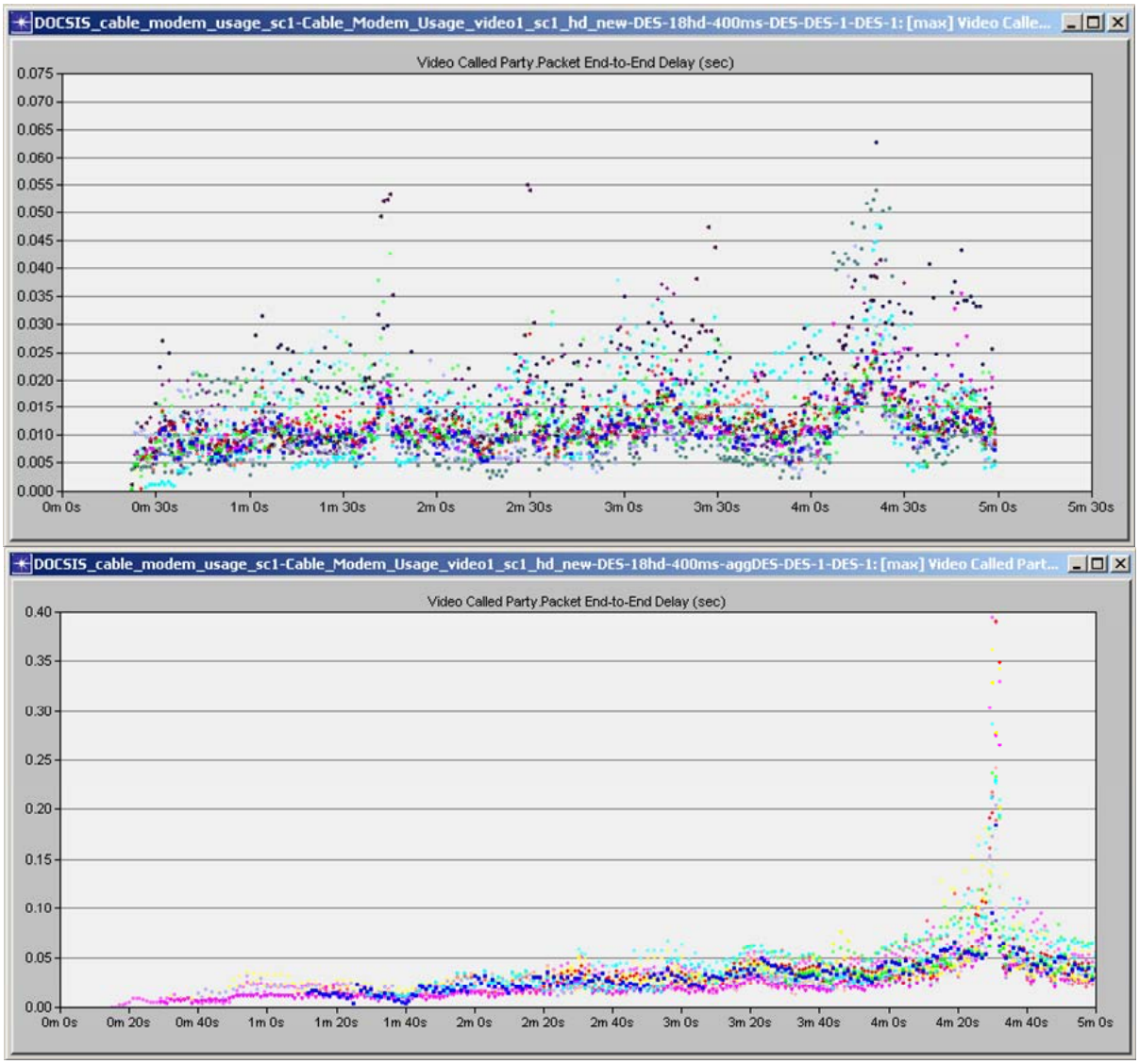


Figure 8 – Packet Delay – Decreased Rate Limiting (T) and Misaligned (B)

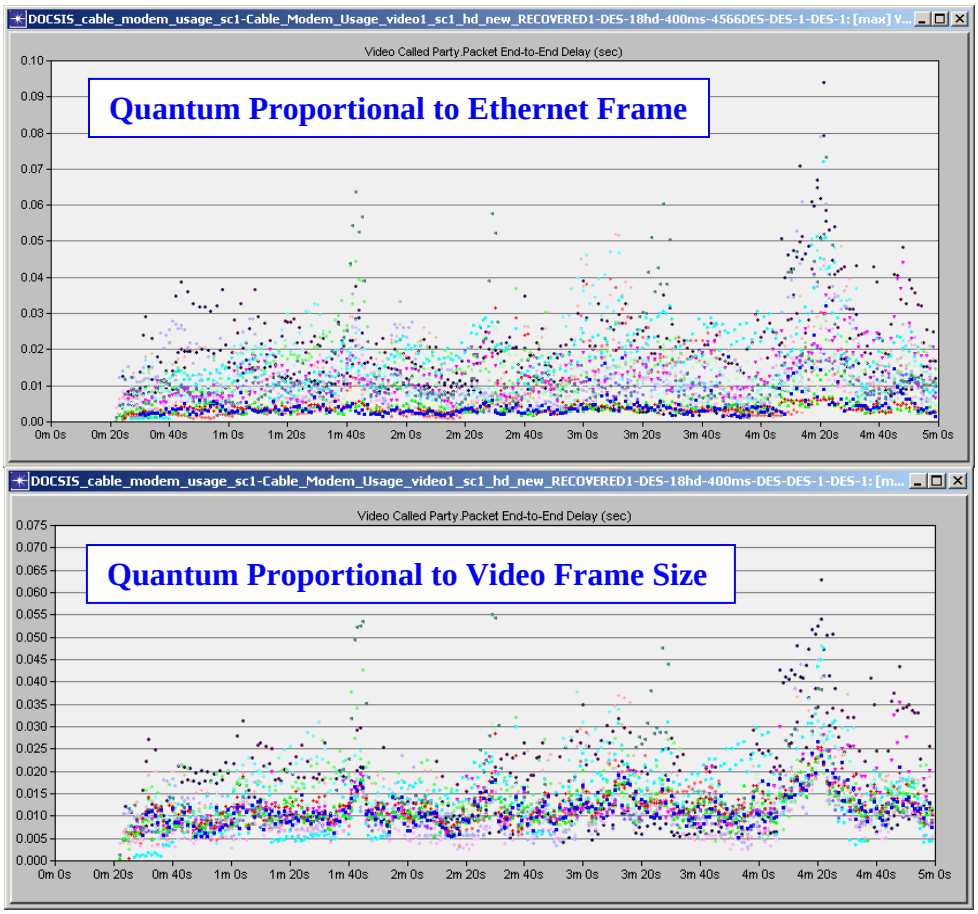


Figure 9 – Packet Delay vs Quantum: Scaled to E-net Frame vs Video Frame

Ideally, video streams would be accompanied by an array of metadata advertising their statistics. And, for stored assets, there is nothing in principle from comprehensively characterizing a stream statistically. However, the statistical variation for video from stream to stream and within a stream is large, and aggregation allows the law of large numbers to come into play. Thus, the added complexity beyond first and second order moments is typically not undertaken.

Some simple constraints, such as maximum size frame and peak arrival rates, can go a long way towards a deterministic network response (or more accurately a deterministically bounded response), if the constraints themselves can be guaranteed – which is a big "if." An example is shown in Figure 10, where we have added to network stress by once again choosing the worst case alignment of streams shown in Figure 7 (R) and Figure 8 (B). Though phase aligned for peaks, we have in this case capped the maximum frame size, set the quantum according to it, and assumed that we know the servicing rate (internal) and arrival rates (external). With that set of constraints, the delay becomes an arithmetic problem. That is, if we know how large the packets can be, how frequently they arrive, and how quickly we can service them, it is straightforward to calculate what's in a queue and the delay in servicing that queue, which is key to delivering on time and under budget. Figure 10 shows the precision for which we can assure a particular behavior under a set of assured constraints for a simulated and calculated queue size.

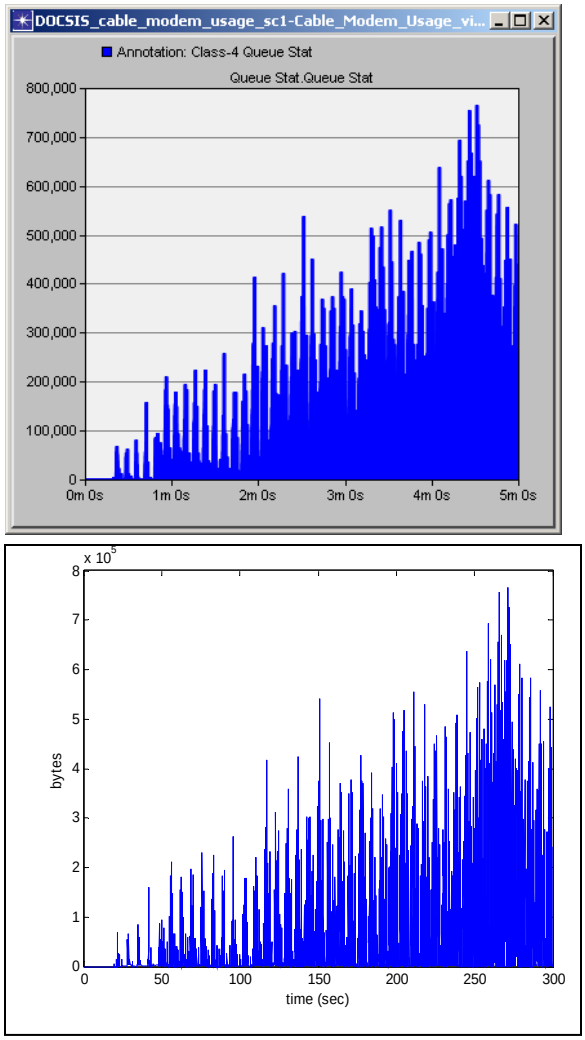


Figure 10 – Simulated vs Calculated Queue Size with A-Priori Traffic Stats

Admission Parameters - Summary

The ability to calculate queue size and delay from a-priori knowledge of the traffic statistics, versus knowing it with statistical confidence, is the difference between assured delivery on admission control decisions based on known delay bounds, and decision with some probabilistic confidence level. In the latter case, where the statistics are not assured, the more confidence is desired, the lower the efficiency of channel utilization will be. CMTS scheduling, and in particular HPRR scheduling [5], offer means to increase the confidence level of the statistical assurances, by supporting a best effort queue when the flow specification constraints are exceeded by a video stream. When servicing the excess video through a default queue, the relative delay will be impacted and be unpredictable, particularly if HSD services are added to the mix, adding stress to meeting the CPE delivery interval. However, the overflow queue provides an opportunity to successfully deliver packets that may otherwise be dropped when the statistics of the incoming streams cannot be assured.

We have quantified and simulated how maximum rate limiters, quanta, time-to-live counters [2], and (not shown here) minimum reserved rates can be combined with traffic characteristics to simulate network performance and result in quantifiable end-to-end network behavior. The multiple permutations of these relationships can be used to guide admission control decisions based on anticipating the impact of a new flow. While admission decisions have not been reduced it to a closed form expression – more of a multi-dimensional look-up table – the makings of the algorithm are as follows: calculate stats and existing workload, evaluate delay bounds/adjust, admit/deny/redirect source. The "redirect" step applies to the upcoming section introducing adaptive streaming into the model. Clearly, the more stream knowledge available a-priori or estimated directly the better, to the point of deterministic behavior for truly known statistics. Completely unknown inputs leave only mathematical characterization of how MPEG-4 AVC streams behave, as described in [2]. Unfortunately, this leaves a huge and impractical statistical range to accommodate.

Introducing Adaptive Streaming

The inclusion of adaptive technology as an emerging IP video tool promises new flexibility through a forgiving answer to the difficult yes/no admission problem, by serving up an answer that, instead of "No," can instead be "Yes, if…." We now take a closer look at how adaptive streaming technology influences IP video streams by incorporating a simple version into the model.

Consider a simple adaptive streaming model, where we are able to rate adjust the video, in this case using two different quantization levels. The basics of adaptive streaming, and in particular how it fits within the cable industry, are described in [3]. The stream multiplex is the same 18 HD VBR multiplex used in Table 6, using a four channel bonded DOCSIS 3.0 downstream. Note once again that the stream library used [4] is described in [2], and offers multiple format and encoding types to choose from. One particular video stream, a clip from *Transporter 2*, has a peak transmission rate of nearly 30 Mbps for high quality (high Peak Signal-to-Noise Ratio, or PSNR) as shown in Figure 11 (Q = 22), and higher still for even finer quantization.

Figure 12 shows the aggregate sent and received traffic with and without the adaptive mode turned on. Note how the initiation of the adaptive mode mitigates the peak burst that must be handled, reducing the "sent" volume. The resulting received traffic sequence now precisely follows the sent pattern through the now-reduced peak excursions. Note that a scale change of y-axis was used to expand on the tracking of the burst peak in the 250-265 sec range. A graphic artifact of the scale change is the ability to identify both sent and received flows on the "Adaptive" figure in red and blue, whereas on the wider, "No Adaptive" scale where there is overlap, the "received" traffic becomes hidden behind "sent" where they track.

.

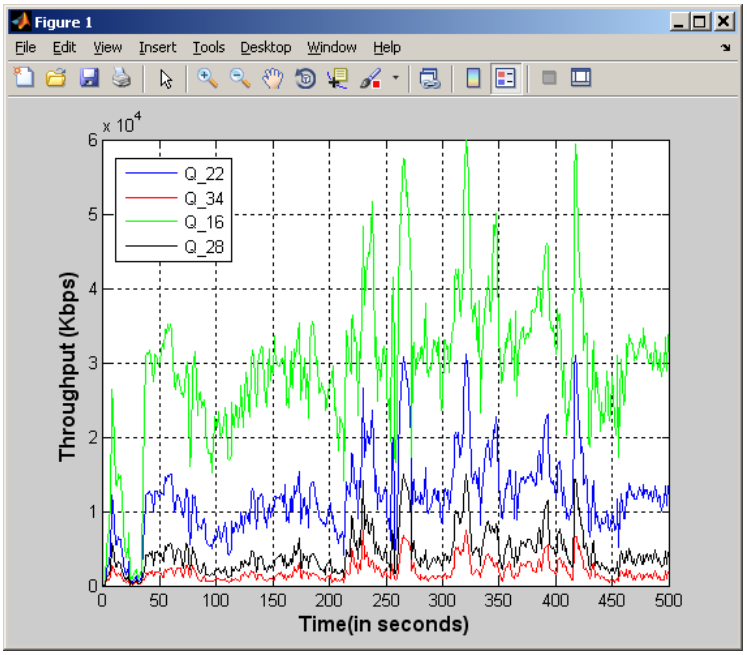


Figure 11 – Bit Rate vs QP for *Transporter 2* Clip

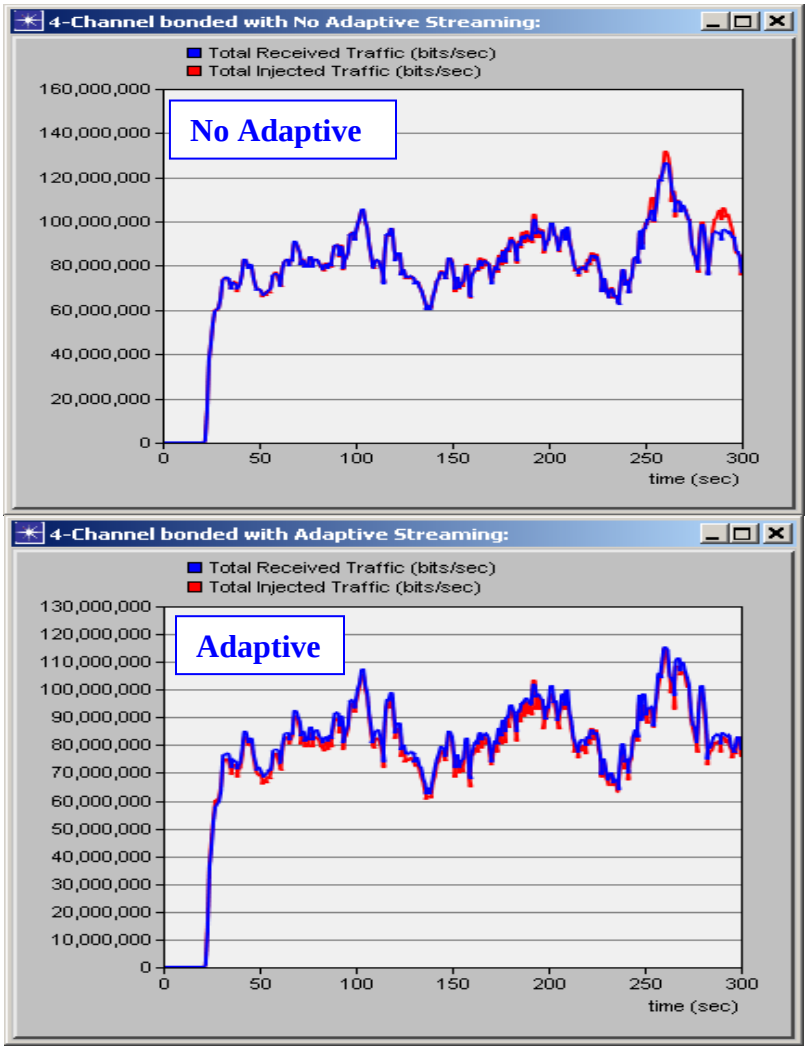


Figure 12 – Total Traffic Sent and Received: No Adaptive vs with Adaptive

The experience for the individual user watching *Transporter 2* is shown in Figure 13. (Editorial note – observing these bit rate plots of *Transporter 2* is actually more entertaining than watching *Transporter 2)*. The user suffers temporary picture loss during 250-265 and 282-297 sec time periods, identified by the top figure of Figure 13. We can estimate from Figure 11 that the latter period is likely not due to his or her movie, but instead likely due to peaks associated with others in the multiplex.

On the other hand, when adaptive streaming is turned on, the video server changes gears and sends a lower rate video clip than the primary stream. In this case, an encoding at Q_28 versus Q_22 is used. The resulting ability of the received traffic of the end user to follow the adaptive sent stream is shown in Figure 13, lower figure. The end-to-end packet delay is also lowered 22% compared to the non-adaptive case. The significance of this decrease is that it is another degree of freedom in system design – the trade-off of adaptive rates, or essentially transient video quality variation, in exchange for shorter buffers on the CPE side. In this example, a 400 msec buffer could have been reduced to nearly 300 msec. This can in turn translate to better user response for IPTV channel change.

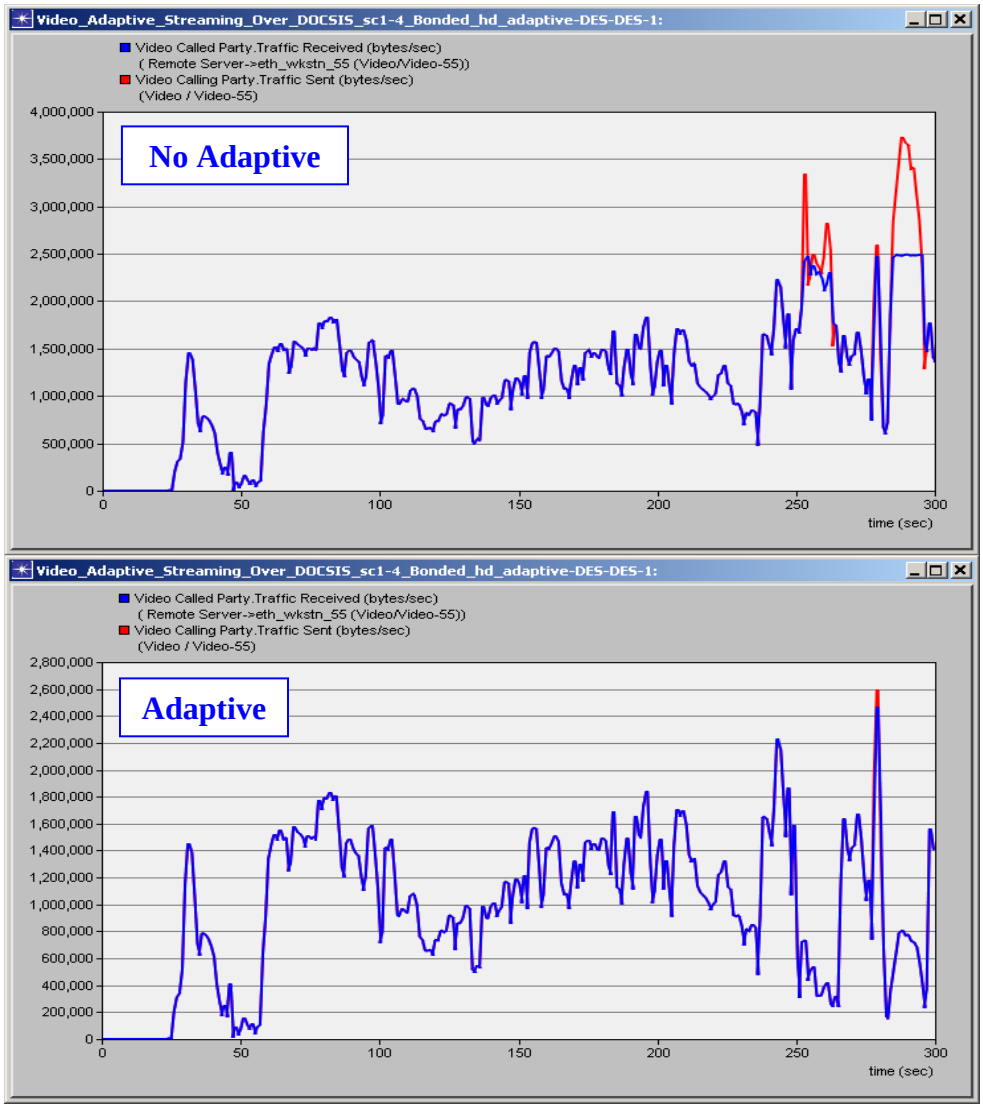


Figure 13: User Received Traffic: No Adaptive vs with Adaptive

VIDEO QoE WITH ERRORS

In the above simulations and the prior results referenced, different content types, formats, network variables, and technologies were permutated to understand the trade-offs involved in delivering low or no packet drop video service to IPTV users. Setting a drop threshold (1e-6), the assumption is that this threshold was chosen low enough to enable reasonable recovery mechanisms to handle clean-up. In the case of buffer size variation, 500 msec was assumed to be the maximum of what could be considered tolerable given system responsiveness needs. It should be noted that buffer sizes in terms of time translate to different memory sizes for different content types.

This section deals with the fallout of imperfect IPTV delivery when errors and drops ensue. We evaluate at the bit, byte, and packet level, where the latter would be the likely manifestation of congestion-oriented errors, and the former physical layer oriented. The byte error case can go either way, depending on other variables.

The current video delivery architecture, based on constant bit rate (CBR) streams encoded in MPEG-2, and using MPEG-2 TS over QAM transport, has some important operational advantages:

1) Simple traffic engineering and bandwidth management (CBR)
2) Low transmission errors (256-QAM)
3) Error resiliency (ITU J.83 encoded)
4) Assured timing/synch control (MPEG-2 TS)

For IP delivery, the advantage of a robust downstream physical layer remains. However, as has been discussed, bandwidth management aspects and timing assurances become more complex because of the use of VBR delivery, the dynamics of IP scheduling mechanisms designed for HSD, and the statistical probability of congestion that is not a component of existing video delivery.

To underscore the intolerance of video to transmission errors and packet drops, for which IP impairment mechanism would be randomized and potentially very harsh, a series of tests were performed whereby bit errors, byte errors, and transport packet loss was introduced into MPEG-2 and MPEG-4 video streams to observe how the displayed stream reacts. The impairments were introduced at steadily increasing rates and/or magnitudes. Subjective assessments were made by deliberately untrained eyes (not Video Quality (VQ) engineers) to better represent an average viewer experience. We do not proclaim equivalence to mean opinion score (MOS) levels of confidence, but the goal was to be more aligned with the home experience rather than the lab "find-the-irregularity" experience. A good lesson learned is never watch TV with a VQ engineer if you want to enjoy a program – they will find things that only Steve Austin (Google it, post baby-boomers) would otherwise identify.

Figure 14 shows a block diagram of the test setup.

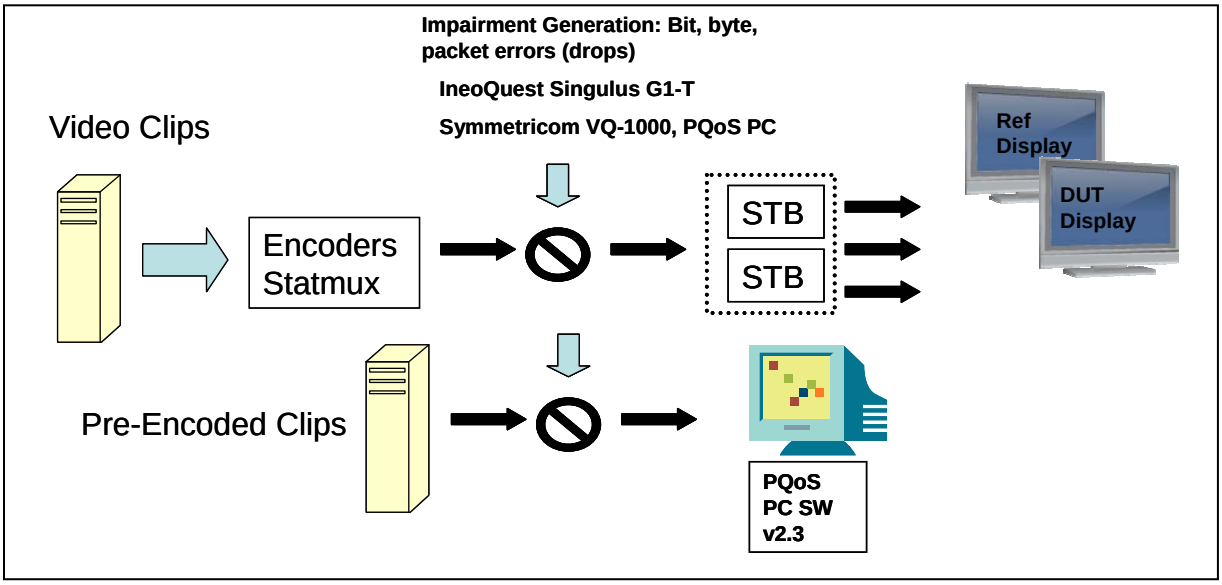


Figure 14 – Impairment Generation and Video QoE

Testing of digital video artifacts as a function of link quality and impairments exists throughout the technical literature. The results discussed here are not meant to recreate years of prior evaluations, but are primarily to provide a basis of observed actual content consistent with the mix and approach used in the simulations for comparison.

In addition, MPEG-4 part 10 encoding, while deployed in telco IPTV architectures, is still relatively early on the learning curve, and has evolved even since current deployments. Thus, any new insight observed adds to the growing library of experience with this standard.

Finally, some of the analysis tools, such as the latest revision of the Symmetricom PQoS video software analysis tool used, are also relatively new. Observations and results based on this tool thus offer potentially new data points in the continually evolving arena of subjective video quality analysis.

In addition to periodic packet dropping identified in Figure 14, implemented using the IneoQuest Singulus G1-T, the device also includes a test mode for rate reduction via dropping by hard peak capping. Though not described herein, this mode is an insightful complement to the testing described above, and represents the starkest possible contrast to the kind of intelligent rate control used by encoders, whose job is to maximize video quality at a particular bit rate allocation. Between the un-informed effects of IP video congestion delay, error, and bandwidth constraints, and VQ-based rate control, we have the ability to compare the best case and worst case ends of the impairment effect spectrum.

Bit and Byte Errors

Because of logistical constraints in the laboratory and VQ analysis tool, only MPEG-2 encoded content was available for the bit and byte error assessments. For MPEG-4 encoding, there are two obvious variables in play with respect to how it would compare, relatively speaking:

1) MPEG-4 is roughly one-half the bit rate on average, and higher in peak-to-average. Therefore, the same

periodicity of errors will effect twice as much, or nearly so, of the content from a time of occurrence and % of errors perspective

2) MPEG-4 has additional sophisticated filtering mechanisms design to reduce blur, halo, motion, and edge effects of block transform compression techniques. It is likely that these filters would act to positively impact potential artifacts (i.e. help to conceal them).

No further research (literary or test) was investigated as to whether, or under what conditions, these two factors cancel one another, or if one carries more weight.

Table 7 describes qualitatively the results of creating bit and byte errors for two types each of selected news-like and sports content using MPEG-2 encoded 720p HD.

For these streams, it is straightforward to observe from Table 7 that once bit or byte error rates stay below the 1e-6 range, viewing is unimpaired. Of course, bit transmission errors of this order and at least a couple orders of magnitude lower are generally well handled by FEC, particularly errors of the random type. When the effect of FEC is included, relatively graceful degradation such as observed in Table 7 gives way to perfect-or-objectionable, due to the nature of the FEC function. While FEC adds dBs of margin at a given error rate, it does so while steepening the error rate curve as a function of SNR.

For example, without FEC, a 1024-QAM downstream needs about 40 dB to achieve 1e-8 error rate. Referencing the "artifact-free" case from Table 7, it can achieve 1e-6 at about 38.2 dB, or about 2 dB lower. For an FEC applied that offered 3 dB of coding gain at 1e-6 (35.2 dB SNR required), we might find that the 1e-8 is achieved post-error correction at 35.7 dB, or a half dB different. The exact amount would vary by FEC architecture, but this represents the "steepness" effect. No specific architecture was called out here, because it is likely that when 1024-QAM arrives, there will be much discussion around deploying newer FEC structures than the now-dated ITU J.83 standard, such as newer low-density parity check codes (LDPC codes).

.

Table 7 – Bit & Byte QoE on MPEG-2 Encoded Content Types

	Content Type	
Bit Errors - One Per N Packets	"Easy" = News	"Hard" = Basketball
N = 100	Multiple simultaneous line and small block artifacts, constant	Multiple simultaneous line and small block artifacts, constant
N = 1000	Same as N = 100, just lesser in quantity, constant	Same as N = 100, just lesser in quantity, constant
N = 10,000	Same effect as N = 100,1000 but at roughly 5 sec intervals	Same effect as N = 100,1000 at ~5 sec intervals, less obvious (masked)
N = 100,000	One or two small block artifacts (mostly) per minute	None observed
N = 1,000,000	Nothing observed	Nothing observed
Byte Errors - One Per N Packets	"Easy" = News	"Hard" = Basketball
N = 100	Same effect as N = 100 "bits," aggravated occasional near break-up	Same effect as N = 100 "bits," aggravated occasional near break-up
N = 1000	Same effect as N = 1000 "bits," increase in block errors vs lines	Same effect as N = 1000 "bits"
N = 10,000	Same effect as N = 10,000, more often ~ every 2-3 seconds	Line artifacts observed every few seconds
N = 100,000	Line and small block artifacts every 15-30 seconds	One or two line artifacts per minute observed
N = 1,000,000	Nothing observed	Nothing observed

Byte errors stress the burst correction and interleaving elements of the receiver processing. Although in this testing we corrupted one byte at a time, multiple or consecutive byte error testing may be required to more finely define a threshold for impairments with this type of impact. Typically, error mechanisms are likely to be plant transient effects such as impulses of interference, power related spikes, or equipment malfunction anyplace there is electronics connected to the coax network. For the most part, byte errors acted like a worsened case of the same bit error rate, as might be expected. In the poorer cases of byte error frequency, another difference was that byte errors have a probability of resulting in a complete display break-up through overwhelming of the decoder's ability to make sense of the incoming information and effectively undo the encoding process.

An important phenomenon associated with the differences between objective measurement and perceptual experience is also apparent in the above table. That is, spot "popcorn" pixilation is more easily masked in some types of complex scenes. An "average viewer" would identify with three key elements on the screen that contribute to complexity:

1) Block-to-block detail granularity
2) High contrast sharp edges
3) Speed of motion

The first item above has the positive perceptual tendency of masking small macro-blocking when the brain is not expecting a pattern in the detail. In the sporting sense, the obvious example is crowd scenes, and even more so in panning crowd scenes as balls, pucks and athletes go past noise-like spectator backgrounds. As has been perceptually discovered over an over, people are wired to identify with patterns associated with prior experience, such as details in the action *on* the field, floor, or ice. Without a pattern to attract attention, fewer disturbances will be recognized. A second contributor to a better perception in this case is simply that the focus is on the match or game most of the time, not the spectators. This explains the better perception of more difficult content in the error injection tests in Table 7..

The part that suffers in the above example of high complexity is motion-related degradation. This tends to be associated, however, with constrained bit rate and infrequent packet loss, rather than the block and line pixelation associated with bit and byte errors. Note, however, that low motion scenes of great detail – also tested but not shown in Table 6 – where patterns *are* expected will get perceived differently than "noisy" detail. Examples are backdrops that involve high structure and high detail, such as cityscape or broad landscape scenes, or multitudes of faces at non-anonymous depth. In these cases, a perceptual expectation of detail is a prevailing factor.

Packet Drops – Effects & Recovery

Packet drops – MPEG or (worse) IP – show the intolerance of video delivery to packet loss. In doing so, it identifies the need for packet recovery mechanism when delivery cannot be deterministically assured, as is typically the case for IP data delivery.

Repeated Packet Drops

Table 8 describes qualitatively and quantitatively the packet dropping results of interrupting MPEG-4 AVC encoded Ethernet/IPv4/UDP transmissions of 720p HD content.

Table 8 – Packet Loss Impacts on MPEG-4 720p HD Streams

	Content Type	
Packet Drops - One Per N Packets	"Easy" = News	"Hard" = Basketball
N = 10	Freeze with no recovery	Freeze with no recovery
N = 100	Freeze with no recovery	SAME AS "EASY" CONTENT
N = 1000	Frames update every 1-2 sec	
N = 5,000	Freeze every ~5 sec with restart	
N = 10,000	Momentary freeze every ~7-8 sec	Momentary freeze every ~7-8 sec
One-Time Drop of N Packets	"Easy" = News	"Hard" = Basketball
N = 1	Macro-blocking and/or momentary freeze	Momentary freeze (no observed blocking)
N = 10	Momentary freeze	Momentary freeze
N = 100	Momentary freeze	Momentary freeze & sometimes artifacts on recovery
N = 1,000	1-2 Second freeze	1-2 Second freeze
N = 10,000	8-10 Second freeze	8-10 Second freeze

Repetitive (in this case with statistical regularity) packet loss creates frozen screens without recovery in the worst case, and periodic freezes of video in the best case – both clearly objectionable, in particular at the rates evaluated here. While unlikely, the repetitive case is valuable to observe in test because it ensures that we are statistically likely to encounter the effect of deleting an I-Frame. Loss of an I-frame certainly makes for more difficult recovery. In addition, while they represent a minority of the frames, I-frames could tend to be overrepresented as cases that cause congestion because of they are inherently larger than B and P frames.

Observing the impact given by the top-half of Table 8 gives a sense of the load that would need to be handled by an error mitigation mechanism, such as packet retransmission, due to link or routing related loss, going as low in this case of 1e-5 packet loss rate. Using a logical extrapolation from 1e-3 through 1e-5 effects, we would anticipate a momentary freeze on the order of a minute, give or take, for a 1e-6 case. Scaled by the user base served by an access device and/or servicing cache, this translates to some scale of processing load, memory, and signaling to manage for using packet recovery as part of a congestion management subsystem. This case (action every minute) would be the relevant relationship for an IP video system engineered based on the packet loss threshold defined in the simulations.

Single Burst of Packet Drops

The case of congestion based errors due to excess delay is more likely to lead to a series of packets being lost. Since behavior given a series of lost packets is important to understand, the bottom half of Table 8 shows the decoder recovery response when a one-time burst occurs that interrupts a packet sequence, varying in size from one to 10,000 packets. From 1-100 lost packets, the decoder recovery is roughly the same – instantaneous – following a momentary screen freeze. Nonetheless, this is an unacceptable experience, but also one likely every digital TV consumer has experienced. (Ironically, this happened to me just last evening watching from my DVR an excellent Episode 7 of the final season of "Lost." Unfortunately, the TiVo DVR in the middle muddies the water as to possible causes.)

Beyond the momentary freeze of relatively low loss events, momentary freezes become seconds of frozen screen when 1000 packets are dropped. Our simulations show that an unmanaged VBR congestion peak can create

a sequence of drops on the order of 100's and in some cases 1000's. These larger sequences of lost packets, again multiplied by the scale of the user base accessing the recovery mechanism, are again indicative of the order of memory, control, and network bandwidth necessary to create effective packet recovery.

There are similar, albeit less apparent, content based differences as a function of the size of the dropped sequence. For a small enough sequence of packet drops, the less likely case for network related condition, we see in Table 8 again that the complexity level contributes to some masking of the blocking effects of the decoder upon recovery. In contrast, as the burst event size increased to N=100, the complexity of the scene – in particular the motion complexity – created added stress on the decoder to recover with the fidelity of the low-complexity news content.

Video Rate Reduction

As previously described, a basic principle of video encoding is to maximize the quality for a given allowable bit rate of delivery. The science and literature is rich with objective and subjective analysis of the multiple variables in play to achieving this, and encoder manufacturers spend time and effort developing solutions that continue to optimize these relationships. In all cases, it is a basic premise to deliver optimal perceived quality (a subject all to its own) at minimum bit rate, which results in VBR streams at a given pre-determined quality.

Now, with the information available to encoders as part of the compression algorithm process – real-time knowledge of scene complexity – the reverse relationship can also be explored. That is, rather than starting with defining a desired level of quality, a given available bit rate can be the independent variable, from which an optimum perceived output can be derived. This basic premise is at the heart of adaptive streaming protocols aimed at improving the experience of Internet video – and more broadly video over IP in general. While our network simulations implement hard capping (i.e. network simulation tools pay attention to link and packet level performance, not video quality) capped VBR conclusions drawn in the simulations to ensure network performance would be applied with awareness just as in the adaptive case, as long as we can assure that a reasonable video quality is achievable within the capping limit defined. In the network modeling world, this correlates to ensuring a cap that is a reasonable rate reduction as a percentage of the peak.

CONCLUSION

Delivering video over DOCSIS, with the same QoE or better, as the existing MPEG-2 TS based infrastructure is critical for a successful transition of services. However, IP traffic has historically always been limited in its ability to provide deterministic delivery guarantees. Data and voice services, to an extent, are robust to some subset of the potential obstacles. Video is robust to none of them. There are some positives compared to voice service – the most notable exceptions being a pure latency advantage (less sensitive) and some added flexibility in dealing with jitter. These come at orders of magnitude differences in service processing bandwidth, however.

We have looked at an array of variables associated with IPTV delivery as the traffic engineering of this architecture begins to take shape. In particular, we observed once again the strong content dependence of video streams on the bandwidth efficiencies.

In addition to the external variables such as CPE buffer size, packet error metrics, and peak-rate capping, we took advantage of some predictability in a video-only service class configuration to see how making use of some of the basics of video packets translates into better network performance. We added adaptive capability to the simulation and observed how this enabled dynamic traffic injection to be more ably followed, packet-for-packet, at the CPE, with no information loss. Finally, we recognized that network simulations gathering numbers do not capture perceptual video effects. Thus, using a similar content mix basis, we took a look at what error and loss mechanisms mean to end user QoE, and how packet loss translates to display and recovery times. This information estimates the scale of the problem for a suitable packet recovery mechanism as a function of a range of packet loss rates and sizes.

Most importantly, we have further developed a very robust model that can be used to comprehensively understand all aspects of DOCSIS delivery of video services, and additionally can be used to evaluate performance for converged services over bonded DOCSIS 3.0 downstreams.

REFERENCES

[1] Dr. Robert Howald, "Fueling the Coaxial Last Mile,"2009 Society for Cable Telecommunications Engineers (SCTE) Emerging Technologies Conference, Washington, DC, April 3, 2009.

[2] Dr. Robert Howald, "Web Surfing to Channel Surfing: Engineering the HSD Edge for Video," 2009 Cable-Tec Expo, sponsored by the Society for Cable Telecommunications Engineers (SCTE), Denver, CO, Oct 28-30, 2009.

[3] John Ulm, Tom du Breuil, Gary Hughes, and Dr. Sean McCarthy, "Adaptive Streaming – New Approaches for Cable IP Video Delivery," Cable Show Spring Forum, May 11-13 2010, Los Angeles, CA.

[4] Fitzek, Frank H.P. and M. Reisslein, "MPEG-4 and H.263 Video Traces for Network Performance Evaluation," (extended version), Technical Report TKN-00-06, Technical University Berlin, Dept. of Electrical Eng., Germany, October 2000.

[5] Patrick, M., "Hierarchical prioritized round robin (HPRR) scheduling," US Patent No 7457313, August 2005.

ACKNOWLEDGMENTS

The authors would like to thank the following individuals for their insights and support in developing this paper: Sean McCarthy, Stan Egger, Xiahan Zhu, and Howie Ton-That, John Ulm, Erik Metz, and Mike Patrick.

DYNAMIC STEERING OF POWER-STARVED CMs, DSG-STBs, & MTAs

Ayham Al-Banna, Tom Cloonan
ARRIS Group, Inc.

Abstract

Power starvation of DOCSIS client devices is a serious problem caused by the insertion of a large number of RF splitters within the subscriber's home coaxial network. This condition introduces significant attenuation in both the US and DS directions, causing the power-starved devices to suffer from degraded performance and service.

Several solutions that address this problem already exist. However, these solutions are either suboptimal or impractical. We propose a novel solution based on dynamic steering of power-starved devices that does not suffer from any of the drawbacks found in the existing solutions.

INTRODUCTION

Since Multiple Service Operators (MSOs) offer various services over the Hybrid-Fiber Coaxial (HFC) network, subscribers tend to have multiple devices in their homes, including Cable Modems (CMs), DOCSIS® Set-top Gateway Set-top Boxes (DSG STBs), Multimedia Terminal Adaptors (MTAs), etc. As subscribers decide to expand their cable access to multiple devices and rooms within their houses, it is common for many new Radio Frequency (RF) splitters to be added within the home coaxial distribution network. This practice can lead undesirable "power starvation" for the devices receiving signals that are passed through these many RF splitters. Power starvation of devices presents challenging problems for MSOs as they strive to offer good Quality of Experience (QoE) service to their subscribers.

Multiple existing solutions to address the above problem are listed in this paper. While these solutions can overcome the problem of power starvation, some of them suffer from serious limitations that make them expensive, suboptimal, or impractical. We propose a novel solution, which does not suffer from any of the limitations present in the existing solutions.

WHAT IS POWER STARVATION?

In this section, the power starvation condition is defined and described. Let us start with a normal and operational scenario where all subscribers' devices (CMs, DSG-STBs, or MTAs) are placed behind a small number of RF splitters as shown in Fig. 1.

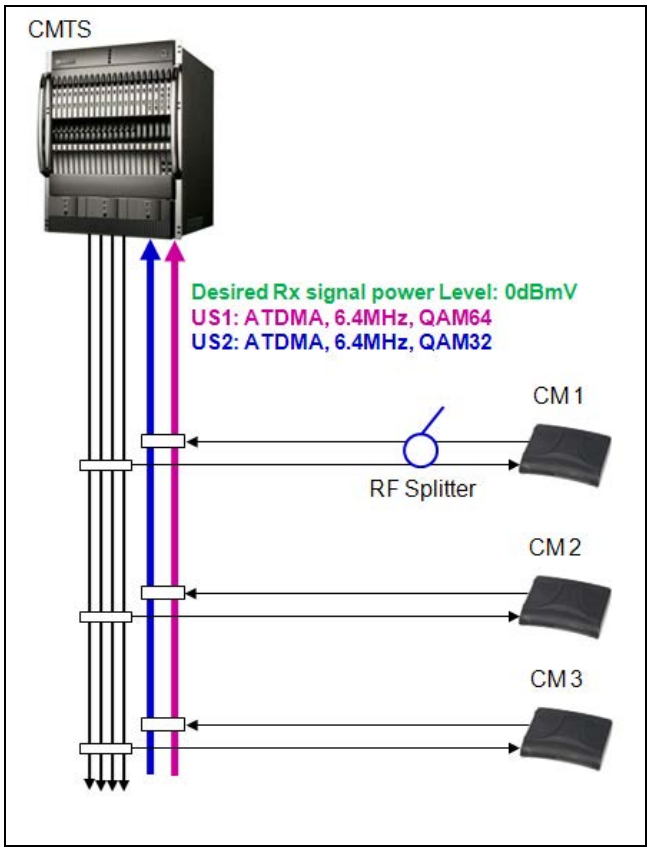


Figure 1. An operational Scenario, where 3 CMs are connected to a CMTS via 4 DS channels and 2 US channels.

All CMs are using Upstream Channel 1 (US1). There is a single RF splitter in the house 1 and no splitters in houses 2 and 3. The CMTS is configured with a receive signal power level of 0 dBmV.

In Fig. 1, we use CMs for illustration purposes while keeping in mind that the discussion also applies to DSG-STBs and MTAs. Herein, we assume that the distance between the CMs and the Cable Modem Termination System (CMTS) is small such that the receive signal power level is acceptable for all CMs when the number of RF splitters installed within each house is small. To simplify the discussion, we also assume that the houses are in close proximity to each other such that the propagation loss between these different houses is negligible.

Figure 1 shows that there is a single RF splitter inside house 1, while no splitters are introduced in the houses of the 2nd and 3rd subscribers. Observe that RF splitters not only introduce loss in the Downstream (DS) direction, but also in the Upstream (US) direction. In Fig. 1, we assume that there are two US channels configured on the CMTS and all CMs are using a single US channel. The two US channels are configured in the following fashion:

US1: ATDMA channel with 6.4 MHz Bandwidth and Quadrature Amplitude Modulation (QAM) 64

US2: ATDMA channel with 6.4 MHz Bandwidth and QAM 32.

All CMs are assumed to be using the US channel labeled US1.

During the ranging process, the CMTS instructs all CMs to adjust their Transmit signal power levels such that their signals arrive at the CMTS at the desired Receive signal power level, which is configured on the CMTS (0 dBmV in this example). This is shown in Fig. 2(a).

In Fig. 2(a), different CMs have adjusted their Transmit signal power level to compensate for the loss between them and the CMTS. We observe that CM1 is sending at higher signal level than that of CM2 and CM3 to compensate for the extra loss introduced by the RF splitter in house 1. In Fig. 2(b), all CMs are hitting the CMTS at roughly the same Receive signal power level. Assuming that there is no channel distortion, all CMs will have comparable Signal-to-Noise Ratio (SNR) values, as depicted in Fig. 2(c), because: 1) All CMs have comparable Receive signal power levels, and 2) The noise level experienced by all receive signals at the CMTS port is identical. Finally, Fig. 2(d) shows that all CMs have good performance because their operating points are well below the maximum acceptable Packet Error Rate (PER) value.

Next, we consider the more interesting scenario of power starvation. Assume that the subscriber that owns CM2 introduces two RF splitters on the cable before feeding it to CM2. Consequently, the CMTS will ask CM2 to increase its transmit signal power level such that the receive signal power level at the CMTS equals the desired value of 0 dBmV. CM2 responds by increasing its transmit signal power level and hits the CMTS at 0 dBmV. Assuming that the noise level did not change, observe that while the transmit signal power level is higher than the value in the operational scenario for CM2, the receive signal power level and SNR values are still similar to the corresponding values in the operational case. This is shown in Fig. 3, where CM2 is still operational with good service since its operating point is well below the maximum acceptable PER value.

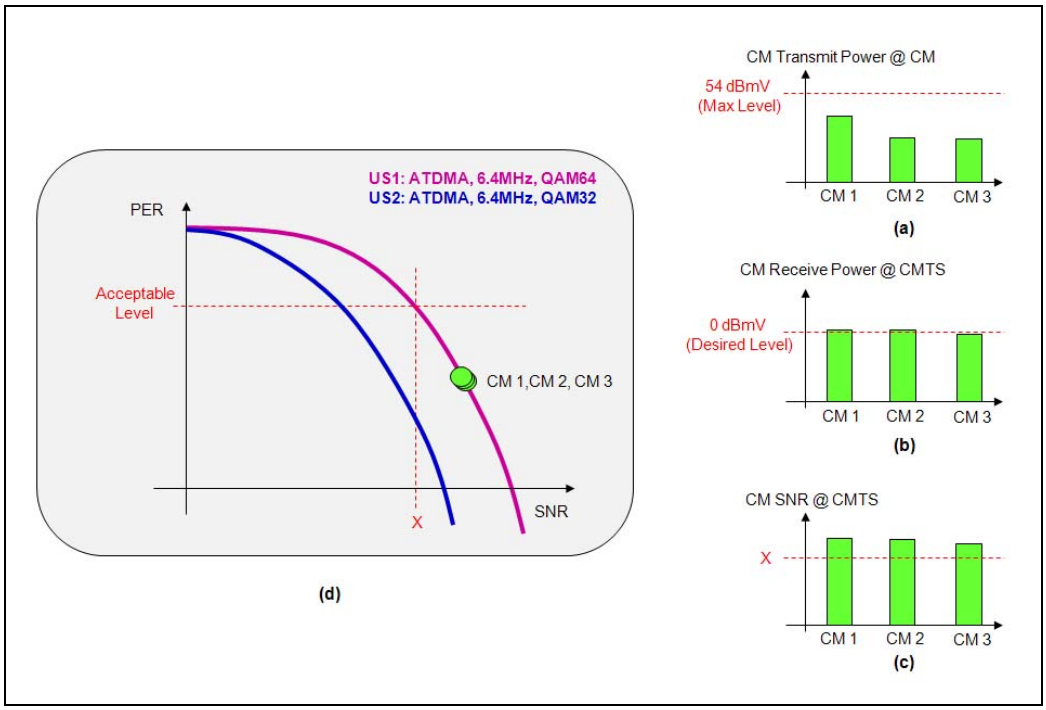


Figure 2. Different curves corresponding to the scenario in Fig. 1. (a) All CMs adjust their transmit signal power level differently to compensate for the attenuation along their way such that their signals hit the CMTS at 0dBmV. While CM2 and CM3 have comparable transmit levels, CM1 has a higher transmit power level to accommodate for the splitter loss. (b) The receive signal power level of all CMs is roughly equal to the desired level of 0 dBmV. (c) All CMs have high SNR values. (d) PER vs. SNR curve showing all CMs have large SNR values and therefore low PER values (below the maximum acceptable limit).

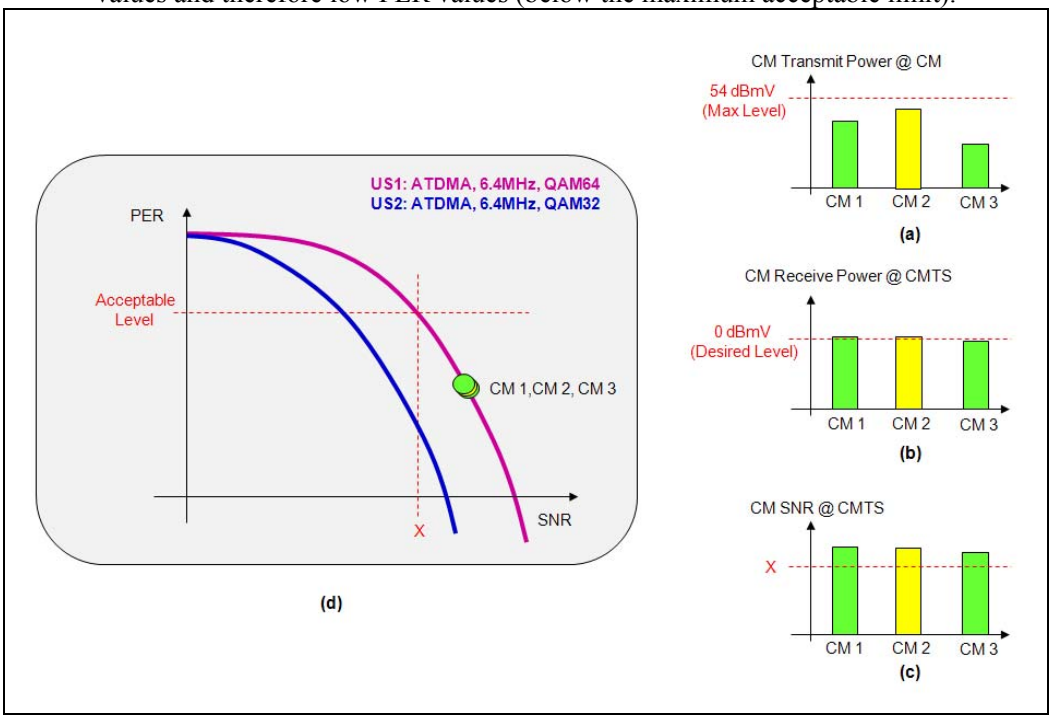


Figure 3. Different curves corresponding to the scenario in Fig. 1 but with 2 RF splitters added along the path of CM2. (a) CM2 increased its transmit signal power level to compensate for the loss introduced by the two splitters. (b) The receive signal power level of all CMs is roughly equal to the desired level of 0 dBmV. (c) All CMs have high SNR values. (d) PER vs. SNR curve showing all CMs have large SNR values and therefore low PER values (below the maximum acceptable limit).

Now, what happens if the subscriber that owns CM2 introduces a third RF splitter along the way, which will further increase the attenuation of RF signals that propagate over the cable. Once the CMTS measures a reduced received signal power level, it runs over the usual behavior of instructing the CM to increase its transmit signal power level such that the receive signal power level at the CMTS equals the desired value of 0 dBmV. As CM2 tries to increase its transmit signal power level to satisfy the CMTS's request, it eventually gets blocked by its own limitation to increase the level because every CM has a maximum permitted limit for the transmit power level (e.g., 58dBmV for QPSK, 54dBmV for QAM32, see Table 6-6 in [1] for more details.) This causes the receive signal power level of CM2 at the CMTS to be less than the desired value of 0 dBmV and therefore CM2 will experience a SNR value that is less than the SNR values of all other CMs on that US channel. Observe from Fig. 4 that all CMs on the US channel obtain good service except for CM2, which has some performance issues as seen from the PER vs. SNR curve. We refer to CM2 as a "power-starved" CM because it increased its transmit power level to its maximum level and yet was not able to hit the CMTS with the desired receive signal power level. Observe that the SNR value that belongs to a power-starved CM can be much lower than the average CM's SNR on that US channel.

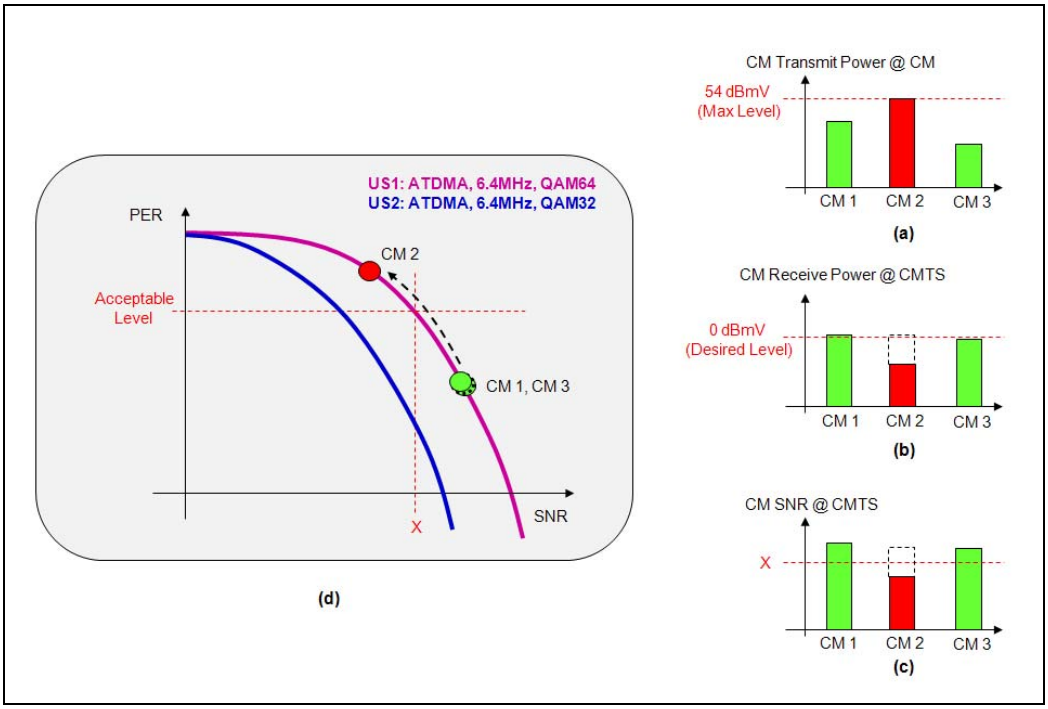


Figure 4. Different curves corresponding to the scenario in Fig. 1 but with 3 RF splitters added along the path of CM2. (a) CM2 increased its transmit signal power level to compensate for the loss introduced by the 3 splitters. However, it got clipped by the maximum limit that CM2 can transmit (54 dBmV in this example.) (b) The receive signal power level of all CMs is roughly equal to the desired level of 0 dBmV except for CM2, where the level is well below the desired value of 0 dBmV. (c) All CMs have high SNR values except for CM2 that has low SNR value. (d) PER vs. SNR curve showing all CMs have large SNR values and therefore low PER values except for CM2 that has low SNR value and hence large PER value (exceeding the maximum acceptable PER threshold).

EXISTING SOLUTIONS AND THEIR LIMITATIONS

There exist multiple solutions for the power starvation problem. However, none of these solutions is optimal as explained in this section. Normally, the MSOs have to choose from one of the several undesirable paths:

1. Do Nothing!

   Since the majority of the CMs on the US channel are receiving good service and only a small fraction of the CMs experience performance issues, one may think that this is acceptable. Unfortunately, this situation is not at all acceptable for the subscribers whose CM is power-starved, and could easily lead to customer churn. Thus, this is an expensive and impractical solution for the MSO!

2. New Modulation profile:

   Another solution to the power starvation problem is to design a new modulation profile (ex: more FEC correction, lower modulation order, narrower channel width, etc.) that can provide adequate PER values even in the presence of the low SNR values of the power-starved CMs on that US channel. This unfortunately yields lower throughputs for all CMs on the channel as shown in Fig. 5. This solution has several disadvantages including: 1) Degraded service (less throughput), for the non-power-starved CMs and 2) lower overall channel bit rates resulting in an upstream plant with lower efficiencies. This solution is not optimal!

3. The SCDMA MSC feature:

   The Synchronous Code Division Multiple Access (SCDMA) Maximum Scheduled Codes (MSC) feature is a good candidate solution for the problem of power starvation. The MSC feature limits the number of active codes used by the power-starved CM while keeping the transmit signal power level unchanged. This

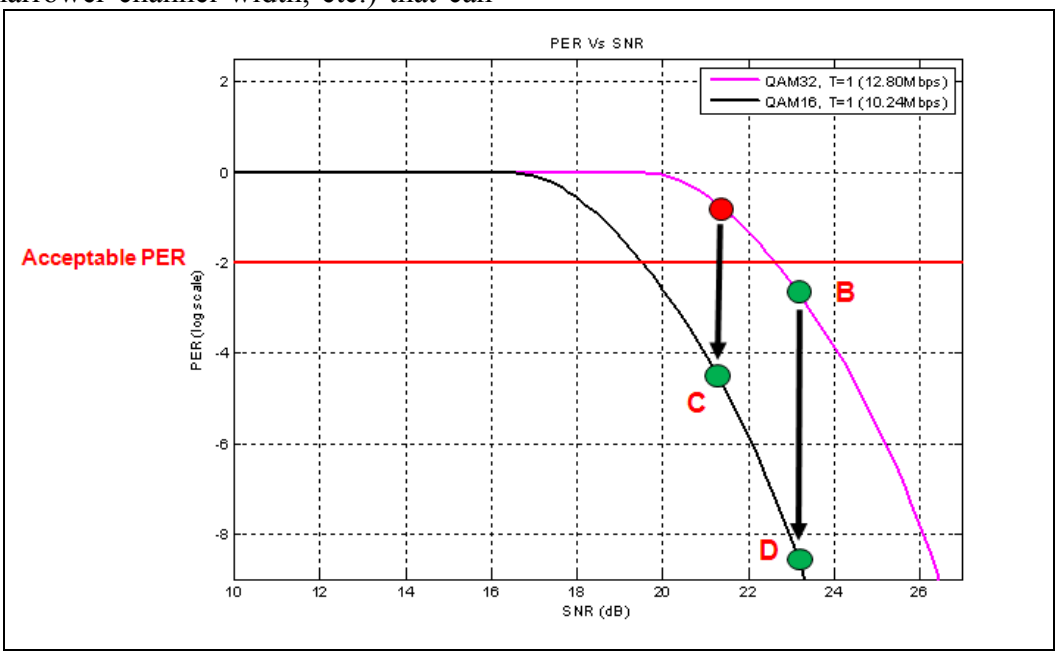


Figure 5. Creating a new modulation profile to accommodate the small SNR values of power-starved CMs is an existing solution. This results in less throughput for all CMs on that US channel. (Channel width is 3.2MHz in this example.)

results in an increased power per code as shown in Fig. 6, which essentially increases the SNR value for that power-starved CM and enables it to obtain good service without changing the modulation profile for the US channel. This solution, however, requires SCDMA-capable devices to be present at both the headend and subscriber's home. Therefore, this solution may not be optimal especially when either the CMTS or the power-starved CMs are SCDMA-incapable.

NOVEL SOLUTION

Power starvation of devices presents a challenging problem for the MSOs because it only affects a few CMs on the US channel. The desired solution needs to be optimal in the sense that it enables the proper operation of power-starved CMs while not degrading the service of non-power-starved CMs or affecting the overall efficiencies of the US channel spectrum.

In this section, we introduce a novel solution for the power starvation problem that does not suffer from any of the disadvantages of the above solutions. The solution is based on an intelligent algorithm that identifies low-SNR Power-Starved CMs and dynamically moves those CMs to channels with modulation profiles that can accommodate the limited SNR values of the power-starved CMs.

In particular, the power-starved devices are first identified using several metrics that can include: Transmit signal power level, receive signal power level, SNR, PER, Modulation Error Ratio (MER), channel noise, DS receive signal level, etc. Once the power-starved device is identified, the system scans through all US channels that are accessible by the power-starved device and identifies an US channel that can provide good performance at the low SNR value of the power-starved device. The power-starved device is then moved (via DOCSIS DCC commands) to the other US channel to obtain good service.

The above algorithm is illustrated in Fig. 7, where the identified power-starved CMs is moved to another US channel that requires smaller SNR values to achieve the same target PER value ($Y < X$). The target channel can be an US channel with smaller bandwidth (less noise in the passband), a channel with a lower order modulation profile, a channel with a modulation profile with higher Forward Error Correction (FEC) settings, or some/all of the above. Observe in Fig. 7 that the power-starved CM2 is moved from US1 (an ATDMA channel whose bandwidth is 6.4MHz and whose modulation profile is QAM64) to US2 (an ATDMA channel whose bandwidth is 6.4MHz and whose modulation profile is QAM32). Observe that moving CM2 to US2 results in an acceptable PER value (even though the modem is still power-starved, though!).

One important attribute of the proposed solution is the ability to identify power-starved devices and *dynamically* move them to US channels that are suitable for their low SNR values. The dynamic feature of this algorithm can be very beneficial especially when the SNR value of the power-starved device improves. This can happen in several ways, including: 1) when the US noise level decreases, or 2) when the subscriber fixes the problem inside the home by removing some of the previously installed RF splitters. Moving power-starved devices back to their original channels once their SNR values have increased helps to provide better subscriber service and easier optimal network management.

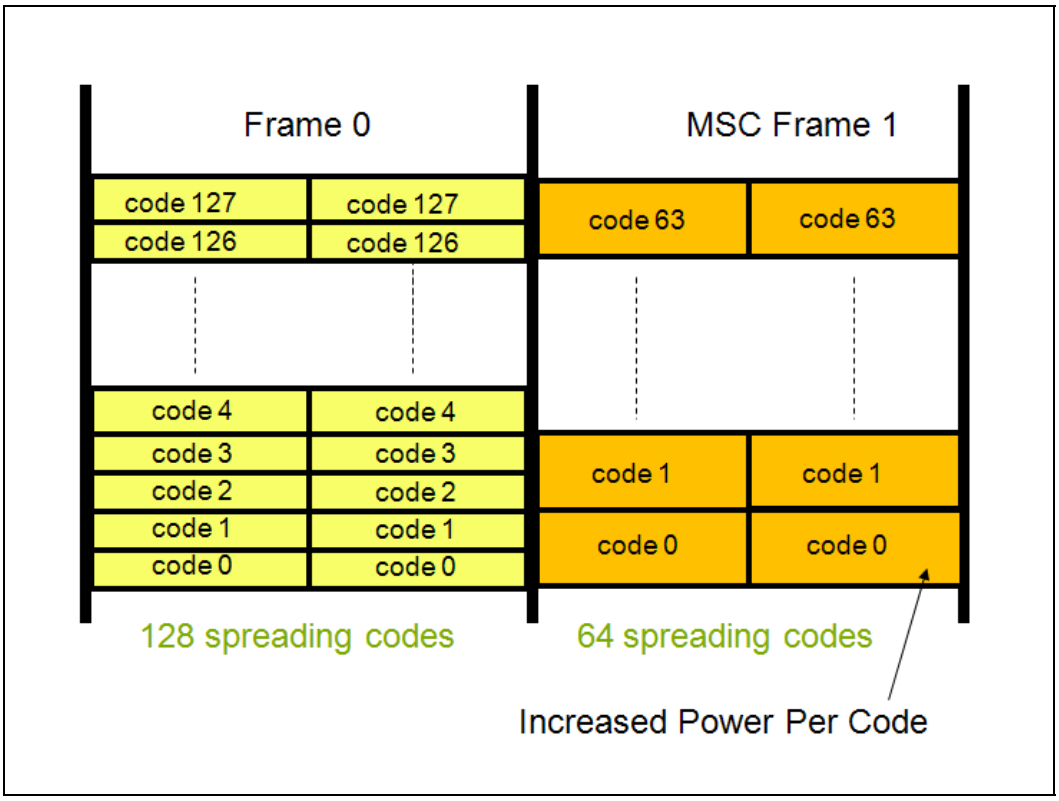


Figure 6. Increasing the code power in SCDMA through reducing the number of active codes is an existing solution for the power starvation problem.

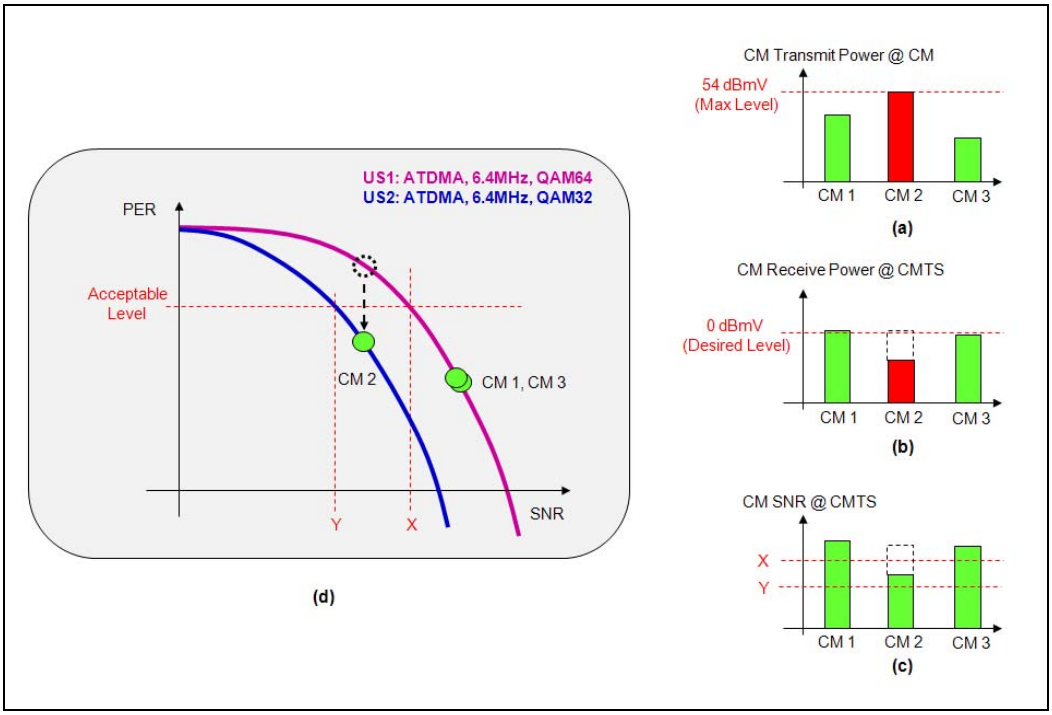


Figure 7. Proposed solution of Dynamic steering power-starved CMs. The power-starved modem, CM2, is moved from US1 (QAM64) to US2 (QAM32) which requires less SNR value to provide for the same target PER. In particular, note that US1 requires SNR=X to provide the desired PER value, while US2 requires SNR=Y (less than X) to provide the same desired PER value. CM2 is still power-starved but its low SNR value is properly accommodated by US2 and therefore no performance issues are encountered.

The network operator may wonder how to locate the additional bandwidth required for a second channel. Most MSOs do not want to consume a large amount of US bandwidth on their plant as they move into a future where the upstream bandwidth will become even more precious. Fortunately, the second US channel can be provided in several ways without wasting bandwidth or greatly affecting the spectrum efficiency. These techniques include:

1. Logical channels. A logical channel is an excellent mechanism to provide the second US channel because it is only used when the power-starved devices on that US channel needs to send data in the US direction. The bandwidth grants for logical channels are assigned dynamically.

2. Narrow channels with robust modulation profiles in the noisy band below 20MHz. This portion of the spectrum is lightly used and can be utilized for supporting the power-starved modems. The same principle also applies to "spectral holes" between other high-speed DOCSIS® channels.

3. Existing DOCSIS®1.0 TDMA channels. Some MSOs already have a TDMA channel present on their network to support legacy devices. This low throughput channel can also be utilized as a home for power-starved devices.

4. MSOs started to deploy DOCSIS® 3.0 US channel bonding which requires multiple US channels to be present. One of these US channels might be adequate to host power-starved devices.

Observe that once the low throughput US channel is identified and selected, it can be efficiently used to host all power-starved devices moved from different US channels within the MAC domain.

Dynamic steering of power-starved devices to other US channels (that are suitable for their low SNR values) is a general solution that has several advantages. These advantages include: 1) the ability to work with all DOCSIS® devices (DOCSIS®1.x, DOCSIS®2.0, DOCSIS®3.0), 2) the ability to work with TDMA/ATDMA/SCDMA channels, 3) the ability to improve the performance of power-starved modems without impacting the performance of non-power-starved modems, and 4) the ability to improve the performance of power-starved modems without impacting the efficiency of the spectrum.

CONCLUSIONS

The problem of power starvation within CMs, DSG-STBs, and MTAs was discussed in this paper. The article illustrates how this problem can occur whenever subscribers introduce many RF splitters into their homes. The existing solutions were listed along with their limitations. In general, the solutions were found to be either expensive, impractical, or suboptimal. A novel solution based on dynamic steering of the power-starved devices to other US channels that can accommodate their lower SNR values was proposed. The paper showed that the offered solution does not suffer from any of the disadvantages experienced by the existing solutions.

REFERENCES

[1] Cable Television Laboratories, Inc., *Data-Over-Cable Service Interface Specifications, DOCSIS® 2.0, Radio Frequency Interface Specification, CM-SP-RFIv2.0-C02-090422,* April 2009.

Authors' Contact Info:

ARRIS Group, Inc.
Address: 2400 Ogden Ave., Suite 180, Lisle,
IL 60532, USA
Tel: 630.281.3009
Fax: 630.281.3362
E-mail: Ayham.Al-Banna@arrisi.com
E-mail: Tom.Cloonan@arrisi.com

Biography:

Ayham Al-Banna, Ph.D.: Sr. Systems Architect at ARRIS Group, Inc., Chicago. His research interests include RF Communication, Traffic Management, QoE, and QoS. Ayham has published a book and numerous publications in the area of Wireless and Cable Communications.

Tom Cloonan, Ph.D.: Chief Strategy Officer at ARRIS Group, Inc, Chicago. His areas of research include QoS for DOCSIS® systems, Traffic Management, routing architectures, and TCP Performance. He holds more than 20 U.S. patents and has written more than 80 articles for technical publications and conferences.

EMPOWERING HD AND 3D VIDEO STREAMING

Benny Bing (benny@gatech.edu)
Georgia Institute of Technology

Abstract

This article presents methods to improve the efficiency and performance of streaming high-definition 2D and 3D compressed videos. Two key methods involve traffic shaping and buffer dimensioning. It will be shown that these methods reduce bit rate variability, which will in turn, minimize packet losses due to buffer overflows at the receiving device. We will also show that with efficient multiplexing and aggregation, further performance gains may be achieved.

1. INTRODUCTION

Streaming live and on-demand digital video content over the Internet, and in telecommunications and broadcast networks, is becoming prevalent. Streaming variable bit rate (VBR) video traffic is a special challenge due to the high dynamic range of the frame sizes that results in high bit rate variability. This is especially so for HD and 3D videos. As an example, Figure 1 shows the high and variable encoded rates for a 720p 3D video compressed using VC-1. Shaping the traffic to reduce the peak rates will minimize packet losses due to buffer overflow at the receiver.

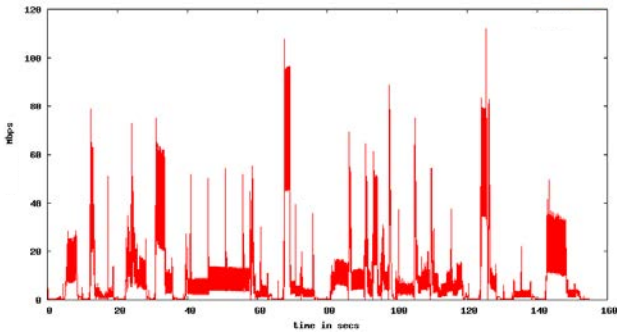


Figure 1: Variable bit rates for 3D 720p VC-1 video.

2. IMPACT OF PLAYER'S BUFFER SIZE

Figure 2 shows the impact of the player's buffer size on the TCP streaming throughput for a 720p H.264 video. A larger buffer of 8 Mbytes allows more information to be stored, hence the transmission can be completed at an earlier time than the case with a 1 Mbyte buffer. Note that the advertised receive window size in TCP may throttle the throughput to a smaller value when the buffer has received sufficient information. This is because the sender will take the minimum of the congestion window (which attempts to avoid congestion) and the receive window, when deciding on the appropriate window size to use. When some of the frames are played out, more buffer space is released, and the receive window will increase again. Figure 3 shows the impact of shaping the streaming throughput to a rate limit of 1.5 Mbps. As expected, the video file is received later than the unlimited case and the player with a larger buffer size achieves more efficient bandwidth utilization.

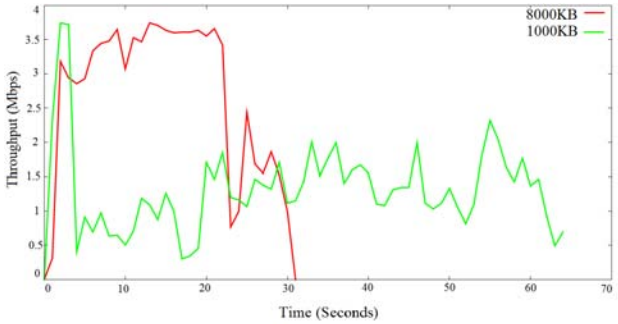


Figure 2: Impact of player buffer size on TCP streaming throughput.

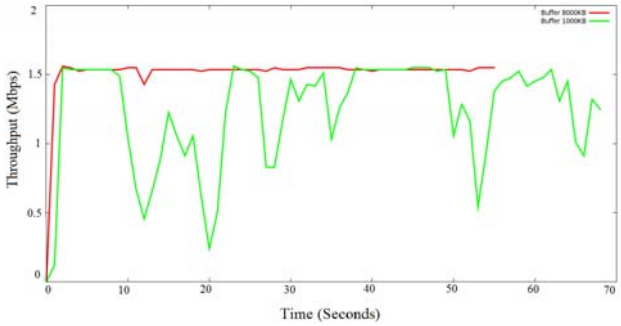


Figure 3: Impact of buffer space and shaping on TCP streaming throughput.

3. IMPACT OF TRANSPORT PROTOCOLS

The performance of traditional TCP and TCP with SACK is shown in Figures 4 and 5. The player's buffer size was set to 600 Kbyte. The theoretical average rate is computed by taking the video file size and dividing by the duration of the video. The initial rate is usually high because the player attempts to buffer as much data as possible for playback (including the first frame, a large I-frame), and thus advertises a large receive window. Subsequently, this advertised window gets throttled to a reasonable value. For the Avatar 720p 3D movie, the initial high rate can be attributed to the nature of the video content (high action at the start) and the player. From Table 1, it is clear that TCP with SACK reduces the overheads and hence, the average streaming rate. This is because TCP with SACK employs aggregated or block ACKs, which reduces the overheads of sending individual ACKs, thereby enabling multiple packet losses to be handled more efficiently.

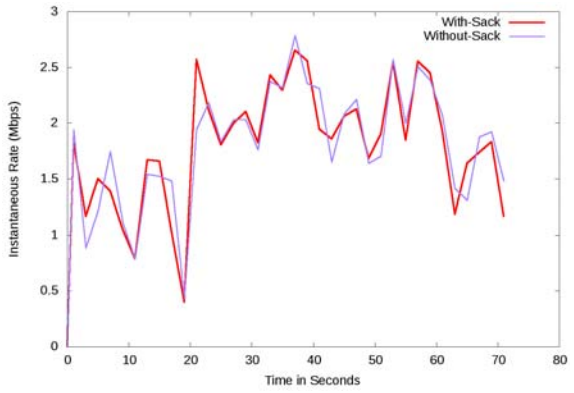


Figure 4: Streaming rates for a 1080p H.264 video.

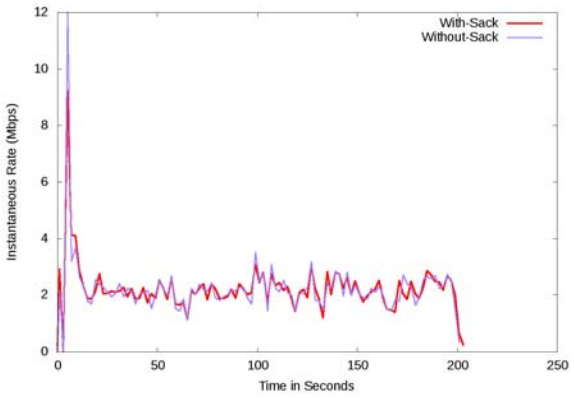


Figure 5: Streaming rates for a 720p H.264 3D video.

Table 1: Comparing TCP overhead (average rates).

	1080p 2D	720p 3D
TCP with SACK	2.011	2.318
TCP without SACK	2.014	2.326
Theoretical	1.836	2.111

Figures 6 and 7 show the streaming of the 720p 3D trailer using UDP. In general, UDP incurs less overheads than TCP since it is a unidirectional protocol. Unlike TCP, there is no congestion flow control in UDP, hence the raw UDP sending rates can be very high (a few orders of magnitude higher than TCP) and the video gets transferred in a very short duration. This may result in unacceptably high loss rates due to network congestion and receiver buffer overflow. These losses cannot be recovered in native UDP. A solution to manage this problem is to send the video at the frame rate that the video is meant to be played back (in this case, 30 Hz). The resulting sending rates are shown in Figure 6. Alternatively, progressive streaming can be employed to shape the rate variation (Figure 7). In both cases, the entire video gets transferred according to its duration (208s). For progressive streaming, the shaping threshold is set to 2.13 Mbps, thus proving that the UDP overheads are lower compared to TCP (Table 1).

3. PEAK TO AVERAGE RATE

The peak to average rate (PAR) normalizes the actual variation of the VBR video rates. This is computed by selecting the instantaneous rate of the compressed video and dividing by the average rate within a predefined interval. For example, if the video frame rate is 25 Hz, then the frame interval is 40 ms. The size of the frame divided by the frame interval gives the instantaneous rate. The predefined interval is chosen to be the duration of the entire video, and the instantaneous rate for each frame interval is averaged over this period, giving the long-term average PAR. Figures 8 and 9 illustrate the instantaneous rates and the long-term average PAR for a 720p H.264 movie trailer.

Note that the peak rate for the 720p VBR H.264 video is over 25 Mbps. Shaping the traffic to a lower rate results in additional buffering delay that has to be accommodated. As shown in Figure 10, choosing a low shaping threshold can result in significant delay. As we will see in the next section, a better way is to aggregate multiple video streams to make more efficient use of the channel bandwidth.

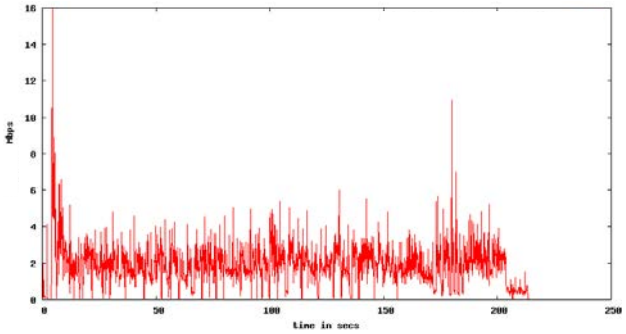


Figure 6: Instantaneous UDP rates for streaming the 720p H.264 3D video at 30 Hz.

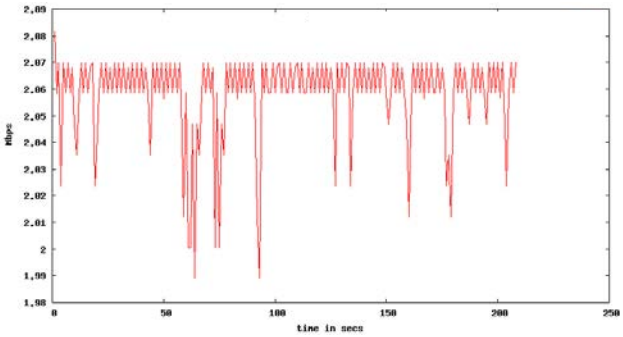


Figure 7: Shaped UDP rates for streaming the 720p H.264 3D video.

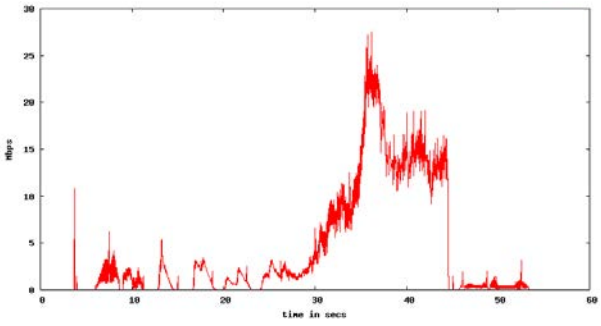


Figure 8: Instantaneous rates for a 720p H.264 video.

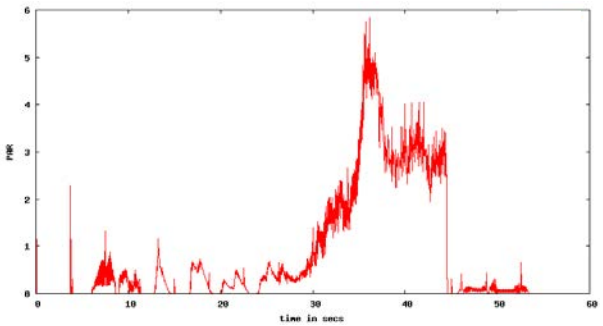


Figure 9: PAR for a 720p H.264 video.

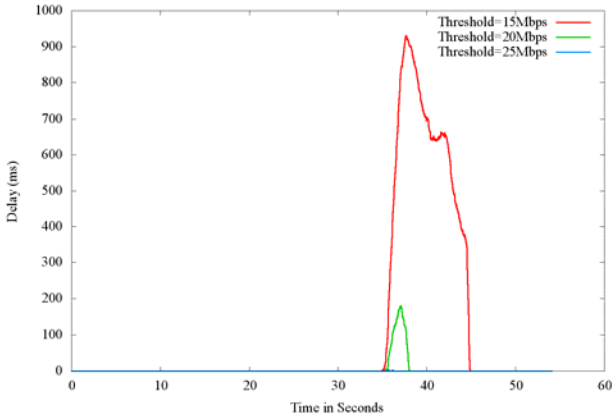


Figure 10: Shaping thresholds versus buffering delay.

The PAR variation is very high. A PAR of 1 is desirable because it achieves perfect utilization of the link bandwidth. A PAR value below 1 implies under-utilization whereas a PAR greater than 1 requires more buffering or bandwidth to accommodate the peak rates of the VBR video. By appropriately shaping the video traffic, a PAR close to 1 can be achieved. Figure 11 shows the streaming of another 720p H.264 video with a fixed shaping threshold. A PAR close to 1 is achieved. Since the video was streamed in real-time using TCP and the duration of the video was not known beforehand, a running average method was used to compute the PAR. The running average is computed by averaging the instantaneous rates since the beginning of the video play back.

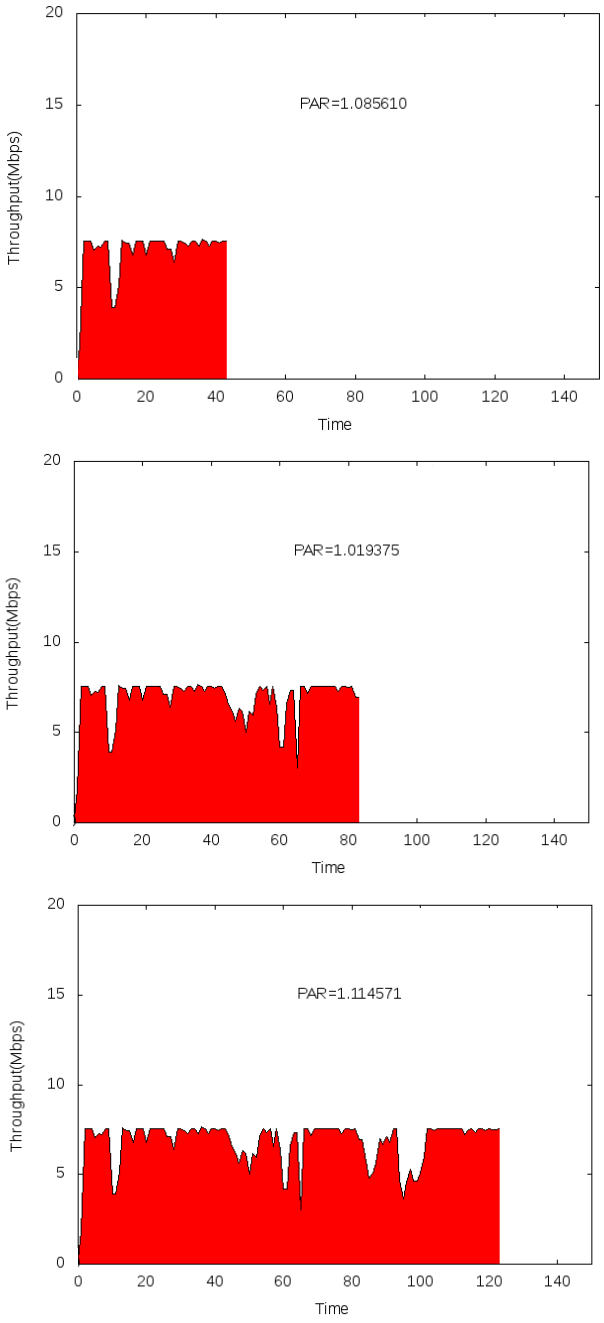


Figure 11: Evolution of PAR for live TCP streaming of a shaped 720p H.264 video.

4. MULTIPLEXING OF VIDEO STREAMS

It is common for a video headend or server to deal with multiple streams. When statistical multiplexing is applied to multiple VBR compressed videos at the headend or server, it can exploit the inherent variations in the instantaneous bit rates and increase the number of video streams within a fixed channel bandwidth while keeping the picture quality constant. For example, if one stream is demanding high bit rate, it is likely that other streams have capacity to spare. A large number of aggregated streams tends to "smooth" to a normal distribution (based on the central limit theorem). Unlike per stream buffer-based traffic shaping, statistical multiplexing introduces minimal delay.

The data rates for a MPEG stream can vary quite dramatically depending on the video content. As shown in Figure 12, the video resolution also plays an important part in the frame size distribution (and hence the data rates). The 480p Dell video contains very fast scene changes and the peak of the frame sizes occurs in the 170 Kbyte range. The 1440 × 1080 Terminator-2 trailer is fast action and the peak of the frame sizes occurs in the 340 Kbyte range. Compare with the slower motion 1080p FCL movie (only 5 scene changes) where the peak of the frame sizes occurs in the 480 Kbyte range. In addition, this video exhibits the broadest range of frame sizes. However, there is no strong correlation between the frame size distribution and the video duration: Dell (75.3 s), Terminator-2 (125 s), FCL (72.2 s).

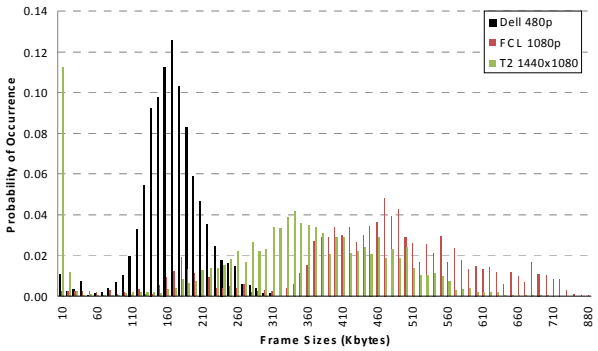


Figure 12: Frame size variability of H.264 videos.

Many compressed videos exhibit the long-range dependent (or long-tail) traffic characteristic. Because of this dependency, the video traffic tends towards clustering and becomes less predictable as the number of streams increases (which is in contrast to Poisson distributions that become smoother as

the aggregation volume increases). To illustrate this phenomenon, we multiplex 19 videos and 38 720p H.264 movie trailers with different content. As shown in Table 2 and in Figures 13 and 14, the standard deviation of the multiplexed rates is almost doubled when the number of multiplexed videos is increased two-fold, thereby proving the increased variability for a higher number of multiplexed streams. Thus, for multiplexed video streams, the buffers need to be larger to accommodate more extreme traffic-burst scenarios and traffic shaping may be needed. However, the standard deviation of the long-term average PAR reduces. This is because the aggregated average rate for 38 streams is larger than 19 streams, and this in turn, masks the effect of the overall variation to some extent. The average rate for the 19 and 38 streams remains about the same – 4.7 Mbps. Note that for 38 streams, a reasonable shaping threshold of say 250 Mbps results in minimal delay and yet, accommodates an average rate of 6.6 Mbps per stream, far lower than the 28 Mbps peak rate in Figure 8. Similarly, for 19 streams, a shaping threshold of 140 Mbps results in a delay comparable to 38 streams, giving an average rate of 7.4 Mbps.

Table 2: Standard deviation for instantaneous rates and PAR of multiplexed videos.

Number of Streams	Instantaneous Rates	Long-Term Average PAR
19 streams	28.797	0.3006
38 streams	50.391	0.2501

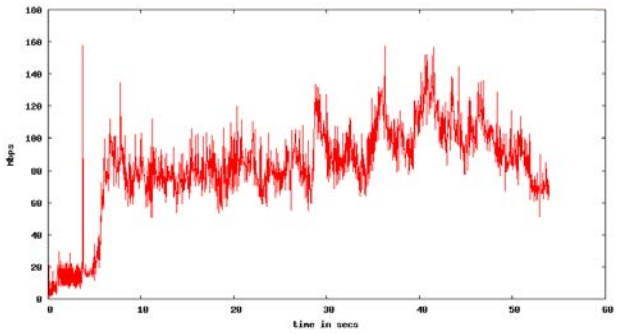


Figure 13: 19 multiplexed 720p H.264 videos.

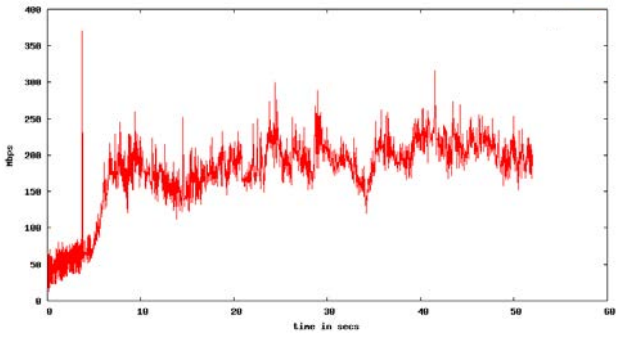


Figure 14: 38 multiplexed 720p H.264 videos.

To sum up, the bandwidth efficiencies that can be achieved using CBR and VBR 720p video multiplexing are shown in Figure 15. With channel bonding, higher efficiencies tend to be possible but not guaranteed. This is because with more streams, the standard deviation of the overall instantaneous rate becomes higher. To attain the same shaping delay as the lower number of streams, some bandwidth efficiency must be sacrificed. Alternatively, more aggressive shaping can be employed but this results in higher delays.

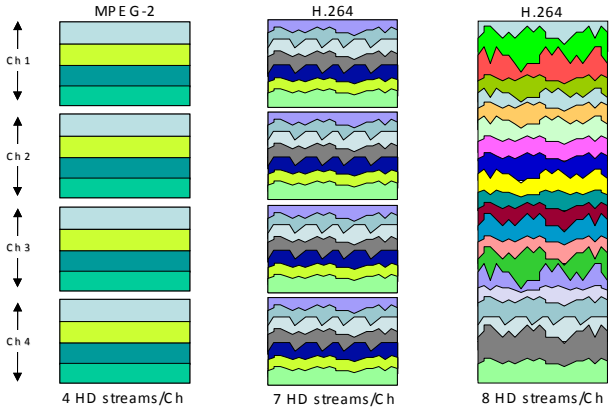


Figure 15: Efficiencies of 720p video multiplexing with and without channel bonding

5. VIDEO CONTAINER FORMAT

We now evaluate efficiencies of the MP4 video container format, which is widely used by online video portals. As can be seen from Table 3, the overhead for encapsulating a H.264 video in MP4 is insignificant, well below 0.01%, roughly 12 bytes per video frame. The MP4 container also incurs less overheads than the MPEG-2 transport stream

(TS) container, which has been wide used in many cable systems (Table 4). The difference in overheads grows as the video file size increases but the average percentage increase is in the region of 4%.

Table 3.6: Overheads for mp4 encapsulation (no audio).

Video	H.264 File Size (Mbyte)	MP4 File Size (Mbyte)	Overheads (Kbyte)
75s Dell, 480p	373.157	373.185	28
72s FCL, 720p	387.224	387.246	22
72s FCL, 1080p	772.062	772.085	23
596s BBB, 1080p	4,213.150	4,213.308	158

Video in 1280 × 544p	TS File Size (Mbyte)	MP4 File Size (Mbyte)	Difference (Mbyte)
54s The Dark Knight	33.395	32.094	1.301
64s The Hangover	45.736	44.055	1.681
73s Fred Claus	51.796	49.936	1.860
83s Night at the Museum	59.716	57.514	2.202
86s Speed Racer	62.907	60.595	2.312
106s 300 Video	54.579	52.479	2.100
128s Star Wars Clone Wars	94.247	90.910	3.337
137s The Astronaut Farmer	94.140	90.663	3.477
143s 10000 BC	98.045	94.540	3.505
141s Brothers Bloom	100.209	96.505	3.704
150s Transformers	110.419	106.415	4.004
191s Tetro	141.744	136.442	5.302

6. CONCLUSIONS

In this article, we have illustrated how HD and 3D VBR TCP/UDP video streaming can be enhanced via appropriate buffer size dimensioning, traffic shaping, and the use of statistical multiplexing to conserve bandwidth for aggregated video streams. We advocate the use of the MP4 container format for improved bandwidth efficiency. We have shown that the use of channel bonding may not improve the multiplexed bandwidth efficiency significantly, when compared to the single channel case. This is due to the long-range dependent characteristic of compressed VBR videos, which leads to an increased variability of the unshaped instantaneous rates when a higher number of streams are multiplexed. Our ongoing work focuses on deriving appropriate formulas for estimating the shaping thresholds and performing a more in-depth analysis of the long-term dependency of the HD and 3D VBR videos. Our longer-term research work focuses on enhancing the efficiency of the DOCSIS PHY layer via orthogonal frequency division multiplexing.

ACKNOWLEDGMENTS

The author is grateful to the support of Cox Communications.

REFERENCE

[1] B. Bing, *Broadband Video Networking – Empowering the HD and 3D Generation*, Artech House, Sept 2010.

BIOGRAPHY

Benny Bing is a research faculty member with the Georgia Institute of Technology since 2001. He has published over 80 technical papers and 11 books, and holds 5 pending patents in the areas of video, gesture, and bandwidth management technologies. His publications have appeared in the IEEE Spectrum and he has received 2 best paper awards. In early 2000, his book on wireless LANs was adopted by Cisco Systems to launch Cisco's first wireless product, the Aironet Wi-Fi product. He was subsequently invited by Qualcomm and the Office of Information Technology to conduct customized Wi-Fi courses. Other books were reviewed extensively by IEEE Communications Magazine (twice), IEEE Network as well as the ACM Networker. He is an editor for the IEEE Wireless Communications Magazine since 2003, where he also heads a special section on Industry Perspectives. He has guest edited for the IEEE Communications Magazine (2 issues) and the IEEE Journal on Selected Areas on Communications. All 5 of his online IEEE wireless tutorials have been sponsored by industry with one tutorial sponsored twice. In October 2003, he was invited by the National Science Foundation to participate in a workshop on Residential Broadband. He also led a team that received the National Association of Broadcasters (NAB) Technology Innovation Award in 2010. The award recognizes organizations that bring technology research exhibits and demonstrations of exceptional merit to the NAB Show. He is the founder of a video startup focused on enhancing and delivering next-generation video entertainment. He is a Senior Member of IEEE and an IEEE Communications Society Distinguished Lecturer.

IMPLEMENTATION OF STEREOSCOPIC 3D SYSTEMS ON CABLE

David K. Broberg, Mark Francisco
Cable Television Laboratories, Inc., Comcast Cable Communications, Inc

Abstract

Three-dimensional television is already upon is. Most of the major TV brands are introducing new models this year and some are already shipping. A tremendous effort has been made to pave the way for the introduction of stereoscopic 3D services over cable systems. This paper describes some of the implementation challenges that have already been overcome as well as some that remain.

INTRODUCTION

This paper divides the discussion of the impacts of deploying stereoscopic 3D (S3D) along the lines of the signal flow. Considerations for the formatting and ingest of S3D signals will be provided along with implications on encoders and headend processing. The importance of proper signal identification and solutions for downstream processing will be explained. Next the implications on the set-top box (STB) will be explored including impacts on firmware, on-screen-graphics and menus as well as closed caption decoding. Finally a summary of TV interface issues will be described.

STEREOSCOPIC VIDEO FORMATS

Frame-Compatible Formats

Stereoscopic video is captured from two identical cameras, one for each eye. To transmit the two pictures would typically require two identical signal paths, two wires, two separate video streams or two channels. Such a system is inefficient and many alternate approaches are available that exploit various aspects of the redundancy between the two images, reducing bandwidth demands by eliminating duplicate components while maintaining sync. Spatial multiplexing is one such mechanism that results in a frame-compatible signal.

A frame-compatible delivery of stereoscopic 3D video makes use of existing MPEG, AVC or VC1 coding standards [1]-[4]. It also uses various spatial filtering techniques to pack separate left and right images into a single frame or stream which can then be delivered through existing systems using only one channel or video stream. A frame-compatible approach delivers the stereoscopic content as if it were a regular 2D video stream and is otherwise compliant with 2D video standards.

Frame-compatible stereoscopic 3D delivery is technically possible using any type of multiplexing technique that is able to repackage the separate left-eye and right-eye images into the space and format normally used for 2D video transmission. Separate left and right images can be spatially reduced horizontally or vertically and squeezed to fit into a single video frame as a side-by-side image or as a top/bottom image, as illustrated below in figure 1 and figure 2.

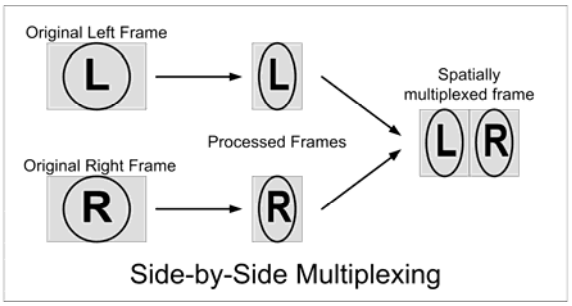


Fig. 1. Side-by-side frame packing is one example of a spatial multiplexing technique used to transport stereoscopic video using existing encoding systems.

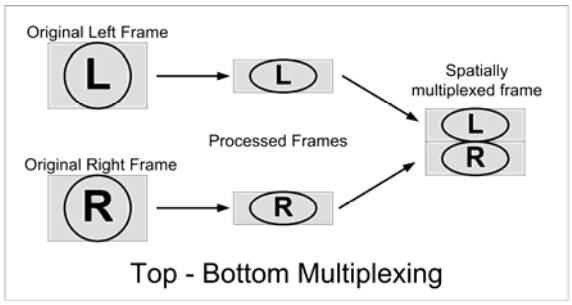


Fig. 2. Top and bottom frame packing is another example of a spatial multiplexing technique used to transport stereoscopic video using existing encoding systems.

The separate left and right images also can be spatially multiplexed on line-by-line, column-by-column or even in checkerboard patterns to be interleaved into a single frame. Different techniques and algorithms can be applied to the spatial reduction filtering with different levels of performance. Each of these methods has the consequence of reducing the spatial resolution overall. This loss of spatial resolution is exchanged for the inclusion of stereoscopic depth, which is conveyed as the horizontal disparity between the left and right images.

There are numerous variations within each of these format subgroups. Variations are created by unique subsampling or spatial reduction filtering. More variations are possible based on whether the left image is first or the right image is first, and whether the reduced images are flipped, mirrored or inverted. Illustrations in Figures 1 and 2 show just two examples of the many possible spatial multiplexing techniques.

In each of these spatially multiplexed variations a processor reduces the frame size of the original left and right video signals from the stereoscopic source so that both images may be packed or combined into a single video frame. The new spatially-multiplexed frame includes both the left and right views.

Separate left and right video sequences also can be temporally multiplexed into a common video stream by alternating frames or fields in a left-right-left-right frame-sequential pattern. Such a method would have the advantage of preserving the full spatial resolution at the expense of compromising the temporal resolution. While left and right stereoscopic signals are used as in the examples, frame-compatible techniques can be applied equally to 2D plus depth or 2D plus difference signals.

Any of these frame-compatible solutions can be compressed and encoded as if they were an ordinary 2D video frame. However, some of these systems are better able to survive encoding and decoding processes without the introduction of new errors or modifications needed to accommodate the new format.

Once these frame-compatible signals reach the 3D display, a stereoscopic processor must demultiplex the combined frame using an algorithm that complements or at least approximates the one used in the encoding process. This stereo-demultiplexing process restores the original left and right views.

Without any effort to narrow the number of usable choices, the market could see some programmers choosing one format while others chose a different format. Choices might even be made on a program-by-program basis finding reasons to favor the merits of one technique over another based on the nature of the content itself. Business arrangements might also influence these choices, as a number of the potential methods may need to be restricted to license agreements with IPR holders.

If so many variations of frame-compatible delivery coexist in the market the complexity of the stereoscopic demultiplexer increases dramatically with nearly unlimited multiplexing variations. To achieve a

successful 3D delivery system, the number of choices needs to be dramatically reduced while preserving the flexibility to work with a variety of existing equipment, content types and video formats. Current plans for cable have limited these choices to just three video formats and two frame-compatible systems:

1. 1280x720p60 – Top-and-Bottom formatting
2. 1920x1080p24 – Top-and-Bottom formatting
3. 1920x1080i60 – Side-by-Side formatting

Full-Frame Stereoscopic Formats

The delivery of two full-resolution frames requires other mechanisms to optimize the transmission and eliminate redundancy between the separate stereo source streams.

Using AVC multi-view coding standards optimized for the carriage of stereoscopic signals [5]-[9] would seem to be a logical choice for delivery over cable TV systems. However, delivery of stereoscopic 3D content using this approach comes with the cost and deployment delays associated with the introduction of new equipment and systems designed for these signals.

For cable operators the cost of replacement STBs and other headend equipment is a sizable consideration. If the demand for stereoscopic 3D content in the home develops rapidly, it is more likely that the cable operators could justify the cost of deploying new MVC solutions. However, without a low-cost interim method of delivery such as the frame-compatible approach that market may never develop.

Within the scope of the MVC coding standard there are also numerous possible options for the delivery of stereoscopic content. These various competing systems must be evaluated as part of the road-map to bring full resolution stereoscopic content for both eyes.

(1.) Discrete Left and Right Signals

The MVC coding system is designed to support a primary or base view along with a secondary (or non-base) view. The base view can be the left or the right view while the secondary view can be the opposite view for stereoscopic delivery. (The MVC extension to H.264/AVC also supports free-view and multi-view image coding with more than two secondary views, but for the purpose of this paper, the analysis is limited to the two-view stereoscopic coding.) The advantage of such a system is its simplicity. Existing AVC decoders that lack the ability to decode more than one stream simultaneously can receive the 2D compatible stream in the main channel (base view) and simply ignore or discard the secondary or alternate view [10]. New receivers would be designed with the ability to decode simultaneously two HD streams (MVC) so that they are able to decode both the base layer and the secondary layer, producing a stereoscopic output.

(2.) 2D plus delta or 2D plus difference

Another variation of MVC coding makes use of a pre-coding subtraction algorithm in order to reduce information in the secondary stream. The main signal is simply the left-eye view while the delta or difference signal is left-eye view minus the right-eye view (L + L-R) [11]. The presumed advantage of such a system is the reduced information content in the secondary stream.

(3.) 2D plus depth, 2D plus depth, occlusion and transparency (DOT).

This variation of MVC coding was primarily developed for the support of multi-view and free-view displays rather than stereoscopic displays. While separate left and right viewpoints can be derived from 2D plus depth and 2D plus DOT successfully, the processing requirements are much greater in the receiver and performance may be lower.

(4.) Frame-compatible with enhancement signals

It is also possible to encode frame-compatible formats using MVC. In this case the primary spatially-multiplexed frame is encoded as the base view while a complementary spatially-multiplexed frame or enhancement signal designed to restore the missing resolution is encoded as the secondary stream or view [12].

In this case a receiver needs to decode both the base view and the secondary stream in order to produce a full-resolution 2D or 3D view since the base view only includes ½ of the available resolution for the left-eye and right-eye views. However, such an approach can still provide a frame-compatible stereoscopic signal to legacy receivers unable to decode the secondary stream. Such a choice enables a more gradual migration if the cable operator uses frame-compatible delivery initially.

3D FORMAT SIGNALING

Signaling and detection of the specific 3D format transmitted to the receiving device are necessary to avoid manual configuration to view 3D. They are also necessary due to the potential for reconfiguration when content changes format between programs in the case of channel changes or within a program in the case of commercials. Providing identification of 3D format within the stream can aid in set-top graphics formatting, television configuration and allowing features such as EAS and closed-captioning to function properly while operating in stereo. MPEG4 Part 10 specifies Supplemental Enhancement Information (SEI) that includes multiplex descriptions for 3D content [13]. A method of extending the SEI to MPEG-2 is being proposed that carries the equivalent descriptors as user data. The descriptors are intended to convey the frame packing arrangement of the content, which is expected to be one of the following:

Side by Side (type 3)
Top Bottom (type 4)
Checkerboard (type 0)

Note checkerboard is not likely to be used due to issues with the 4:2:0 chroma subsampling[1] used in most MPEG profiles, but it is included to align with the HDMI supported formats. Each frame configuration can map to a supported HDMI signaling should the content pass through the STB unaltered.

STB CONSIDERATIONS

The addition of SEI or MPEG-2 signaling in 3D content may allow the STB to format graphics accordingly, scale video appropriately and signal the television operational mode without user interaction. This can enable automatic adaptation to stereoscopic content, and in-program switching if interstitial 2D content is present. Note the term STB (set-top box) is used to represent any device that receives MPEG-2 transport streams and provides a digital display interface compatible with HDMI.

3D Signal Detection & Mode Switching

An MPEG section filter is required to detect 3D signaling within the video bitstream. Once implemented, a program change due to channel change, program boundary or interstitial, may trigger a mode change within the STB. This is likely to occur without user interaction, but may involve on-screen or front-panel messaging. The resultant mode change may depend upon the type of television detected. If a 3D TV is detected that supports HDMI 1.4 or 1.4a, automatic switching will occur. If no 3D capabilities are detected, an on-screen message can be displayed to request the consumer to change the TV mode appropriately. Options to allow 2D viewing

[1] Since the color space is half the resolution of the luminance, quincunx, or checkerboard sampling results in 3D image degradation.

of the content can be presented dependent upon STB capability. Upon automatic or manual progression to 3D operation, the STB will format subsequent graphics appropriate for the 3D frame packing arrangement.

3D Graphics Rendering

In order for STB-generated graphics to appear properly when viewing 3D content, the graphics must be formatted with an equivalent frame multiplex. For example, if top bottom video is being received, the set-top generated graphics should be formatted top bottom prior to bit blending in the destination video buffer. Additionally, it may be desired to shift the images horizontally to have the graphics appear slightly in front of the video plane, providing a natural viewing experience. This may be particularly important when affecting captions, which tend to stay on-screen for longer viewing durations.

3D Video Scaling

Video scaling is sometimes used to allow the currently viewed program to remain on screen while viewing a program guide or other interactive features. Scaling 3D video involves a choice of associated television mode and maintenance of 3D imagery of the video. It may be desirable to revert the television from 3D viewing to allow a simple 2D projection of scaled video and guide. This may be required due to the complexity of the video scaling, such as multiple thumbnails projected, or due to limitations in the STB processing and memory. Alternatively, a single eye image of the video may be scaled linearly and copied to the appropriate frame multiplex format. A second alternative maintains a 3D projection of the scaled video, which requires the left and right components of the frame to be scaled linearly and copied to the appropriate locations for the frame multiplex. Video scaling options are illustrated in Figure 3.

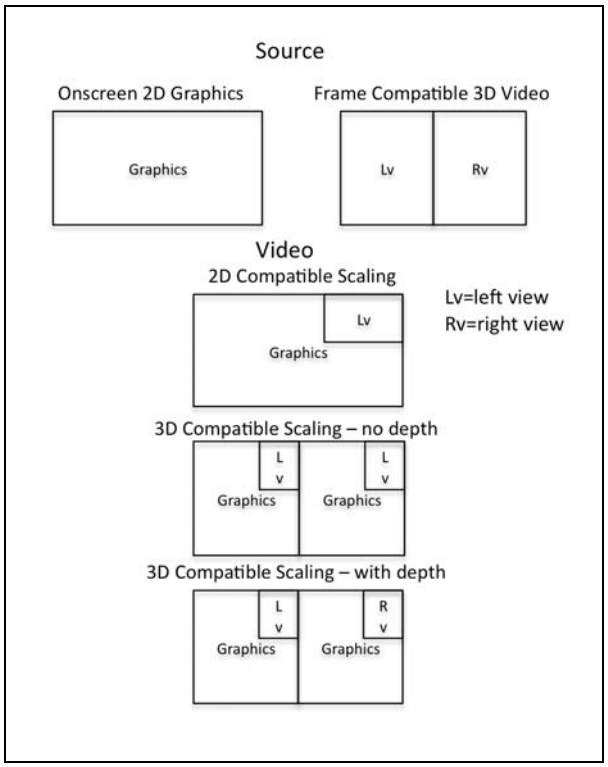


Fig. 3. Methods of combining graphics and video in a scaled window for 3D presentation

Television Signaling

Televisions supporting HDMI 1.4 and 1.4a 3D format signaling are currently available. Testing has shown the response to 3D signaling of currently available models to be without perceivable delay. This indicates the ability to support 3D mode switching at program start and stop, channel changes and within program boundaries. The ability to add interstitial 2D elements is possible without consumer perception of a program disruption. 3D eyewear may behave differently dependent upon implementation. It is generally a better experience to deliver a uniform program format to avoid the variation in light transmission that occurs when active 3D glasses temporarily stop shuttering. Formatting 2D advertising in a frame-compatible format may provide the best overall experience.

IPTV Considerations

While the methods described in this paper are primarily focused around delivery of 3D assets on QAM-based multi-program transport streams, application to IP and Internet delivery systems is readily achieved. IP encapsulation of MPEG-2 transport streams is typically used for network transport between source and edge QAMs, DOCSIS® delivery is readily achieved as delivery increasingly moves to IP oriented Content Delivery Network (CDN) architectures. Alternatively file-based delivery is possible although equivalents to the defined user data and SEI structures of MPEG are needed. One advantage of IP delivery is the ability to independently carry left and right eye images for adaptive edge multiplexing. Care must be taken to maintain synchronization of left, right and audio streams due to possibilities for propagation variation across IP networks.

Output Formatting and Rescaling

Another important consideration in the STB is the output formatting or rescaler. For 2DTV, the typical operation of the STB has been to adjust the various transmission video formats into a single video format preferred by the display through the use of a rescaler. For example HD video signals delivered as 1280x720p60 may be rescaled by the STB into a 1920x1080i60 format to drive the monitor. Conversely, depending upon the monitor, 1920x1080i60 HD content may be downscaled to 1280x720 and de-interlaced before being sent to the display.

For 3D signals such conversions can be far more detrimental to the 3D picture performance. Since the frame-compatible signals are already spatially reduced either horizontally or vertically, further spatial rescaling may add cumulative losses to the resolution, as well as destroy certain pixel relationships necessary to be decoded accurately. The interlace-progressive relationships must also be carefully preserved to assure the optimal 3D representation.

For these reasons it is essential that the STB operate in a video pass-through mode where the STB does not rescale when 3D video is being delivered. When a 3D signal is delivered as 1280x720p60 it must be output as 1280x720p60 by the STB and signals received as 1920x1080i60 must be output as 1920x1080i60 accordingly.

This video by-pass mode was not always possible with 2D televisions because many were not equipped to support a wide range of video input formats and scan rates. However this is not the case with the modern 3DTVs, and virtually all of the new 3DTVs are able to handle this wider range of video formats as an input signal.

3DTV INTERFACES

Legacy HDMI

Today's deployed HD STBs include the HDMI interface based upon the previous version 1.3 or older specifications. These specifications included no reference or provision for stereoscopic 3D video.

STBs that support this interface are still able to deliver stereoscopic 3D content, with certain limitations. First, the formats must be frame-compatible and use the same exact timing and signaling as any 2D video signal. Second, there is no direct provision for any automated detection or switching from 2D to 3D with such a system.

HDMI v1.4

The version 1.4 of the HDMI specification was release in June of 2009 [14] and added specific support for the carriage of stereoscopic 3D formats. It also added additional signaling to enable the discovery

and identification of stereoscopic capabilities as well as 2D and 3D signal identification.

Unfortunately, the HDMI v1.4 standard failed to mandate that 3DTVs support the frame-compatible formats necessary for cable delivery. This initial version also failed to identify or specify the needed Top-and-Bottom format. Without these necessary provision in the standard the market place uncertainty would make it very difficult to reliably deliver 3D video services to a wide range of products and models.

HDMI v1.4a

With the update to version 1.4a of the HDMI standard [15] the Top-and-Bottom format was added along with mandatory support for the three needed frame-compatible formats by 3D displays.

Another important change was also added at the same time as a change to the license restrictions. This change enabled the STBs that are limited to legacy HDMI v1.3 or older implementations to be able to selectively support the frame-compatible modes along with the format signaling and self-discovery features, without any obligation to support the higher bit rate full-resolution 3D formats mandated by version 1.4 [16].

This important change paved the way for firmware updates to be possible in existing STBs so that fully automated 3D support could be enabled.

Analog Component

Most new televisions, including the new 3DTV models, are still equipped with analog component interfaces. Many subscribers continue to use systems connected using these outdated analog connections. There is a risk that some who use these systems will upgrade to a new 3DTV and reconnect the existing analog component interfaces out of habit or to avoid the cost of upgrading to HDMI.

Most of the new 3DTVs don't offer true 3D viewing from the analog components. However, some include built-in 2D to 3D converters that can be operated on the analog component inputs. Since there is no bi-directional hand-shaking on this interface it is impossible for the STB to recognize the presence of the 3D capable TV. There is also no provision for 3D signal identification, so at best the experience would require a totally manual 2D/3D switching function.

3D-capable systems that are connected this way will only lead to disappointment and customer service issues and should be avoided.

RF or Direct QAM

Many of the new 3DTVs entering the market this year continue to include support for "clear-QAM". These sets don't include the CableCARD™ slot and are not fully qualified UDCPs. Nonetheless they are often able to decode SCTE-07 compliant QAM modulation when no conditional access encryption has been applied (clear-QAM). However there will be additional challenges using this approach for 3D delivery or reception beyond the usual problems, such as channel mapping and EAS support that plague 2D Clear-QAM sets.

Some of these sets can support the frame-compatible (broadcast) formats even from the tuner input, but are limited to MPEG2 video decoding and are not built compliant with SCTE standards for decoding AVC/H.264 or VC1. These sets will not be able to detect any supplemental data used to identify the 3D signals or format types, forcing them into a manual operational mode for 2D/3D switching at best. Some may actually stumble when the 3D signals are

received by failing to properly ignore the supplementary data signals.

To avoid disappointment and customer complaints the clear-QAM connections should be avoided for 3D services.

IR Signaling

Finally we can't overlook the need for infrared (IR) signaling between the 3DTV and the electronic shutter glasses. While this is not an interface that is provided by the cable operator or the STB, it does use a shared physical media with the IR-remote control.

Presently the market for 3DTVs predominantly uses IR signaling to activate and sync the 3D glasses. These systems are non-standardized and a variety of techniques, protocols and formats are used. Some use a subcarrier or pulse modulation, while others use base-band signaling. Some are broadband and others are narrowband filtered.

The risk of this chaotic range of non-standard implementations is to interference with the control of existing STBs. This interference can potentially be in either direction. For example, the STB remote could cause sync disruptions to the 3D eyewear or, more likely, the 3D sync signals from the TV could disrupt the IR remote operation of the STB.

Until standards are fully developed in this area, it is likely that some updates, patches or other field fixes may be necessary in the STBs, the 3DTVs, or both, to avoid these problems.

CONCLUSION

Cable can deliver 3DTV programming today in formats compatible with the latest generation of 3DTVs. This paper has described a variety of technical challenges that must be overcome to ensure success, maximum performance and easy operation. Proposed transmission standards are presented here that may provide a more seamless 2D to 3D transition for customers. Progress has been made and will continue so that the 3DTV experience of cable customers in the home can offer a rich new dimension in TV viewing never seen before.

REFERENCES

[1] ISO/IEC 13818-2, International standard (2000), MPEG-2 video.
[2] ANSI/SCTE 128 2008 AVC video systems and transport constraints for cable television.
[3] ANSI/SCTE 157 2008 VC-1 video systems and transport constraints for cable television.
[4] ANSI/SCTE 43 2005 Digital video systems characteristics standard for cable television" section 5.1.2, table 3, p3.
[5] ITU-T Rec. & ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services," March 2009.
[6] Yip, P.Y.; Malcolm, J.A.; Fernando, W.A.C.; Loo, K.K.; Arachchi, H.K., "Joint source and channel coding for H.264 compliant stereoscopic video transmission," Canadian Conference on Electrical and Computer Engineering, 2005., vol., no., pp.188-191, 1-4 May 2005.
[7] Martinian, E.; Behrens, A.; Jun Xin; Vetro, A.; Huifang Sun, "Extensions of H.264/AVC for multiview video compression," Image Processing, 2006 IEEE International Conference on, vol., no., pp.2981-2984, 8-11 Oct. 2006.
[8] Wenxian Yang; Ngan King Ngi, "MPEG-4 based stereoscopic video sequences encoder," Acoustics, Speech, and Signal Processing, 2004 Proceedings. (ICASSP '04). IEEE International Conference on , vol.3, no., pp. iii-741-4 vol.3, 17-21 May 2004.
[9] Yang, W.; Ngan, K.N.; Cai, J., "An MPEG-4-compatible stereoscopic/multiview video coding scheme," Circuits and Systems for Video Technology, IEEE Transactions on , vol.16, no.2, pp. 286-290, Feb. 2006.
[10] Ying Chen, Ye-Kui Wang, Kemal Ugur, Miska M. Hannuksela, Jani Lainema, and Moncef Gabbouj, “The emerging MVC standard for 3D video services,” EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 786015, 13 pages, 2009.
[11] Ethan Schur, "Digital Stereoscopic Convergence Where Video Games and Movies for the Home User Meet," HDdailies.com, Mar. 2009.
[12] Walt Husak, "Dolby 3D White Paper", unpublished.

[13] Amendment to ITU-T Rec. H.264 | ISO/IEC 14496-10: "Constrained baseline profile and supplemental enhancement information," in-progress.
[14] HDMI Licensing, LLC: "High definition multimedia interface specification version 1.4", June 5, 2009.
[15] HDMI Licensing, LLC: "High-Definition Multimedia Interface Specification Version 1.4a Extraction of 3D Signaling Portion", Mar 4, 2010.
[16] HDMI Licensing, LLC "HDMI Licensing, LLC communicates further detail on 3D requirements within the HDMI specification version 1.4" Sunnyvale, California, December 23, 2009.

ABOUT THE AUTHORS

David Broberg
Vice President, Consumer Video Technology
CableLabs

Mr. Broberg has been with CableLabs® since 1999 and was a principal developer of the OpenCable™ hardware architecture. As a member of the System Architecture & Design Group, he is currently responsible for identifying, analyzing and developing key strategic video technology and related specifications and standards. Mr. Broberg has more than 30 years of technical and business experience in the television industry including broadcast, cable, satellite and consumer electronics and holds four awarded patents in the field. Prior to joining CableLabs, Mr. Broberg served in several key strategic engineering management and product development roles for Mitsubishi Electric including digital broadcasting and consumer electronics.

In 2009, Mr. Broberg received the Excellence in Standards Award from the Society of Cable Telecommunications Engineers. Mr. Broberg is a member of the SMPTE Task Force on 3D, the CEA 3D Task Force and chairs the SCTE Digital Video Subcommittee 3D Ad Hoc Group. Mr. Broberg is a Sr. Member of IEEE and a veteran of the US Air Force where he began his career in video technology and obtained an Associate in Applied Arts & Sciences, Electronic Technology from Mountain View College in 1980. Mr. Broberg is also a member of the International Stereoscopic Union and enjoys stereoscopic photography and video as a hobby.

Mark Francisco
Fellow, Office of the CTO
Comcast

Mr. Francisco has been with Comcast since 2001 and has been developing premises devices and services for Comcast's video, Internet and telephone customers. As a member of the Office of the CTO, he is currently responsible for video product architecture supporting broadcast, switched broadcast, on demand and internet delivery networks. Mr. Francisco has over 25 years of experience in design and systems integration with RF communications systems and holds three patents in the field, with four applications pending. Prior to joining Comcast, Mr. Francisco led systems integration and test activities for Motorola's New Jersey digital mobile phone design center and led the development of telemetry receivers and transmitters for Lockheed Martin Astro-Space Division. Mr. Francisco received a Masters of Science Degree in Electrical Engineering from Drexel University in Philadelphia, PA, and a Bachelors of Science Degree in Electrical Engineering from Rutgers College of Engineering in Piscataway, NJ.

METADATA DRIVEN GRID ENCODING

Dan Holden
Comcast Media Center

Abstract

This paper addresses a new transport methodology that will reduce bandwidth consumption on a cable plant. Cable faces a unique challenge — how to support heritage set top boxes (STB) concurrently with new Consumer Electronic (CE) devices that have exponentially greater capabilities which include: increased resolutions, frame rates, 3D, interactive applications, two-way/IP protocols, and the ability to reach beyond a traditional cable plant. The cable industry must find a way to maximize the video viewing experience, while minimizing the bandwidth and storage requirements associated with compressed video streams. The complexities associated with advanced services are daunting. Numerous standards try to address unique opportunities, but the industry currently lacks a unified methodology to associate all of these disparaging technologies into a single, integrated delivery mechanism. An approach to resolve this issue is dividing video, audio, and associated data into a grid. Rows are responsible for horizontal partitioning of enhanced video resolution, 3D, applications, data, subtitling, and audio tracks. Vertical partitioning through columns will allow for packetization along Group of Picture (GOP) boundaries, which will service IP streaming, ad insertion, start over, and other container-based services.

Current techniques for addressing the increased capabilities of STBs include simulcast for linear content and multiple file-based encodes for VOD assets. For 3D content, side-by-side and over/under frame compatible formats are being adopted, which will substantially increase storage requirement in the VOD plant. In addition, these 3D formats will require simulcast techniques when new formats such as 1080p60 are introduced into linear broadcast. Horizontal segmentation of compressed video, in conjunction with an enhanced transport container, will provide a new method to deliver 3D in a scalable fashion. A single video package will be able to contain all information required to produce multiple formats, e.g. 720p, 1080p24, 1080p30, 1080p60, 3D 1080p24, 3D 1080p30, and 3D 1080p60.

This delivery mechanism will allow for scalable video delivery that can expand with CE device capability. Allowing CE devices to make two-way requests of specific containers utilizing metadata will reduce storage costs and bandwidth consumption on the local loop. This generalized approach will significantly reduce VOD storage requirements. For linear broadcast, PID filtering and video processing at the STB will eliminate the need for simulcast.

INTRODUCTION

All Multiple Systems Operators (MSO) face similar challenges of how to implement advanced and re-usable business intelligence on heritage Consumer Premise Equipment (CPE). In the past three years, it has not been feasible for MSOs to move from MPEG-2 to H.264. EBIF (Enhanced TV Binary Interchange Format) was selected over tru2way for initial deployment at Comcast for one simple reason – there are tens of millions of deployed DCT2000s in the home. Implementation of 1080p will likely happen on MPEG-2 before H.264. The first release of 3D video will be using

MPEG-2. In order to stay competitive a paradigm shift in video encoding must take place. Video compression must not only save storage and transport bits, but there must also be a clear roadmap between existing deployed technology and the next generation of video display devices. Innovation has consistently been the key to cable success. The foundation of maintainable technology is based upon a clear upgradeable and supportable roadmap. Incompatibilities between core infrastructure and edge devices prevent the deployment of new technologies when they are available. By introducing an abstraction layer and proven Information Technology (IT) techniques, it will be possible to offer features in the home without the need to replace legacy devices. The future of encoding needs to target a system that allows for loose coupling of the encoder and decoder, and ensures all audio, video, and data is tightly bound. The proposed grid approach will allow compression experts to continue to do what they do best; save bits on the plant. This will allow other teams of experts to extend the functionality of CE devices in the home.

The cost to store and processes bits at the edge of a Hybrid fiber-coaxial HFC network is exponentially more expensive than in a super headend. Given this equation, STB tend to be basic devices with limited storage and processing power. Truck rolls for edge-device replacements are normally the option of last resort for MSOs because of the high cost and the number of affected households, which can be in the millions. For this simple reason, heritage devices tend to live on the network beyond their engineered life span. As a result, MSOs tend to deploy technology that meets the requirements of the simplest device on the network (currently a DCT2000). In order to stay competitive, a methodology must be adopted that allows new, innovative technology to exist concurrently on a shared network with heritage CE devices in the home. Like your older brother's hand-me-down clothes, old TVs and other CE devices never make it to the dumpster, they migrate from the living room to the bedroom, which has a dramatic effect on the network. The extensible nature of grid encoding will allow this bedroom TV to continue to generate revenue far beyond its life expectancy, without impacting innovation.

Grid encoding is a new science that will be deployed in the super headend in order to reduce the number of bits transported and stored at the edge of the network. It will reside between the encoders and decoders, and should not require significant changes to current encoding specifications or technologies. Video pumps and CE devices will need to be built in a fashion that will leverage the two-way infrastructure of the modern HFC plant. This cutting edge technology has the ability to extend video, audio, and interactive TV specifications; and most importantly, it will extend the life of CE devices in the home; therefore, reducing the operational cost and maintenance.

ENCODING

Blob Encoding

Current encoding technology relies on encoding assets to a single "blob." A blob encompasses video, audio, and other Packet IDs (PID) such as interactive TV applications that have been compressed and identified by metadata. If we examine a typical asset encoded using this methodology it might contain the following representative PIDs.

Table 1

PID Type	Values
Video	720p60, 1080p24, 1080p30, 1080p60, 1080i, Flash, WM9
Audio	AC3, AAC
Data	EBIF, tru2way, Packet Cable

Using the sample data in Table 1, it is possible to calculate the number of blob assets that would be generated by combining the different PIDs. A simple combination, without repetition can be expressed as:

number of blobs = n! / [(n-r)! r!] = 13!/[(13-3)!*3!] =
(13 choose 3)
= 286 Blobs

If we were to take these 13 unique PIDs three at a time, we would derive 286 separate assets. Each asset would effectively be the same movie just implementing different compression or interactive TV technologies. Our single movie when encoded using blob technology would require unique encodings or packaging, in order to support seven different video 'coder-decoder' CODECs, two audio CODECs and three interactive TV data PIDs. Using this primitive encoding technique for 100,000 movies would result in 286,000,000 assets that would need to be managed, distributed, and streamed individually.

GRID ENCODING

Unlike current blob encoding technologies which do not leverage the capabilities of a two-way network infrastructure, grid encoding seeks to move the complexity of video transport off the CE device to the network. Old encoding technologies singularly focus on reducing the number of bits required for transport and storage. Now the focus shifts to solving issues surrounding migration paths from heritage technologies (MPEG-2, 1080i, 2DTV video) to 'better' technologies (H.264, 1080p, 3DTV). Significant improvements can be achieved by leveraging typical TCP/IP communications and information technology (IT) topologies. When a CPE device is granted the ability to announce itself on the network and broadcast its capabilities, it is possible to generate significant changes to the encoding specifications. It will effectively be able to tell a VOD system the decoder capabilities, and the VOD system will be able to respond to a request with a tailor made video encoding package.

Encode

Encoding is the process of creating a self-synchronizing stream of signals against a known timeline. For example, we will assume encoding with MPEG-2. The process will work equally well for VC-1, H.264, Flash, or any other CODECs that create horizontal separation of the data. In order to keep the example simple we will use "Theatrical Release" as the title of a 3D movie . For audio we will choose AC3 in order to ensure the encoded asset is compatible with currently deployed STB. A representative output stream from an encoder is depicted in Figure 3. This elementary stream will now be passed to the transcoding step of our process.

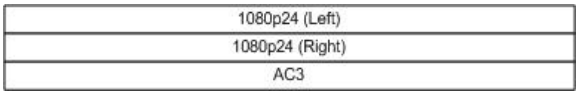


Figure 1

Transcode

During the transcoding process multiple CODECs, resolutions, and/or bit rates will be generated. This process will enhance the movie so that it can play on multiple CE devices at multiple bit rates. Support for multiple audio CODECs will also be added in the transcoding process. This process will prepare the asset for fragmentation at various bit rates.

iTV Striping

For this example the enhancement layers will carry EBIF and tru2way applications. These applications are carried in Package Identifiers (PID) which is added through a process called iTV striping. These data PIDs are bound to the video and provide new functionality on the CPE device. They add interactive features to the video stream such as voting-and-polling applications, advanced advertising, and other enhanced features. A representative output is depicted in Figure 2.

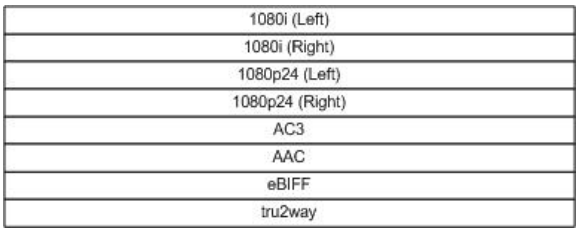


Figure 2

Fragmenting

Fragmentation is the process used to separate the transport stream vertically in order to prepare the video for loading into the grid. Packetizing the video stream into PIDs is an effective means for separating the audio, video, and data into fixed or variable, size units. This process of placing the packets into fixed or variable duration units will be utilized for adaptive streaming. The number of frames loaded into each cell of the grid does not have to be consistent, as the timeline will be maintained in the client buffer. Breaking the video along the GOP Boundaries will help facilitate ad insertion. Each fragment will require a unique identifier to facilitate the loading of the grid and orderly retrieval of video, audio, and data for placement into the decode buffer. Multiple bit rates for the video PIDs will be provided in order to support adaptive streaming technology. This functionality is not expressed in Figure 3 for the purpose of simplicity.

Fragment 1	Fragment 2	Fragment 3	Fragment 4	Fragment n
		1080i (Left)		
		1080i (Right)		
		1080p24 (Left)		
		1080p24 (Right)		
		AC3		
		AAC		
		eBIFF		
		tru2way		

Figure 3

GRID

Let's start by examining a single fragment that has been produced by the fragmentation process (Figure 6.) It may contain multiple video CODECs. In reference to the "Theatrical Release" movie example, it contains two fragments with the same sequence of video frames. The left eye 1080i and the left eye 1080p24 cells are by definition not equivalent, even if they contain the same footage. They may contain the same number of frames and their time code must be synchronized with the audio cells and iTV cells. The grid represents a new type of structured video. Rather than thinking of the grid as fragmented video, it will be treated as a simple data structure that has been partitioned and described with metadata for optimized storing and retrieval of video objects. Each video cell in the grid is effectively equivalent. The timeline will be maintained in the buffer. It is natural to think of this type of data structure as an enhanced database that has been optimized for storing and retrieval of video, audio, and other objects of interest to a typical CE device.

Load the Grid

Examination of a single fragment (Figure 4) reveals multiple cells with video, audio and data. These cells can be loaded into the grid one column at a time. The first

column of the grid represents time 0 on the timeline or the beginning of our sample movie title. Each cell of the grid is loaded until the last column is complete.

On video ingest, audio, and other components of the video stream are parsed and loaded into appropriate cells. These cells are then accessed by the VOD pump using Structured Video Query Language (SvQL) at the request of a CPE device.

Query the Grid

After the data has been placed into the grid, it is possible to extract the data using SvQL. Below is an example that could be used to represent a 3D movie request from an STB:

SELECT * FROM movies WHERE movie_title='Theatrical Release' and left_eye_video=1080p24 (Left) and right_eye_video=1080p24 (Right) AND audio=AC3 and iTV=EBIF

In addition, this sample query could be used to retrieve a 2D movie on a Personal Computer without interactive features:

SELECT * FROM movies WHERE movie_title='Theatrical Release' and left_eye_video=1080p24 (Left) and right_eye_video=NULL AND audio=AAC

Fragment n
1080i (Left)
1080i (Right)
1080p24 (Left)
1080p24 (Right)
AC3
AAC
eBIFF
tru2way

Figure 4

Extend the Grid

Extending the grid is a straightforward process that can be achieved by simply adding additional rows to video before fragmentation. These rows will represent new columns in the grid. As a sample, 1080p video at 60 frames per second will be added. Additionally, IP Multimedia Subsystem (IMS) and MP3 audio will be added to the CPE device. The features will be added to the encoding, transcoding, and fragmentation farms. A representative fragment is shown in Figure 5.

Fragment n
1080i (Left)
1080i (Right)
1080p24 (Left)
1080p24 (Right)
1080p60 (Left)
1080p60 (Right)
AC3
AAC
MP3
eBIFF
tru2way
IMS

Figure 5

Extending the fragment has no effect on retrieving the data. The SvQL statements on legacy devices will not query the new, extended rows of the grid. It is possible to update the SvQL on legacy devices in order to extend their respective life and functionality. This type of extension will ensure CPE devices will expose the greatest amount of functionality and therefore will not lock the MSO into legacy technology.

Video On Demand (VOD)

Grid encoding for VOD leverages the inherent capabilities of a two-way infrastructure. Beginning at the left of Figure 6, the video is transformed, segmented, and loaded into the grid. From the right of the figure, the CPE device identifies itself and

capabilities on the network, and requests only the cells of data that it knows how to process. The end-to-end VOD workflow loads the grid and exposes the data to the home.

VOD systems require significant innovation, and changes would be required to the CPE devices in order to take advantage of the new technology.

Late binding or polymorphism allows our CE devices to attach or request a stream that matches the exact features and hardware of the device.

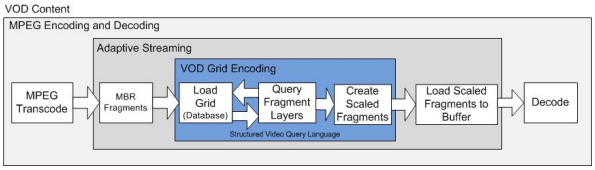


Figure 6

Linear

Linear content is not loaded into a static grid for a future query; rather the data is processed real-time and loaded into a "topic grid," which does not send video to a specific receiver. The published video is characterized into specific types of video without knowledge of which device will consume the video. This queue will live on the CPE device or another network location that is localized to the home. Subscribers then expresses interest in specific types of video (1080p24, AC3, EBIF) and receives only the video of interest, without knowledge of what, if any, video publishers that exist. This technique allows the live, streamed data to be loosely coupled and tightly bound to our encoding technology. All data retrieved from the grid is re-assembled as fragments and loaded into the buffer of the CPE device for decoding. Figure 7 represents end-to-end flow for linear content.

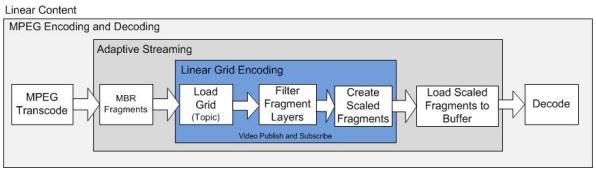


Figure 7

METADATA

Metadata is the key component to drive the entire encoding and decoding process. It is the glue that brings the architecture together. Original content is described by metadata. It is used to drive the orchestrations to encode, transcode, fragment and load the grid. All business logic is encompassed in metadata such as XML. The CE device utilizes metadata request relevant cells and the entire grid may be persisted on a device such as a DVR for future playback. Typical Electronic Program Guide (EPG) is another form of metadata that can be used to help parse the video and load it into the grid. PIDs use metadata to characterize their value to the ecosystem. Metadata is used to describe the grid and is integral in retrieving the correct components in the grid. The grid is self-describing metadata comprised of rows and columns similar to a database.

While SvQL may be used for VOD applications, linear content is in flight. A message bus should be utilized due to the data latency associated with database applications. In order to subscribe to the message bus, the following type of Java Message System (JMS) XML metadata could be utilized:

Sample Message Bean XML:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<tv-ejb-jar
xmlns="http://www.objectweb.org/tv/ns"

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
xsi:schemaLocation="http://www.objectweb.
org/tv/ns
    http://www.objectweb.org/tv/ns/tv-ejb-
jar_4_0.xsd" >
 <tv-entity>
  <ejb-name>VersusChannel</ejb-name>
  <jndi-name>VersusChannelHome</jndi-
name>
  <jndi-local-
name>ExampleTwoLocalHome</jndi-local-
name>
  <jdbc-mapping>
   <jndi-name>jdbc_1</jndi-name>
   <jdbc-table-name>MoviesTable</jdbc-
table-name>
   <cmp-field-jdbc-mapping>
    <field-name>MovieTitle</field-name>
    <jdbc-field-name>dbMovieTitle</jdbc-
field-name>
   </cmp-field-jdbc-mapping>
   <cmp-field-jdbc-mapping>
    <field-name>VideoCODEC</field-
name>
    <jdbc-field-
name>dbVideoCODEC</jdbc-field-name>
   </cmp-field-jdbc-mapping>
   <cmp-field-jdbc-mapping>
    <field-name>AudioCODEC</field-
name>
    <jdbc-field-
name>dbAudioCODEC</jdbc-field-name>
   </cmp-field-jdbc-mapping>
   <finder-method-jdbc-mapping>
    <tv-method>
     <method-
name>findByMovieTitle</method-name>
    </tv-method>
    <jdbc-where-clause>where
dbMovieTitle = 'Theatrical Release'</jdbc-
where-clause>
   </finder-method-jdbc-mapping>
  </jdbc-mapping>
 </tv-entity>
</tv-ejb-jar>
```

COMPARISON TO SCALEABLE VIDEO ENCODING (SVC) AND MULTIVIEW CODING (MVC)

SVC suffers from one inherent flaw: it adds significant complexity to the edge of the network, the very edge that is the most complex to maintain, upgrade, and support. The cost of a single byte of memory in a super headend is minimal. As this byte is propagated into the network the cost soars exponentially. For example, if the cost was $20 per GB, the memory would cost $20 if deployed in the super headend, or nearly $500MM if deployed at the edge. As indicated in table 2, the aggregate cost for installing a GB of memory in every set top box in a major market cable system could exceed $490 million, assuming a cost of $240 per set top box in a market supporting more than 200,000 set top boxes.

Table 2

Location	Cost for 1 GB RAM
Super Head	$20
Application Point of Presence (APOP)	$240
Headend	$3,000
Set Top Box	$ 490,000,000

The same calculations can be made for CPU, software support, and other technologies deployed in the STB. Moving complexity upstream saves capital expenditure and future support expenses. SVC video topology leads to significant cost increases for the operator. However, it does offer several advantages, for instance; SVC is a more effective approach than simulcast for reducing the bandwidth associated with the delivery of multiple CODECs. From a VOD perspective, it may be advantageous to use SVC to the video pump then to convert the stream at the pump to the specific CODEC requested by the CE device, which would be close to the grid encoding technique described earlier.

Multi-view coding (MVC) is an attempt to address encoding requirements for 3D and multiple camera views. Again, this technology is targeted at linear broadcasts and has limited application in an On Demand environment. While the cost of a general CE device may be relatively cheap, the cost to support and deploy the devices can be astronomical. MVC as a standalone technology may bring value to 3D content; it is certainly a better compression technology than the current side-by-side frame compatible approach that will be deployed in first generation 3DTV broadcasts.

Grid encoding can be used to enhance various types of encoding, including SVC and MVC. It may also be used as a replacement for both SVC and MVC encoding standards. SVC will utilize fewer bits to represent the video, but it does not currently have the ability to support adaptive steaming technology. SVC binds the encoder directly to the decoder which is a disadvantage. It will add complexity to the decoding process at the CPE device than grid encoding, which will drive up the cost of the CE device.

Grid encoding moves the complexity of decoding from CPE to the network, which should provide better architecture for scaling. Grid encoding is better suited for streaming On Demand assets as each CPE device has the ability to request only the information that is relevant to the device. VOD pumps are in a better position to handle the complexity of multiple CODECs than the STB. Pump upgrades are possible without intrusion into customer's homes. Grid encoding will require more bits on plant than the SVC. Note: the bits will be located on the "cheap" and easy-to-maintain distribution network, rather than the expensive and inaccessible network in the customer's home.

CONCLUSION

Grid encoding addresses a key component that is lacking in our current video compression architecture: how to make our services extensible. Within a few short years 1080p60 video will be available on STBs. Whatever technology is deployed today will ultimately have a very short life span. Ensuring a clear path to the future is a core component of any valid architecture that an MSO should consider deploying.

The Motion Picture Experts Group is highly regarded for "doing compression the best.," as evidenced by industry-wide adoption of such standards as MPEG-2, MPEG-4 and H 264. However, the group's' work on H.264 did not incorporate a mechanism that will allow MSOs to implement the technology. The technology addressed in this paper may belong to the Society of Cable and Television Engineers (SCTE) as they claim responsibility for everything in between the encoder and decoder. The future of technology is certainly in doubt, thus a flexible and extensible architecture is a vital component for a successful cable future.

Multiple CODECs will exist concurrently on the network. The network will evolve and building an architecture that allows for expansion is integral to our future success.

MOVING TO AN ALL-IP, ALL-AVC HFC ACCESS: OPPORTUNITIES AND CHALLENGES:

Carol Ansley and Mark Bugajski, ARRIS Group

Paper Outline:

1. A typical MSO HFC downstream usage – "staying the course, or not"?
    a. "Staying the course" path option
    b. Flash-transition to all-AVC in HFC
    c. Bandwidth usage comparison
2. VBR H.264 AVC over DOCSIS3.0 option benefits
    a. VBR SPTS UDP streaming
    b. HTTP FMP4 streaming
3. Demarcation Gateway (DGW)
    a. Architecture
    b. Support for legacy home CPE
    c. Building blocks
4. Conclusions

Stay the Course or Jump Ahead?

Multiple System Operators (MSOs) have evolved over the last ten years from delivering only video services, originally in analog, to providing a triple play of video, voice and data. Along the way their capital investments have been directed at each service in turn, usually success-based. But this paradigm is starting to show its age, and may become unsustainable if recent trends in HSD growth and increasing competition from Telco and satellite sources continues.

A typical HFC plant today provides some analog services, a large digital tier of broadcast linear programming, and an expanding tier of narrowcast video and data services. Voice services take a small portion of upstream HSD bandwidth, but are generally negligible in the downstream direction. The remainder of the upstream bandwidth is used for HSD and STB return signals.

With the ever increasing pressure to provide more HD program content, particularly from the satellite providers, MSOs must find ways to increase the number and quality of video programming options for their subscribers. There is also substantial pressure on the HSD capability over the HFC plant to match FTTx deployments with their high downstream and upstream capabilities.

To increase video capacity many operators have deployed, or are considering deploying, switched digital video deployments, as well as decreasing the amount of spectrum allocated to inefficient analog services. Both of these approaches present public relations obstacles and require capital outlays. For example, to deploy switched digital services one must install a control plane to manage the switching of the channels and deploy narrowcast QAMs for each service group, as well as commit operational resources to carefully monitor and balance the SDV program loads over the channels committed to the narrowcast SDV services. To reduce analog services would seem to be a simpler option technically, but it has business implications due to must-carry agreements, franchise agreements, and other business arrangements that may prevent a channel from being moved to a digital tier or removed outright. In the aftermath of the digital transition, MSOs have acquired some subscribers who specifically wanted to be able to continue to use their analog televisions. To accommodate these implications, MSOs have sometimes chosen to also deploy DTAs as a part of the analog removal, which requires capital investment.

Another trend that is having a profound effect on MSO capital and operational investments is the spread of personal video recording technology. Recently it was reported that PVR technology has spread to greater than 40% of North American households. The consumer has become accustomed to being able to time shift their consumption of video content. To accommodate these expectations, the MSOs have had to deploy advanced settop boxes that can provide these services, with the requisite increase in their subscriber capital costs. Another option that championed by at least one operator is moving the PVR functionality into the headend servers. This option decreases the capital outlay necessary for widespread PVR service offerings, but does have a concomitant increase in bandwidth usage as subscribers move from the broadcast program stream over to the unicast streams.

One last area that is expected to provide increasing pressure on the MSO's plant is the upstream speeds required to remain competitive in HSD service. DSL and FTTx services are typically symmetric, while the present cable plant is profoundly asymmetric. To

address the asymmetry will require either extensions to the high end of the cable plant to provide increased upstream service, or changes to the downstream plant range, converting some portion of it to upstream use. Either of these changes has implications for the customer premise equipment, and the basic hardware of the HFC plant. The MSOs have not identified a preferred solution, but market pressures are expected to eventually force the MSOs to choose one of these alternatives.

Recent FCC proposals may also lead to fundamental changes in the home network architecture for the delivery of video and data services. The recently released Broadband Plan calls for the industry to rapidly adopt a Gateway architecture with the goal of providing a more open marketplace for CPE equipment. This proposal has a long way to go before it is adopted, but it would be best for MSOs to determine how they might cost effectively comply with that new requirement if it survives the legislation and oversight process.

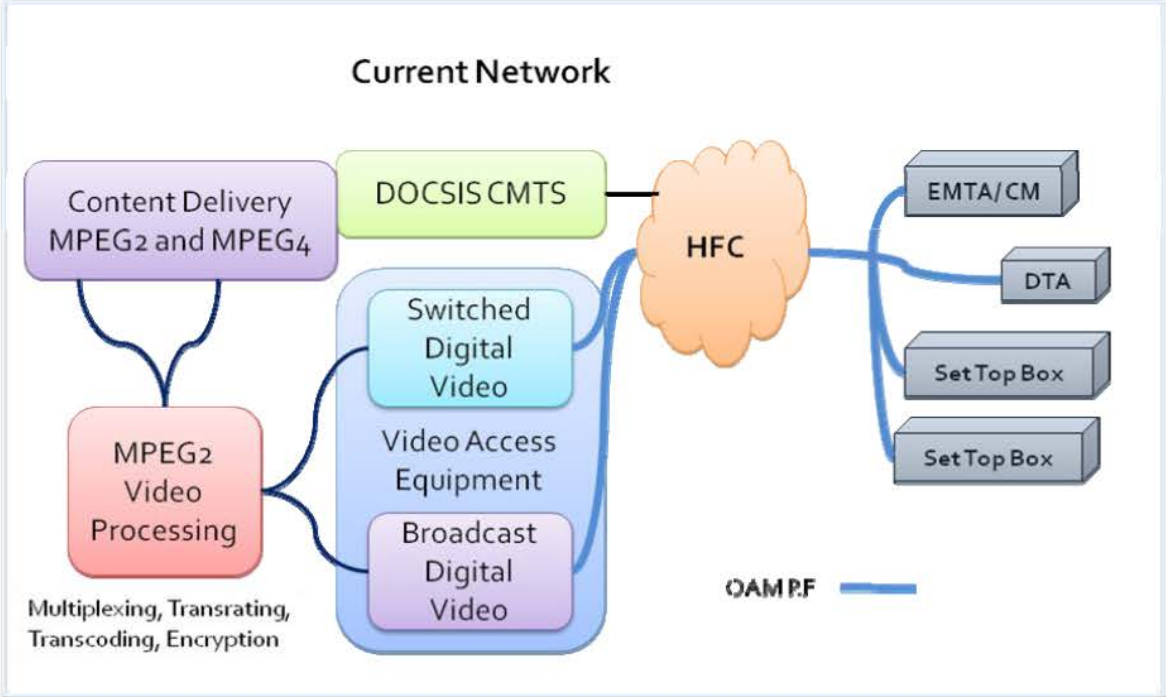
This paper compares the costs of continuing these activities with a seemingly radical alternative and rapid movement to an IP to MPEG gateway called the Demarcation Gateway to allow the legacy equipment to continue without replacement and which separates the network evolution from that within the subscriber's residence. It will compare the head-end and access changes necessary to provide the market demands of advanced video services and increased HSD bandwidth.

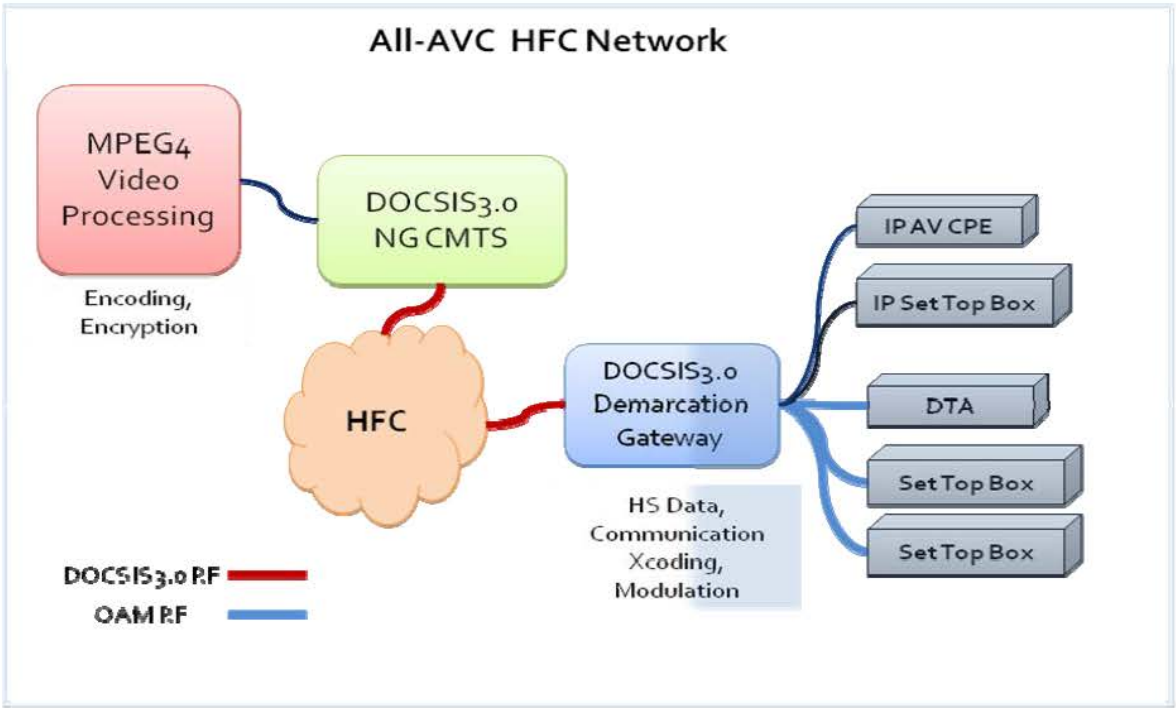
There have been proposals made periodically to move to a deployment model that makes use of a gateway to interface between the outside plant network and the subscriber's premises. The Demarcation Gateway proposed in this paper can disconnect the outside plant from the inside network, offering several advantages. The MSO network can evolve separately from the Home network. RF ingress from sources within the home can be blocked from outside network, improving network performance and reliability. Recent regulatory pronouncements also indicate that, within the US at least, regulators may impose a Gateway-based architecture on the industry. Assuming for the moment that this initiative has a fair chance of success, this paper analyzes some advantages to be had in accepting this direction, and using it to reach an advantageous long-term position.

The overall video industry has been developing and utilizing new compression technologies beyond the traditional MPEG2 encoding that is most common in MSO plants today. The H.264 AVC codec offers substantial bandwidth reduction over the MPEG2 codec in use today, but the millions of legacy set top boxes deployed already have hindered the widespread deployment of AVC within the CATV industry. AVC has been used to great effect within the satellite industry so decoders, encoders and robust CPE silicon solutions are readily available at reasonable cost points, but it has been the upfront cost of swapping out the CPE devices that made the business case for that swap out unacceptable within today's financial markets, even though there are many technical advantages.

The Demarcation Gateway architecture eliminates the legacy swap-out concern while allowing the MSO network to take advantage of the latest industry developments.

The network block diagram below shows the new architecture. The transport mechanism over the HFC transitions from MPEG2 over QAMs to IP Video over DOCSIS. Also there are advantages to transitioning from the constant bit rate (CBR) encoding commonly used for SDV deployments now to a variable bit rate encoding scheme that the wider data transmission pipes over DOCSIS 3.0 have enabled.





The new architecture presents an opportunity to take advantage of recent technological improvements (technology update) in several key areas without wholesale stranding of the current Capex investments. It will be possible to:

- Gradually and seamlessly transition to all-IP transmission in the access for full service and equipment convergence

- Use VBR formatted H.264 AVC encoded streams for significant improvements in content transmission efficiency
- Deploy highly integrated SoICs that enable development of a Demarcation Gateway
- Instantiate the viewing device specific high quality, while cost effective transcoding inside the DGW

We will next expand on some of these technical advantages ….

VARIABLE BIT RATE H.264 AVC CONTENT DELIVERY OVER IP ACCESS.

As stated earlier, the H.264 AVC encoding brings about significant expansion of the compression toolkit and allows the HD 1080i content to be encoded for streaming at average bit rates ranging from 5 to 8 mbps. The average bit rate is dependent on how the encoders (transcoders) are provisioned and what type of content is being encoded. As with MPEG2, content with static screens and slow-action (talking-heads) can be encoded at the lowest bit rate and sports programming requiring the highest bit rate for the same format.

ARRIS has performed very extensive studies on methods of unicast (VoD) streaming using Variable Bit Rate (VBR) SPTS formats that represent a departure from the traditional Constant Bit Rate (CBR) method. The VBR transport streams are compliant with the MPEG specs but have not to-date been used in the HFC networks because of the complexities and challenges associated with implementation of streaming servers and the downstream spectrum management. The 6MHz channelization of the HFC RF spectrum creates 40mbps delivery "pipes" that are too narrow to enable "self-averaging", the natural (not forced) statistical multiplexing of peaks and valleys of VBR streams.

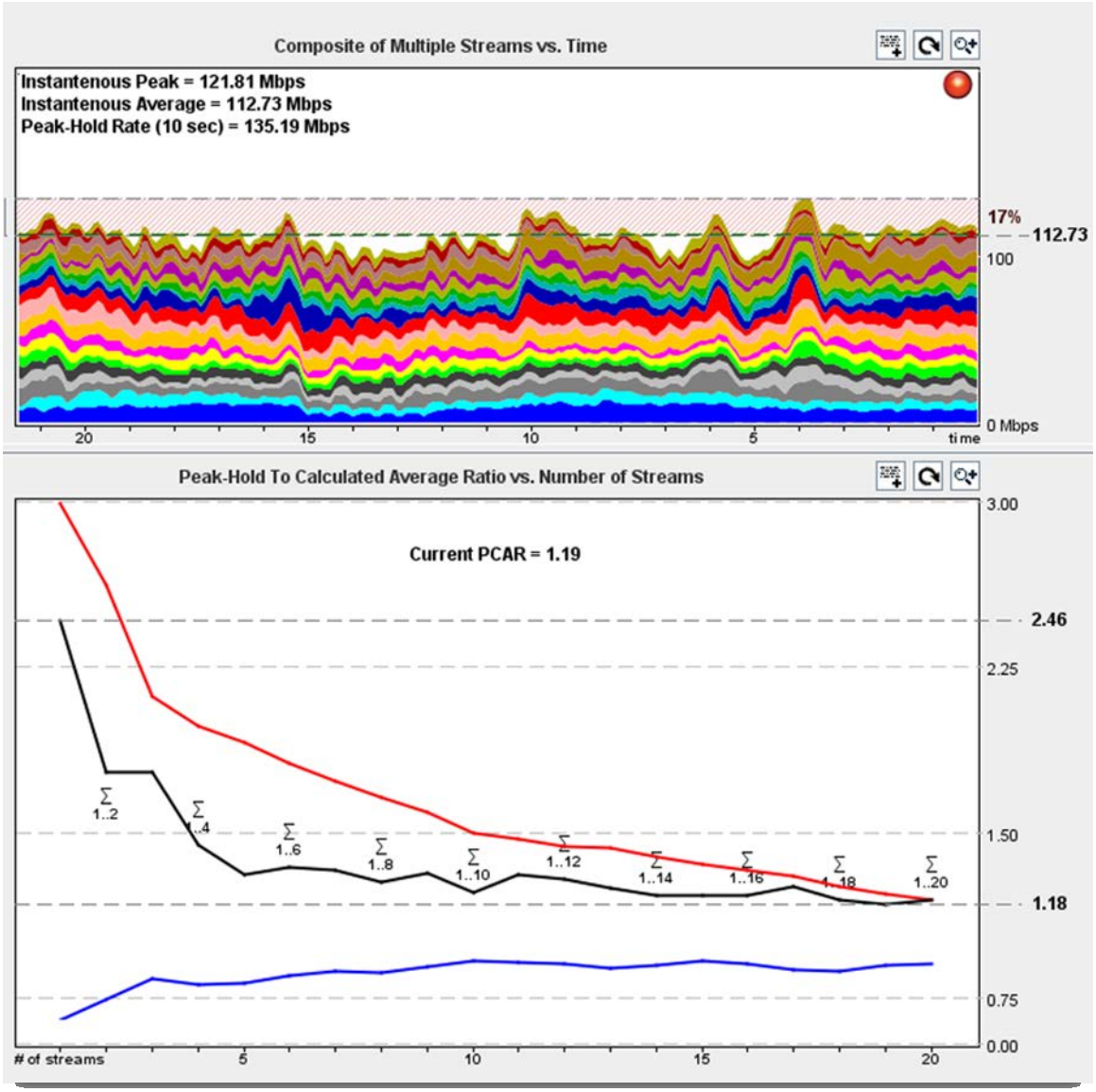
The channel bonding feature/functionality of DOCSIS3.0 breaks the 40mbps limit and allows for establishing delivery pipes that are 4, 8, 16 or more times wider that a single 6MHz channel. We have developed sophisticated research and simulation tools that allow us to simultaneously stream hundreds of combinations of continuous and fragmented VBR streams. We studied combinations of the most variable content (movie trailers and commercials) as well as several categories of time-shifted cable programming. During our research, the streaming sessions were run over extended periods of time to investigate random occurrence of cumulative peaks of the combined streams. The results of these studies have been published and presented at previous cable conferences.

In summary, VBR streams when allowed to self-average inside a D3.0 bonded-downstream pipe create a combined flow that is fairly predictable (as measured by the peak to average (PAR) indicator) and the bandwidth allocation for this flow is manageable with simple provisioning rules. Actual cable content was transcoded from MPEG2 to H.264 AVC format and used for the purpose of these studies.

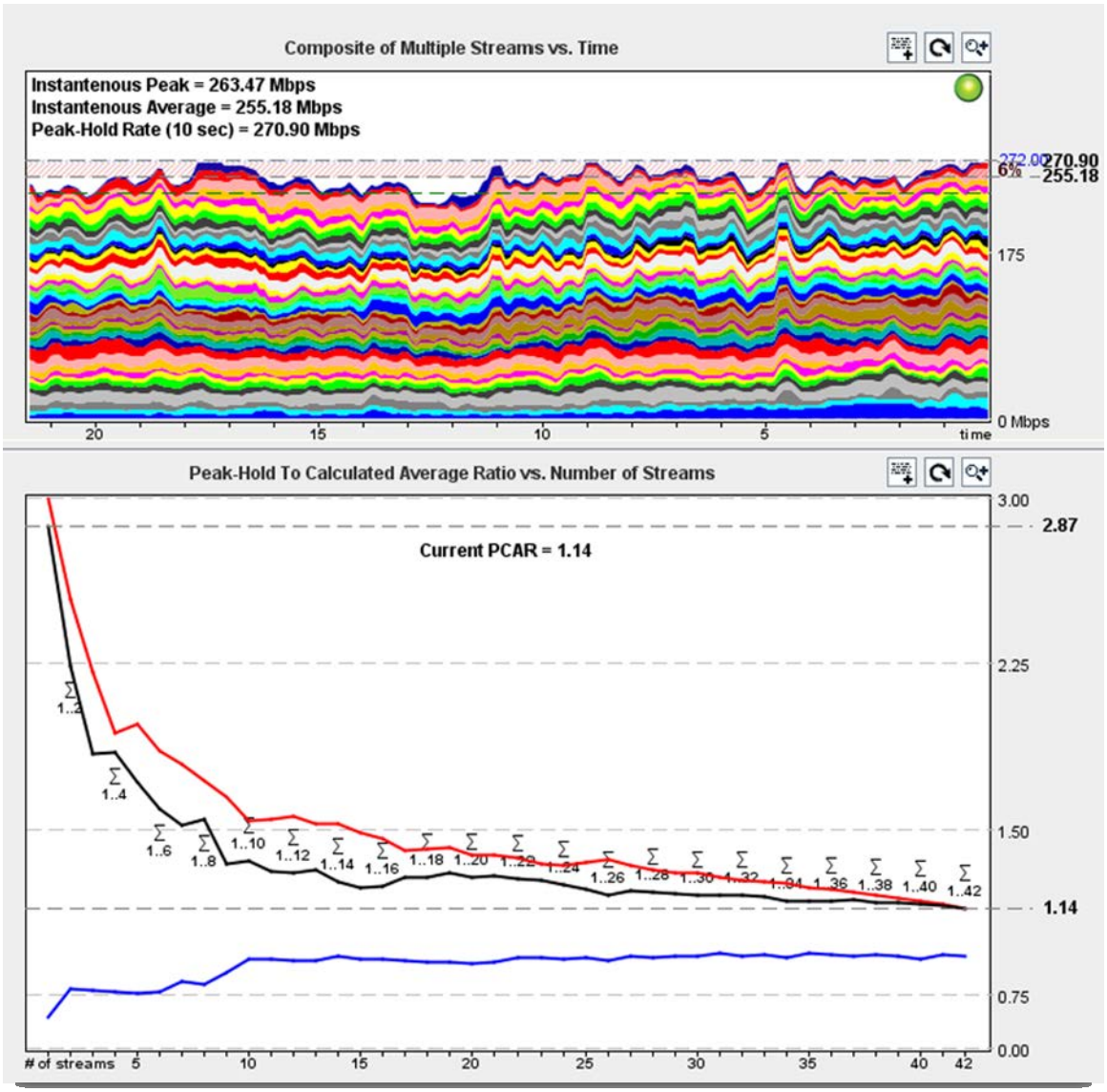
The following plots illustrate the results of one of our studies. A Peak to Calculated Average (PCAR) parameter has been defined as a ratio of the peak bit rate (calculated as a peak-hold inside a 10ms time window) of the VBR stream to the average bit rate that is calculated from the media file size divided by the duration of its playback. Both parameters are easily obtainable for on-demand assets and for the multicast (or time-shifted TV) and can be predicted from the encoder settings and type of content streamed.

The plots below developed during the study illustrate that:

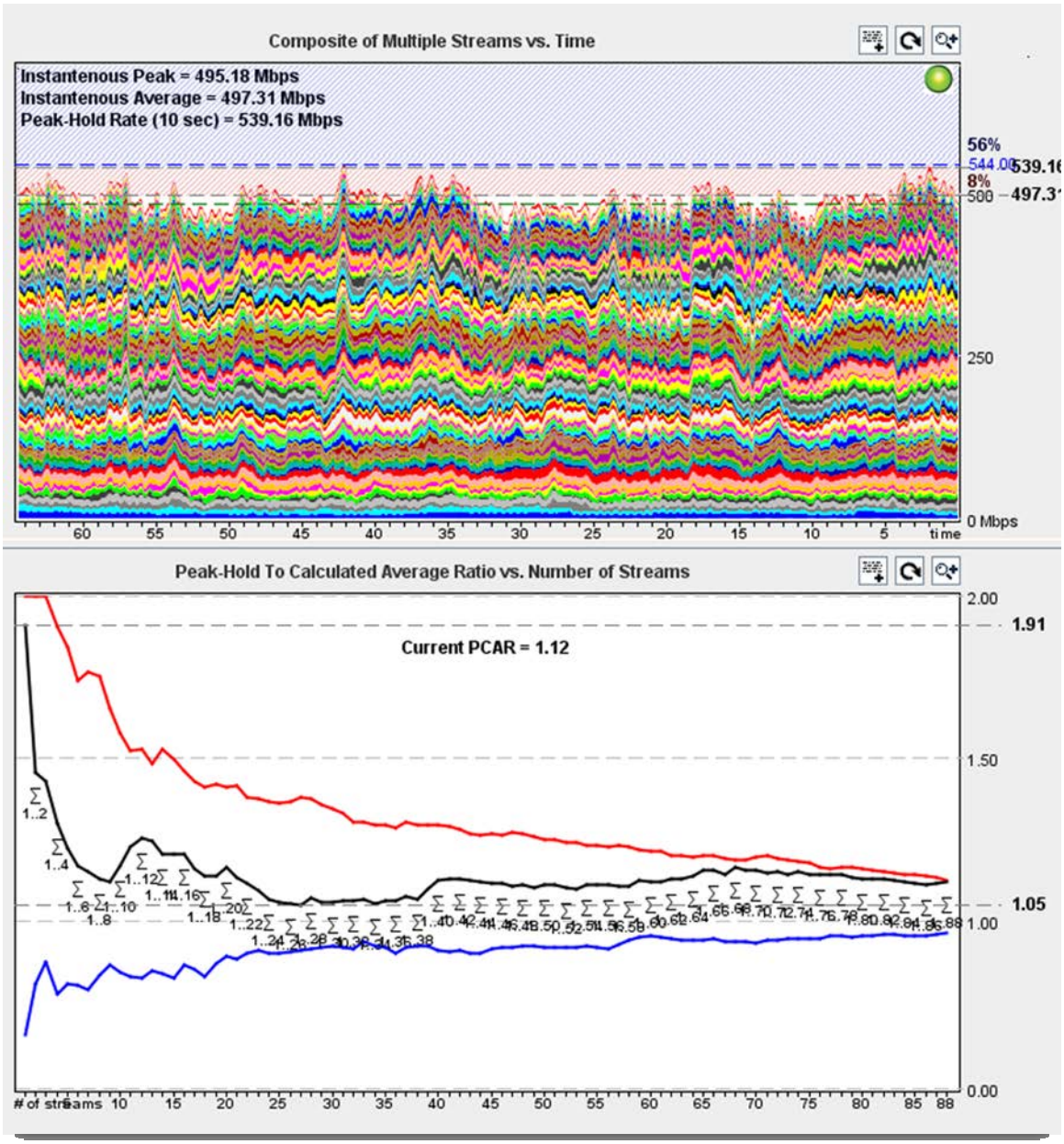
- More streams can be safely combined in a larger bonded pipe than in a smaller bonded group (as shown on X-axis of the plots)
- The PCAR variability is asymptotic to a value of 1 as more VBR streams are being combined thus assuring that peaks will not exceed the maximum bandwidth available



Plot 1. Example of 20 VBR streams inside one quad-bonded D3.0 pipe. The PCAR value (max hold value in red, min hold value in blue) is plotted vs. the number of streams combined.



Plot 2. Example of 42 VBR streams inside one octal-bonded D3.0 pipe. The PCAR value (max hold value in red, min hold value in blue) is plotted vs. the number of streams combined.



Plot 3. Example of 88 VBR streams inside one sixteen-bonded D3.0 pipe. The PCAR (max hold value in red, min hold value in blue) is plotted vs. number of streams combined.

In our study, multiple clients receiving the VBR streams were simulated in Java environment. The streaming method was linear and used the UDP IP protocol. The streams' time positions were regularly randomized to simulate user interaction with the content.

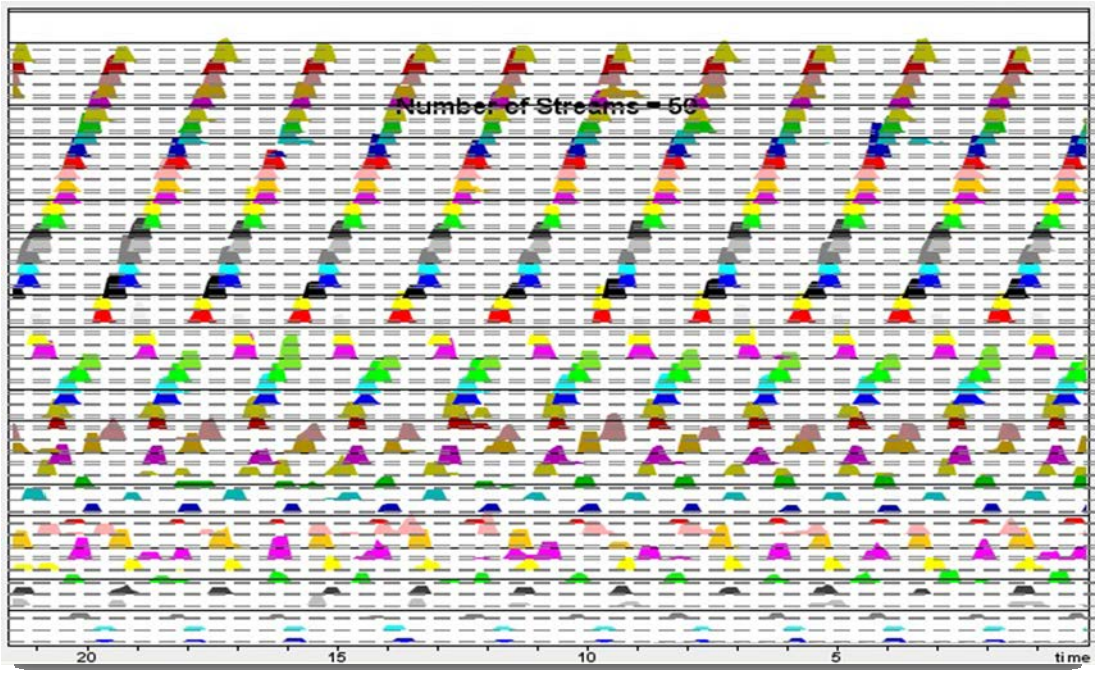
Peak to calculated average (PCAR) of individual streams was typically between values of 2 and 3 as the encoders that generated the content under study were set to 12mbps peak rate.

The emerging, standards-based Fragmented MP4 method of streaming of VBR formatted "chunks" of content lends itself even more to delivery over D3.0 CMTS than the continuous VBR (MPEG2TS UDP). The sizeable video buffers inside FMP4 compliant clients allow for increased flexibility of the inter-

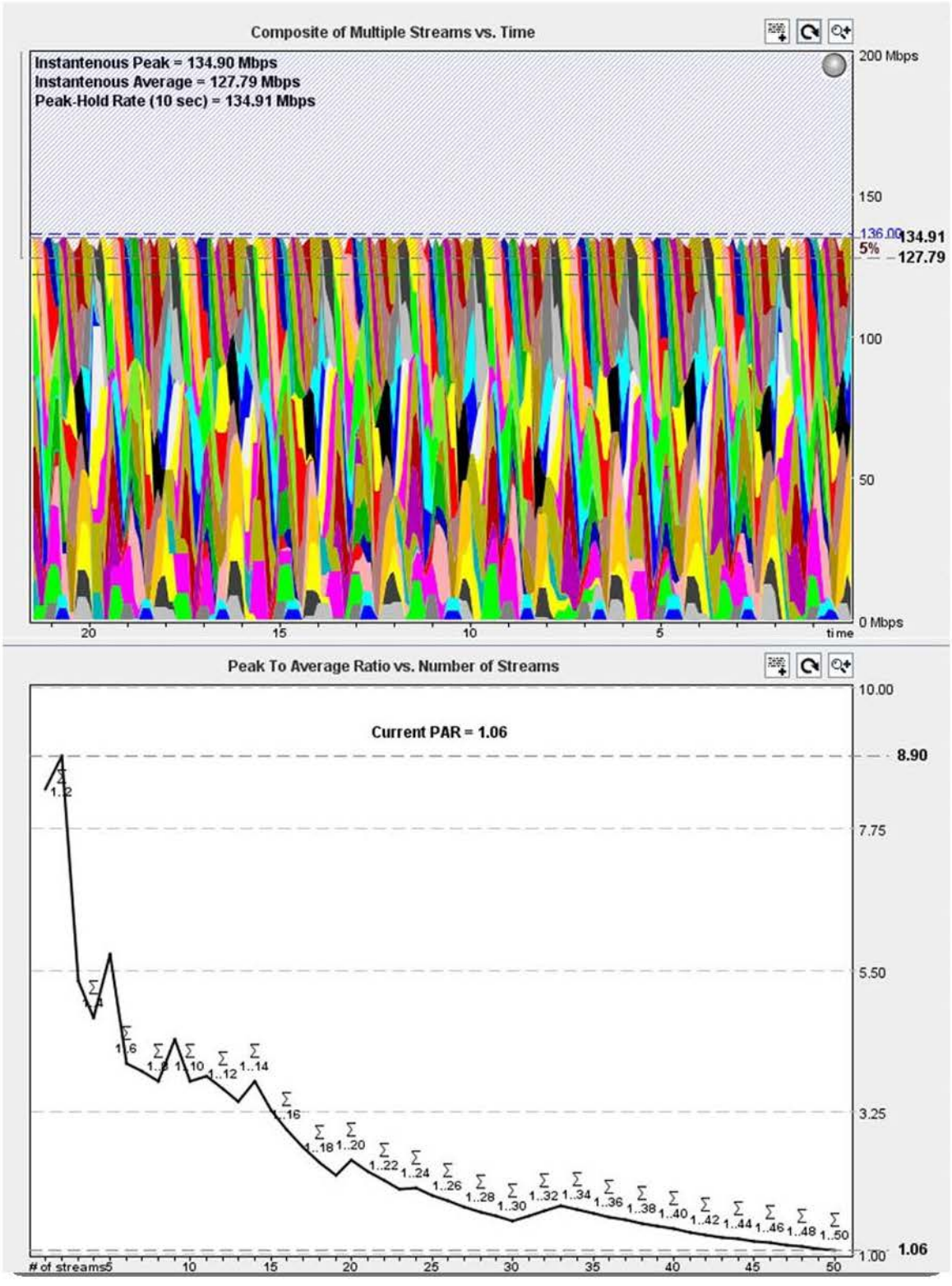
arrival time of the fragments. The player will wait for the fragments arriving from the server for much longer durations of time, while continuing to play content from its buffer. As a result, the streaming process is very tolerant to queuing in the CMTS and the overall Content Distribution Network (CDN). The queuing function, a built-in DOCSIS feature, acts as a traffic shaper that uses all transmit opportunities to send the variable fragments (as they arrive from the FMP4 servers) towards the client devices. As shown on the two plots below, the CMTS can act as an elastic buffer that spreads and smoothes the peaks of the combined downstream flows.

The FMP4 studies showed that the Peak to Calculated Average Ratio (PCAR) values of many sequential bursts (responses to clients' "get" requests) that are combined inside the quad or octal bonded D3.0 pipe stays very close to the ideal value of 1 for long durations of time. Plots 4 and 5 below illustrate how peaks of transfers of individual content fragments are smoothed out by bandwidth shaping applied to the entire flow.

Plot 4 shows the bursts for individual client devices and Plot 5 shows all of the bursts combined inside a quad-bonded downstream pipe. The PCAR value of the combined FMP4 flow traverses faster towards the value of 1 as more bursts are combined than in the case of VBR UDP streaming described earlier in this section.



Plot 4. A view on 50 FMP4 individual http transfer flows towards the client devices combined inside the wide-band D3.0 downstream pipe. The peaks are smoothed-out by a queuing mechanism.



Plot 5 50 FMP4 flows (consisting of regular bursts) to individual clients combined inside a quad-bonded downstream

In managing the UDP or FMP4 VBR streaming it is very critical that the average value of the combined flow (calculated or estimated as a sum of the averages of the individual streams over an extended time (continuous or fragmented)) does not exceed the bandwidth available in the DOCSIS pipe. Simply speaking, the laws of FIFO "physics" have to be respected and any bandwidth smoothing buffer will overflow and some packets dropped if more packets enter the FIFO than can be sent to the HFC.

Ours VBR streaming studies from analyzing many streams of various CATV content can be summarized as follows:

- The average bit rate (ABR) of H.264 AVC 1080i stream encoded with a real-time commercial coder ranges between 5 and 9 mbps. This range can be shifted lower by 1 mbps average when off-line, file based encoding is performed
- The instantaneous value of Peak to ABR of any stream varies between 0.6 and 3. Again, for off-line encoding the PAR value can reach a much higher value.
- Combining of multiple streams reduces the variability of P/ABR of the composite flow (VBR self-averaging effect) to the range of 0.9 – 1.1 for 500 mbps flows
- Efficiency of VBR self-averaging strongly depends on several factors:
    - Number of VBR streams combined, up to a point of diminishing returns
    - Encoder (transcoder) settings, maximum bit rate and the strength of compression
    - Mix of content in the combined flow (ratio of talking heads to movies and to sports)
- A "light touch" of delay applied to some or all streams further smoothes out the peaks and maximizes the effects of self averaging while preventing loss of video packets. Attention must be paid to keep the delay at a value lower than that of any receiving CPE
- The fragmentation of VBR content (files or live streams) and delivery of such "chunks" in response to regular HTTP "get" requests by clients with very deep (up-to 20 seconds) input buffers further maximizes efficiency of content delivery over a D3.0 system.
- DOCSIS overhead has to be included in the actual HFC bandwidth provisioning but it does not have a significant impact on the improvements in the overall transmission efficiency.

VBR bandwidth can be managed in the backbone and on the HFC by using the simplified provisioning parameters

- A conservative average value of 8mbps can be used for provisioning VBR VoD/SDV (H.264 AVC @ 1080i content) over octal-bonded D3.0 downstream facility. This number is significantly lower than today's rate of 15mbps of HD VoD MPEG2 encoded CBR streams.
- The same rate of 8 mbps rate can be safely used for provisioning of Broadcast VBR. Today's 3 into 1 HD stat-muxing results in an average bit rate of 13.3mbps MPEG2 CBR (one third of ~40mbps MPTS)
- Use of jumbo Ethernet frames (14 and higher MPEG2TS packets in a frame) further improves the transmission efficiency (reducing the 3% (ETH+IP+UDP) overhead)

Significant additional savings in CAPEX and OPEX, or expenditure avoidance can be realized with the transition to the H.264 AVC because no video processing related to improving efficiencies of content delivery using 6MHz channelized HFC will be needed:

- No stat-muxing will be required for Broadcast streams to fit a maximum number of them into 40mbps bandwidth of a single 6MHz channel
- No grooming will be needed to CBR for VoD/SDV services

The content will traverse through the delivery system in its "native" format as generated by the encoders. Its quality will be maintained as generated at the source.

RF Bandwidth Savings/Recovery Strategy Opportunities

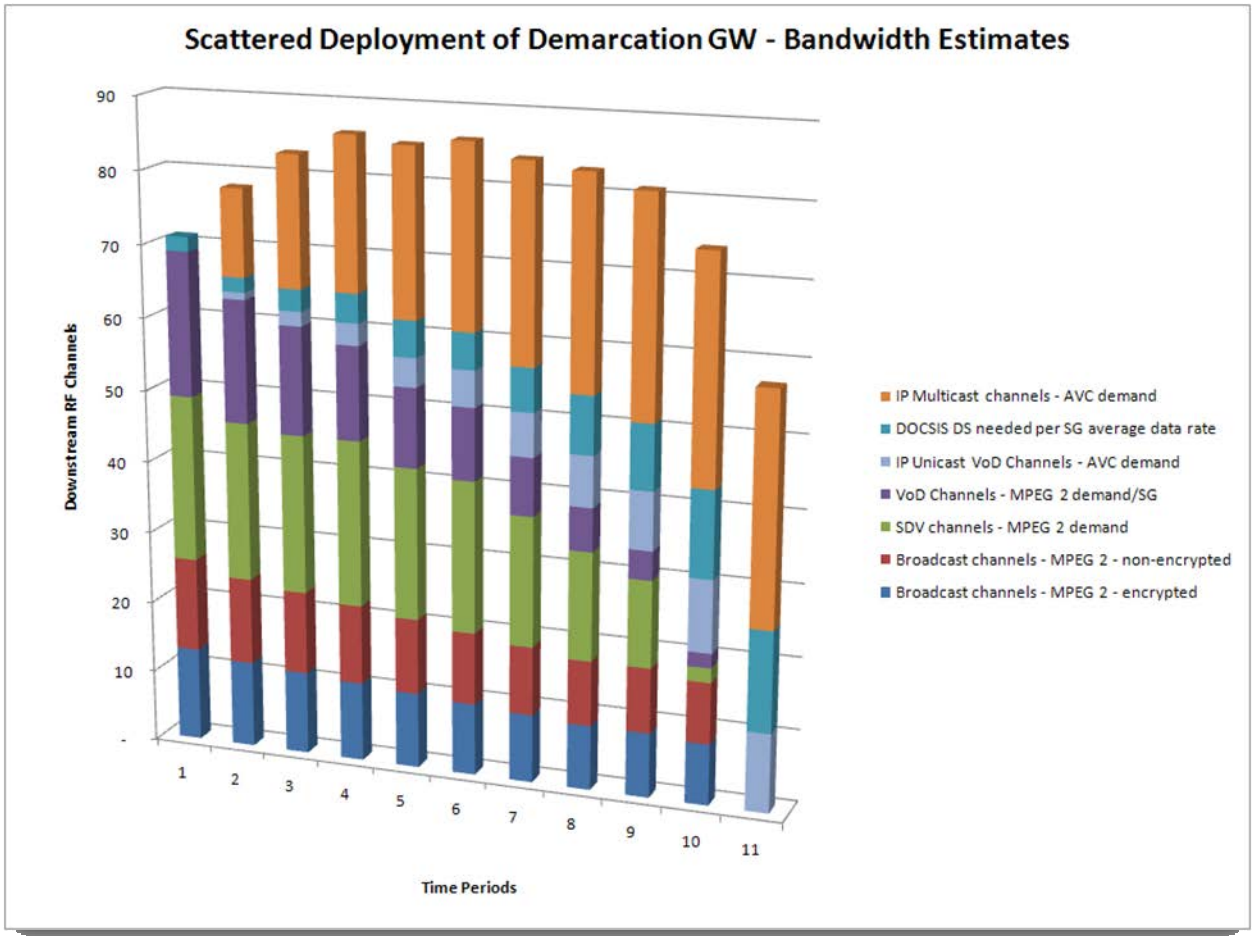
According to our on-going analysis of bandwidth usage trends by large MSOs, continuing with the traditional deployment paradigm may yield a possible bandwidth map as shown in the following diagram. The current MPEG2 channel distribution is contrasted with a possible MPEG4-based channel

distribution that would provide all current subscribers with the same services, except that the video is delivered over using H.264 AVC over IP.

It is clear and intuitive that once the operator has transitioned from MPEG2 to H.264 AVC over DOCSIS, the bandwidth usage gains and recovery can be very significant thus allowing the operator to expand current offerings and aggressively grow new services.

The transition to an all-IP AVC environment is the most difficult part of the strategy.

As illustrated below, a gradual transition that is based on slowly "retiring" the MPEG2-only STBs, while introducing the new AVC CPE unfortunately results in a substantial simulcasting penalty. The serving areas with a mix of customers who have been moved to IP Video, and those who have not, still must carry a duplicate suite of video services. Within each node this could lead to a transition with increased RF usage over a long period of time.



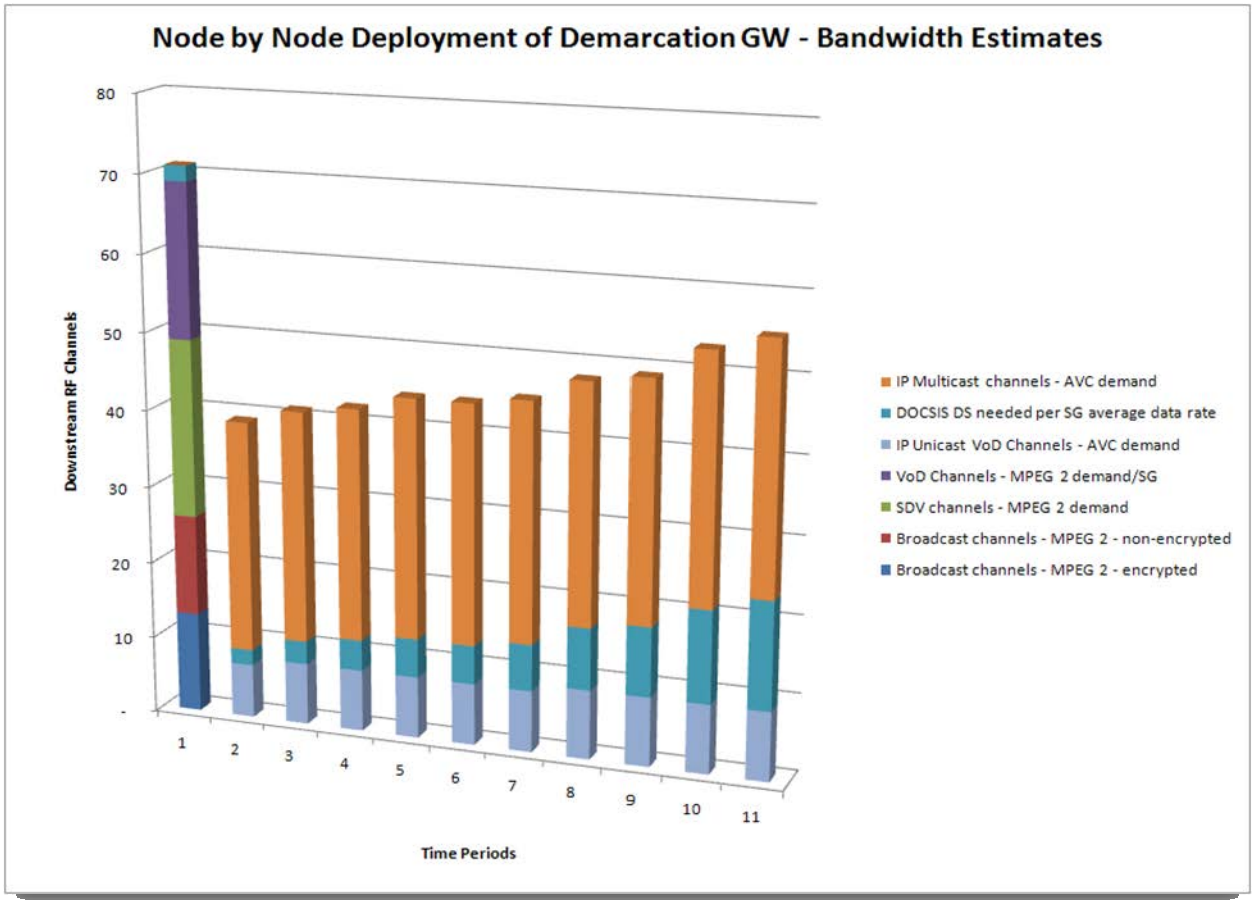
Such a gradual transition does not make efficient use of capital or operational investments.

In contrast, a node-by-node, wholesale change to VBR AVC in the access strategy can be enabled by pre-installation of the Demarcation Gateway in the entire serving area followed by a flash cut-over to all-AVC HFC. Such a strategy will allow the operator to switch quickly from one channel map to another and start expanding the offerings without a need for simulcasting.

The Demarcation Gateways can be deployed within a targeted node, without disturbing the existing

services with the help of an RF pass-through path. The operator may deploy the DGWs outside of customer premises at the HFC/home demarcation point, thus eliminating the need to work around subscribers' schedules and improving overall operational efficiencies later on.

When the head-end AVC infrastructure is ready, and the DGW deployment is completed, the DGWs can be commanded to disconnect the pass-thru path and begin providing modulated RF signals to the home STBs to replace the MPEG2 services that are now decommissioned in the HFC.



The biggest advantage of this approach is when the costs of the transition are considered. The ability to avoid replacing the existing in-home CPE provides a significant cost benefit. Lower cost IP STBs can still be deployed once the Demarcation Gateways are in place, but that investment is not required immediately. The STB replacement investment can be transformed into a success-based selective investment that happens at a slower pace.

THE DEMARCATION GATEWAY

In the early 1990's ARRIS pioneered a two-way HFC specific telephony Network Interface Device as a part of the Cornerstone product portfolio. More than ten million Voice Ports were installed on the side of houses served by MSOs around the world. Design of such an active demarcation point of deployment device offers engineering challenges that were proven to be surmountable. Combining the know-how from the best of telco and cable RF worlds and some hard-core engineering work resulted in a device that was five 9's robust while consuming less than a Watt of power most of the time.

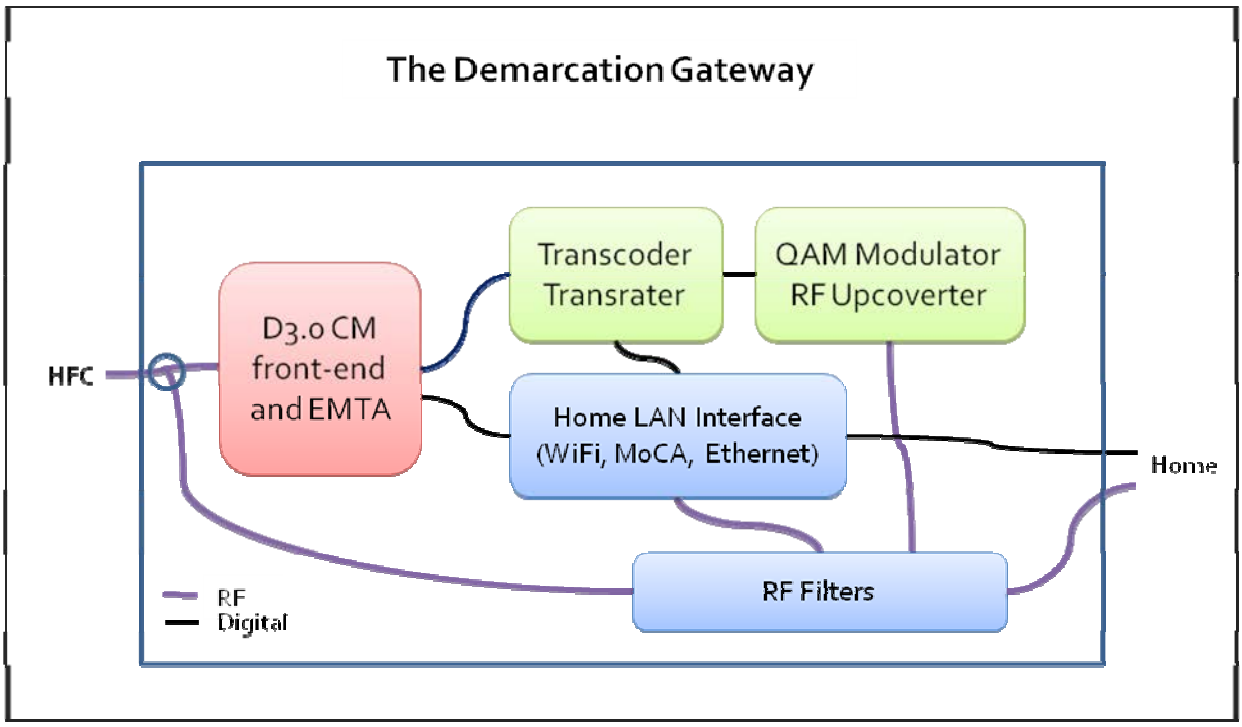
The subsequent proposals and even some trials have attempted to address a whole home Gateway device, and have even explored on-site transcoding possibilities for MPEG2 support, but cost and performance have been barriers that prevented serious consideration of these ideas.

The relentless advance of silicon integration and processing capabilities has overcome these barriers. The same advances that have empowered the latest generation of DOCSIS3.0 capabilities can also be used to enable the Demarcation Gateway. We all can relate to the shining example of integration and true multi-functionality of Apple portables -- today's functionality would require racks of equipment to implement not so long ago…

The bonded channels of DOCSIS3.0 have opened a wide pipe to deliver all content over "bursty" IP to the home. The resulting statistical multiplexing gains due to the change from standard MPEG digital channels to bonded DOCSIS channels are significant. The technology that enables these RF advances was developed in response to D3.0, taking advantage of technical developments in the cellular phone and wireless technology markets.

Silicon integration has also been continuing on the video processing front. Only a few years ago an MPEG2 to MPEG4 HD transcoder was a top of the line 2RU encoder, then it was shrunk into a stand-alone co-processor chip next to a set top box video processor. The upcoming generation of set top box silicon will commonly include this feature within the main set top chip. The processing requirements of a Demarcation Gateway will be met at a reasonable price point, far less than would be required to replace the existing MPEG2 Set top boxes within the home. It is worth stating that the bandwidth at home is significantly cheaper than on a crowded HFC network and the transcoder can generate much higher b/w MPEG2 stream for the same resolution as today's rack-mount, HE based xcoders

The Demarcation GW can be architected as illustrated below:



Key components of the DGW are in place already although implemented as separate subsystems and located in different HFC network elements and home devices.

The two new functional blocks, additions to what we can find in currently designed cable transport GWs, are the micro-QAM and RF up-converter and the transcoder/transrater.

The QAM+RF design can be easily "borrowed" from next generation super-dense EQAM devices currently under development by several vendors. Our estimates show that the two 6MHz channels that can be generated in this subsystem will provide sufficient transmission capacity for the very popular, multi-tuner HD DVR. The digital modulator can synthesize QAM carriers directly onto the desired RF channels, thus avoiding the need for a complex and power-hungry RF upconverter.

Integration of AVC to MPEG2 transcoding is critical to enabling the continual use of tens of millions of MPEG2-only STBs. The functionality proposed for integration within the DGW is already widely deployed in home AV appliances from CE vendors like Sharp, Hitachi, Toshiba, NEC, etc. Such coding conversion features are being widely deployed in the home for a variety of reasons.

(a) Storage expansion on PVRs to increase storage capacity on a fixed-size HDD

(b) Enabling smart phones to tap to HD broadcast (scaling)

(c) Allowing PVR based content to be viewed on smart phones

The prices of stand-alone silicon chips that can perform these functions are moving below $20. Integration with other building blocks on a same die of a SoIC will translate into a significantly lower cost of the transcoding function. Silicon fab house such as TSMC and UMC have consistently been shrinking silicon geometries from 180nm through 130nm, 90nm, 65nm to 40nm now. As these geometries decrease, we are seeing that the silicon area to integrate the many hardware blocks has shrunk approximately 20 times with a transition from 180nm to 40nm. This brings an astounding opportunity to integrate; power and cost reduce the needed functionality onto a single SoIC.

The cable industry has the footprint, a large deployed MPEG2 STB base and if united, can create production volume requirements for its vendors which can make the investment of creating a DMG on a hardened SoIC a viable business case.

Conclusions

ARRIS believes that today, more than ever before (and enough) "stars have aligned" for a successful development of a Demarcation GW. Let us re-iterate the convincingly influencing factors for making the DGW (as outlined above) case:

- The deployment of DGW may free-up 50% of the HFC RF spectrum, the strongest "weapon" that MSOs possess. Delivery of VBR formatted, AVC encoded content in the DOCSIS3.0 downstream bonded "pipe" brings about this saving.
- Use of the NG DOCSIS3.0 technology in the HE (to drive the DGW's front-end), pioneered by among others, the Comcast's CMAP initiative, will bring the access equipment cost down as a significant step, an order of magnitude similar to the transition from D1/2.0 to D3.0.
- The DGW may enable a flash transition to the long-awaited, all-IP access on a serving area by serving area or node by node basis.
- The DGW installed on the side of the house will allow an operator to non-intrusively and transparently meet all communications and entertainment needs of the customer in a powerfully "sticky" way.
- The technology for fast-prototyping and fine-tuning the DGW is here now.
- The silicon geometry to integrate the DGW functional blocks is already in place at the foundries that serve the cable industry.

NEW HFC OPTIMIZATION PARADIGM FOR THE DIGITAL ERA

Jan de Nijs (TNO), Jeroen Boschma (TNO), Maciej Muzalewski (VECTOR) and Pawel Meissner (VECTOR)

Abstract

A cost-effective way to expand the capacity of HFC networks is a most efficient use of the amplifier power. The maximum output level of components is limited by the non-linear behaviour. In the current practice, 2nd (CSO) and 3rd (CTB) order non-linear behaviour is thought to limit the performance, and thus the maximum signal load. Studies of the ReDeSign project though, shows that in case of digital loads the performance is not limited by 2nd and 3rd but by 4th and 5th order non-linear behaviour. In this paper we present proof for the above, followed by a first specification of the 4th and 5th order non-linear amplifier parameters. To conclude we show and discuss the match between the measured performance of an amplifier and the simulated performance using a 5th order non-linear component specification.

INTRODUCTION

Considering the rapid growth of the customer bandwidth demand, the management of the network capacity must be considered business crucial. In practice, the network capacity is the resultant of the system balance of Figure 1. Given an RF power budget, an operator has to establish an appropriate balance between the network load and the quality of the delivered signals. Currently, network capacity related decisions like adding digital channels, replacing analogue channels by digital ones or raising the signal level of the digital channels are based on the practical experience of the RF engineer, basic RF calculations and measurements using a test cascade in a laboratory. In the ReDeSign[1] project we have studied the possibility of numerical network performance simulations as an advanced alternative for RF network calculations.

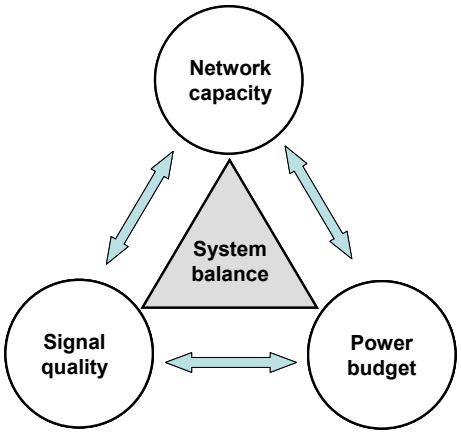


Figure 1 HFC Network system balance

The basic concept of a network performance simulation is given in Figure 2. As an input, the network, the active components and the network load have to be fully specified. Next, the algorithm calculates all signal levels and the distortion signal levels at the system outlet.

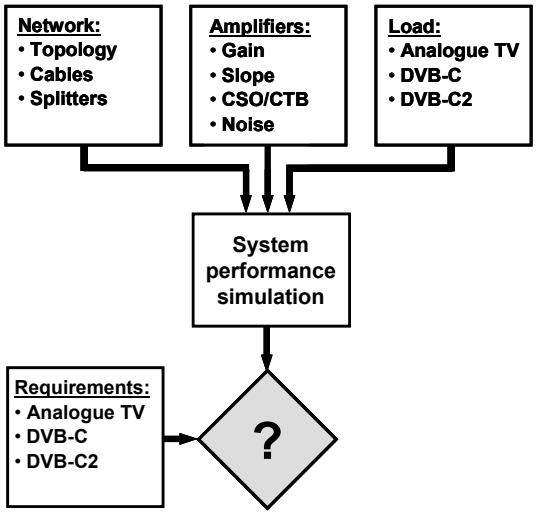


Figure 2 Network performance simulation

In this treatise, we will discuss the feasibility of the above approach to predict the signal levels and in particular the distortion signal levels in case of single components. We will compare results from measurements and simulations. The specification of the active components is the most challenging aspect.

The studies presented in this treatise is part of the ReDeSign project. Further details can be found on the ReDeSign web site.[2]

THEORETICAL FRAMEWORK

In agreement with standard signal theory, the non-linear behavior is described using a Taylor expansion:

$$y(t) = a_1x(t) + a_2x^2(t) + a_3x^3(t) + a_4x^4(t) + a_5x^5(t) + \ldots \qquad \textbf{Eq. 1}$$

In case of a Gaussian input signal $x(t)$, this time-domain response function can be restated into a frequency-domain description using the Price Theorem[3]:

$$Y(\omega) = A_1X(\omega) + A_2X(\omega) \otimes X(\omega) + A_3X(\omega) \otimes X(\omega) \otimes X(\omega) + A_4X(\omega) \otimes X(\omega) \otimes X(\omega) \otimes X(\omega) + \ldots \qquad \textbf{Eq. 2}$$

In the current cable approach, only the 2nd (CSO) and 3rd (CTB) order terms are taken into consideration. Different measurements have been defined to specify the 2nd and 3rd order parameters, amongst others a measurement of the 2nd and 3rd order distortion products when applying a load of unmodulated carriers on a periodic frequency grid. For example, in Europe a CENELEC load of 42 unmodulated carriers is used to specify the CSO and CTB performance of amplifiers.

Eq. 2 states that the distortion products are generated by the convolution of the input signals. Therefore, these distortion products often are referred to as intermodulation (IM) products. In case of a mixed analogue-digital load, the composite IM signal encompasses distortion signals generated by intermodulation of *i)* analogue TV with analogue TV signals (IM_{AA}), *ii)* analogue TV signals with digital signals (IM_{AD}) and *iii)* digital with digital signals (IM_{DD}). Likewise, third order IM_{AAA}, $IM_{AAD,}$… IM_{DDD} products are generated.

The analogue TV signal can be considered as a narrowband signal because of the dominance of the unmodulated carrier signal during the blanking periods. In contrast, digital signals have a broadband nature. Because of this different nature, also the intermodulation products have a different nature; IM_{AA}, IM_{AAA}, IM_{AAAA}, etc. are all narrowband distortion products (the CSO and CTB composite cluster beats) whereas any intermodulation product with at least one digital carrier (IM_{AD}, IM_{AAD}, $IM_{ADAA,}$…) has a broadband nature with a Gaussian signal level distribution.

SIGNAL DEGRADATION STUDIES

The signal degradation for a load of *i)* 42 unmodulated carriers and *ii)* 95 digital carriers was studied for a number of amplifiers. The amplifiers 2nd and 3rd order behaviour was specified using a standard CENELEC CSO/CTB measurement with a load of 42 unmodulated carriers.[4]

Load of 42 unmodulated carriers

For a CENELEC load of 42 unmodulated carriers, the narrowband CSO and CTB cluster beat levels at different frequencies using a spectrum analyser with 50 kHz measurement resolution. These measurements were performed for increasing carrier levels and for different components. A typical result of the signal-to-distortion signal ratio (SNR) curves for a low, mid and high frequency is shown in Figure 3.

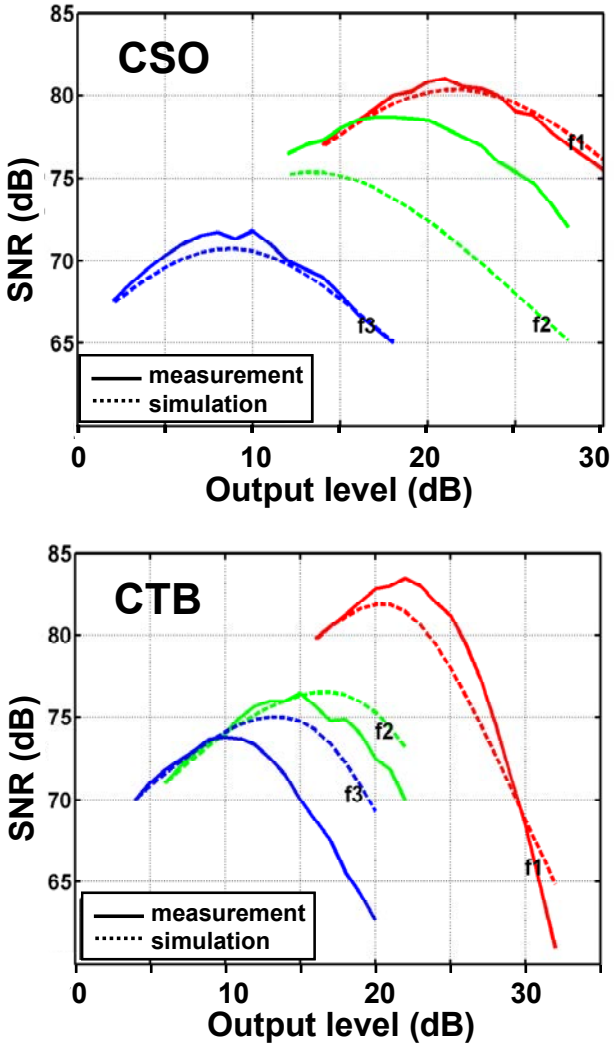


Figure 3 Signal-to-distortion signal ratio (SNR) for the 2nd (CSO) and 3rd (CTB) order distortion products in case of a load of 42 unmodulated carriers and when gradually increasing the carrier level. The SNR curves are shown for 3 frequencies, f_1 = 120 (CSO) and 119,25 (CTB) MHz, f_2 = 424 (CSO) and 423,25 (CTB) MHz and f_3 = 856 (CSO) and 855,25 (CTB) MHz.

With the aid of a dedicated simulation tool the SNR curves were simulated as well.[5] For these simulations we used a 3rd order component model and the CSO and CTB specification data as obtained from the specific components. As such, the simulations can be considered as a smart extrapolation of the CSO/CTB distortion levels for lower and higher carrier levels than the level of the specification measurement.

All measured SNR curves show the expected behaviour; for a low carrier level the SNR increases with slope +1 reflecting a constant (thermal) noise level, whereas for higher carrier levels the curves decline with a slope -1 (CSO) and -2 (CTB), respectively associated with the generation of 2nd and 3rd order intermodulation products. A comparison of the measured and simulated curves reveals a qualitative agreement. All curves have a congruent shape and in particular the slopes of the negative asymptotes match well.

Summarizing this result we can conclude that the simulations provide a fairly good prediction of the SNR curves and in particular the order of the degradation mechanism is correctly predicted; however, the measurement and simulations certainly do not match exactly.

Load of 95 digital carriers

Next we studied the signal degradation in case of a digital load of 95 DVB-C carriers using the same amplifiers. The distortion signal level was measured in a vacant frequency channel using a spectrum analyser with an 8 MHz bandwidth resolution.

shows a typical result of such a measurement. With the aid of the regular CSO and CTB specifications of the specific components and using a third order component model, we have simulated the SNR curves as well.[5]

Comparison of the measured and simulated curves shows that the simulation does not provide a proper prediction of the measured curves. Clearly the simulation provides a too optimistic SNR curve, and in particular it does not predict the correct slope for high carrier levels. The measured curves approach an asymptote with slope -4 associated with a 5th

order degradation mechanism. An equal result was obtained for different amplifiers.

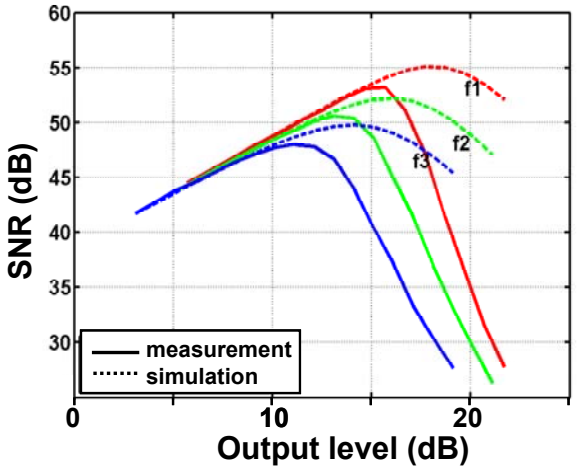


Figure 4 Signal-to-distortion signal ratio (SNR) for an amplifier with a digital load of 95 digital carriers, when step wise increasing the carrier signal level. The curves were recorded for the frequencies f_1 = 120 MHz, f_2 = 416 MHz and f_3= 856 MHz.

The above result demonstrates that in case of a digital load the degradation is caused by 4th and 5th order distortions and not by 2nd and 3rd order terms. As said, this result was obtained for different amplifiers. Moreover, as a rule the SNR curves of amplifiers with all digital loads reveal a steep decline of the SNR for high carrier levels, comparable to the curves Figure 4.

Analysis

The observed dominance of the 4th and 5th order degradation mechanism in case of digital signals can be explained in a straightforward manner. Suppose that an amplifier is first connected to a load A of 95 unmodulated carriers with a composite power level P and next to a load B of 95 digital carriers with the same composite power level P. In both cases, the same composite distortion signal power will be generated, although in case A the distortion signal is concentrated in a limited number of narrowband cluster beats with a high spectral power density whereas in case B the distortion signal is smeared out over the full frequency band. In case B a broadband distortion signal with a low spectral power density is generated. Next, we have to take the thermal noise level of the amplifier into consideration. This noise level specifies a minimum spectral power density level below which non-linear distortion products cannot be detected. Evidently, narrowband cluster beats with a high spectral power density of case A will surpass the thermal noise detection level for relatively low composite signal power level P. In contrast, for load B, a much higher composite power level P is needed to raise the smeared-out broadband distortion signal beyond the thermal noise level. At this level, 4th and 5th order distortion products may dominate the non-linear character of the component.

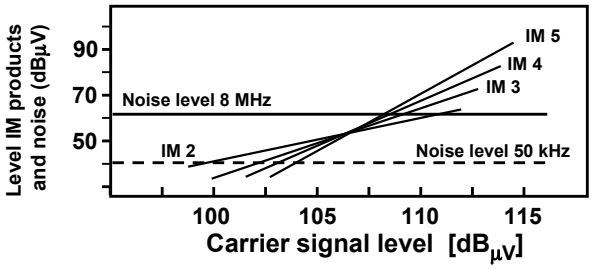


Figure 5 Illustration of the impact of the measurement resolution bandwidth on the measurement sensitivity for non-linear intermodulation products. See text for an explanation.

In the measurement of the non-linear distortion signals in case of a load of unmodulated carriers and of digital carriers, the different nature of the distortion products (narrowband versus broadband) results in a measurement with different measurement bandwidth of the spectrum analyzer, 50 kHz

versus 8 MHz. This results in a 22 dB lower measurement sensitivity in case of the digital loads as compared to a load of unmodulated carriers. In case of such a lower measurement sensitivity, a much higher carrier level is needed to detect the non-linear distortion signal. Logically, at this higher signal level the magnitude of the higher order distortions will have increased (much) more than the lower order distortions. This effect is illustrated in Figure 5.

The above analysis provides a consistent and logical explanation of the observed dominance of 4th and 5th order degradation in case of digital signals. For completeness we have to note that in case of broadband digital signals the higher order degradation terms do not dominate by definition. Conceivably, in case of amplifiers with a very low noise figure or very weak 4th and higher order behaviour, the 2nd and 3rd order degradation can still dominate the 4th and higher order mechanisms.

PERFORMANCE SIMULATION FRAMEWORK

In the above paragraph we have argued that for proper network RF planning, the 4th and 5th order non-linear terms have to be taken into consideration. Thus, a methodology is needed to specify the 2nd, 3rd, 4th and 5th order non-linear behavior. In this section we report our preliminary findings of the ReDeSign studies.

Currently in standardization, the issue of specification of components in case of a digital load is discussed. In these discussions it is proposed to use an all-digital load for the specification measurement. Although this provides a straightforward characterization of the performance, and as such a good figure of merit to compare components, it can be argued that the conventional specification using a load of unmodulated carriers offers some fundamental advantages. These advantages are:

- Even and uneven order non-linear terms are measured separately,
- Because of the narrowband distortion products, a narrowband measurement with a better sensitivity can be used.

Apart from these two specific technical advantages, the general approach of using an advanced HFC performance simulation tool in combination with a full specification of the component non-linear behavior offers the possibility of RF planning in case of cascades with mixed analogue-digital loads.

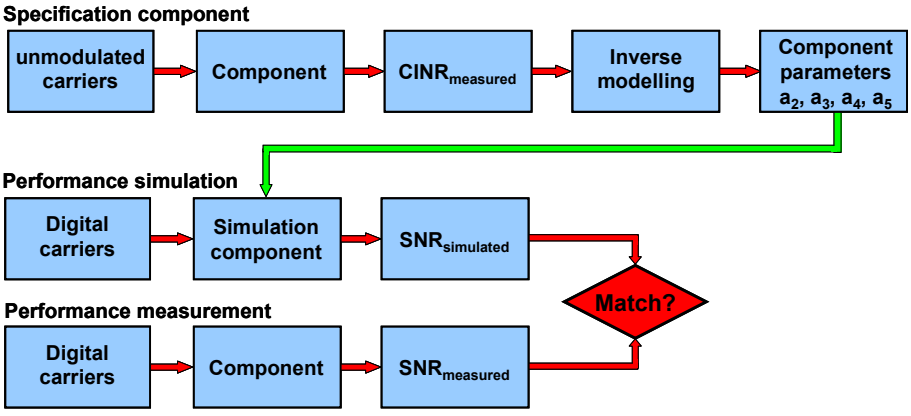


Figure 6. Schematic performance simulation framework for the specific case of a single component.

In Figure 6 we show a performance simulation framework that relies on a component specification measurement using a load of unmodulated carriers. First, the carrier-to-interference noise ratio (CINR) for CSO and CTB cluster beats is measured when applying unmodulated carriers. Using inverse modeling, the coefficients a_1, a_2, a_3 ,a_4 of the component model or A_1, A_2, A_3, A_4 of Eq. 2 are extracted. Next, we can use these component parameters to predict the SNR curve when applying an all-digital load. To conclude, as a validation, the SNR curve for an all-digital load is measured and compared with the simulated SNR curve.

Within the ReDeSign project we have made a first exploration of the feasibility and consistency of the aforementioned approach to model components and to predict the performance. Unfortunately, the result is not conclusive; however, it is encouraging and it provides most interesting insights in the current methods of specification. In the following we will discuss the current status of the studies.

FIRST RESULTS FRAMEWORK

The proposed framework to predict the performance of a component is far from straightforward, and as such susceptible to different errors. Part of the challenge concerned the disentanglement of different error sources. In this process, we discovered several sources for errors that eventually will propagate to the calculated performance value. In short, we found the following critical aspects:

- When applying a load of unmodulated carriers, all even (2^{nd}, 4^{th}, …) and uneven (3^{rd}, 5^{th}, …) intermodulation products contribute to the narrowband cluster beats in a mixed manner,
- The magnitude of the non-linear distortion products strongly depends on the probability density function (PDF) of the composite signal load. A deviation from a Gaussian distribution of either the specification load of unmodulated carriers or the digital load has a large affect on the distortion signal level,
- Measurement artifacts can have a large impact on the measured SNR figure.

In the following we will subsequently substantiate the above points.

Accurate measurement of non-linear terms

In case of a grid of unmodulated carriers with equidistant frequencies $f_n = \delta_f + n\Delta f_c$, all distortion products will coincide at the frequencies $f_n + m\delta_f$, $f_n + (m-1)\delta_f$,… $f_n - m\delta_f$, with n the ranking number of the carrier and m the order of the intermodulation product. Each carrier thus is accompanied from a set of composite intermodulation clusters (cluster beats) as specified in Table 1.

Table 1 Offset frequencies of m^{th} order intermodulation products with reference to f_n.

IM order	Frequency offset (δ_f)										
	$-5\delta_f$	$-4\delta_f$	$-3\delta_f$	$-2\delta_f$	$-\delta f$	0	δ_f	$2\delta_f$	$3\delta_f$	$4\delta_f$	$5\delta_f$
1						x					
2					x		x				
3				x		x		x			
4			x		x		x		x		
5		x		x		x		x		x	
6	x		x		x		x		x		x

In principal, the 4^{th} and 5^{th} order intermodulation products can be measured in an isolated manner at the frequency offsets $\pm 3\delta_f$ and $\pm 4\delta_f$ with respect to the carrier position f_n. In our studies we have verified the technical feasibility of this approach; however, we found that even at a very high carrier level the signal level of the 5^{th} order cluster beat at offset $\pm 4\delta_f$ remains too low for an accurate measurement. Because of the very limited number of beats contributing to the clusters, the signal level of the composite beat is about 20dB beneath the cluster beat at $\pm 2\delta_f$ and below the sensitivity threshold of a

spectrum analyzer. Therefore, we concluded that this approach would not work.

Alternatively, 4th and 5th order terms can be assessed at the frequency offsets 0 and $\pm\delta_f$, albeit that at these offsets the composite contribution of 2nd and 4th (and 6th ..) and of 3rd and 5th (and 7th…) is respectively measured. Next, the 2nd and 4th (and 6th ..) and 3rd and 5th (and 7th …) terms have to be separated properly. This can be accomplished for example by measuring the cluster beat levels for a series of carrier levels.

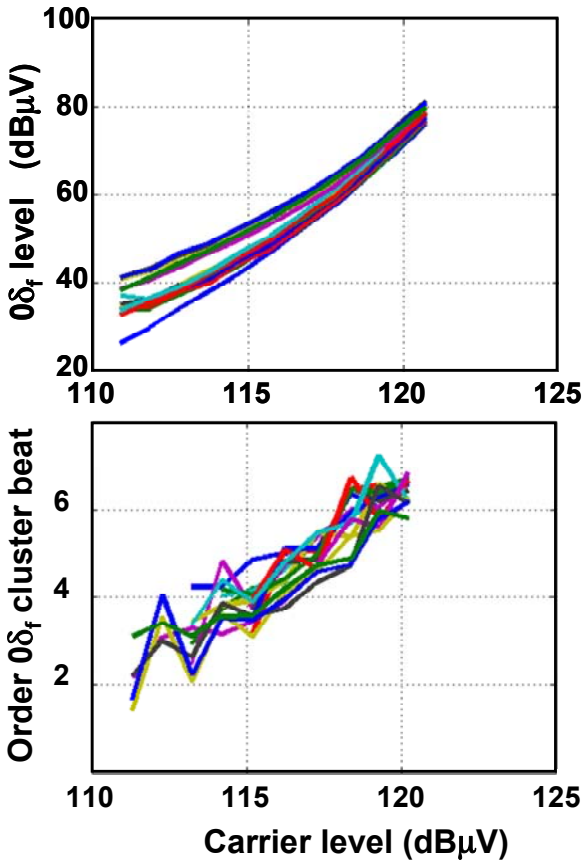


Figure 7 Level (top) and order (bottom) of the composite cluster beat versus the carrier level at frequency offset $0\delta_f$. When switching off the carrier, this cluster bear can be measured accurately. The different curves are recorded for different carrier frequencies f_n.

The top window of Figure 7 shows the signal level of the composite cluster beat with 0Hz frequency offset (the CTB cluster) as measured when temporarily switching of the carrier and for increasing signal level of the unmodulated carriers. The curves with different color refer to the different carrier frequencies f_n of the composite signal. The slope of the curves immediately reflects the effective order of the distortion. The bottom panel shows the effective order of the cluster beat as obtained from the slope. The curves show that the effective order gradually increases from a value 3 toward a value 7. The figure shows that there are no clear level ranges where a single specific order dominates, but that instead for lower carrier levels 3rd and 5th order coexist next to each other and at higher carrier level 5th and 7th. For carrier levels below 111 dBμV, the cluster beat level could not be resolved because of thermal noise floor of the set up.

The same measurements were performed for different devices (Si, GaAs and GaN) and for the cluster beats with $0\delta_f$ (CSO) and $\pm\delta_f$ (CTB) offset. In all cases comparable results were obtained.

This result places the existing method of specification of the amplifiers in a new light; it shows that not the pure 2nd and 3rd order non-linear terms are characterized, the CSO and CTB beats, but a mixture of the 2nd and 4th and of 3rd and 5th order terms. Because of this, these CSO and CTB values can not be used to reliably calculate the narrowband cluster beat levels for composite network loads because they may represent mixed values of 2nd and 4th order non-linear behavior (CSO) and of 3rd and 5th order (CTB).

Signal coherency effect

The proposed framework and the use of Eq. 2 for the calculation of the distortion products both require an input signal with an exact Gaussian-shaped probability density function (PDF). Following the central limit theory, a signal composed of a sufficient number uncorrelated processes will develop

toward a Gaussian distribution. Deviations of a Gaussian distribution will be most pronounced for high signal levels of say 4 – 6 times the average signal level σ. Either the PDF may exhibit a tail associated with a too frequent occurrence of such high signal values or alternatively it may decline too steep as compared to the Gaussian distribution which would infer a too low occurrence of high signal level events. Evidently, because the non-linear behavior only happens at high signal levels, a deviation of the PDF of the input load will have a large impact on the distortion signal level.

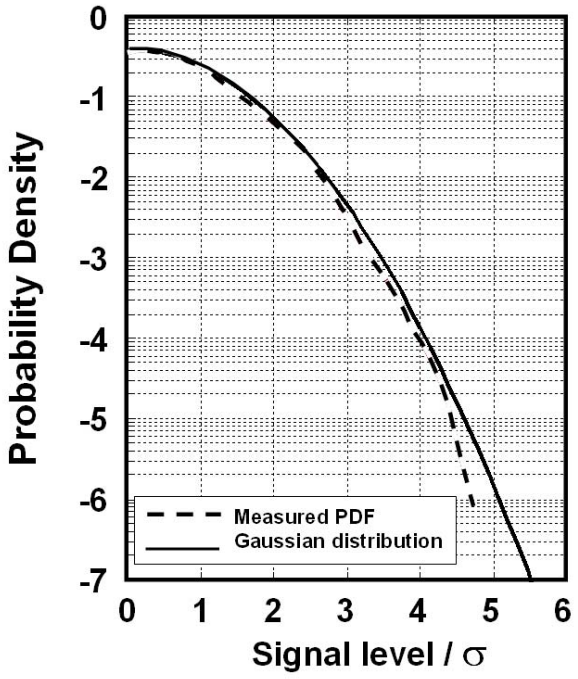


Figure 8 Probability density function of composite test signal of 42 unmodulated carriers measured using a fast memory scope. The 42 unmodulated carriers were generated by commercial test system for a CENELEC CSO/CTB measurement.

To verify the shape of PDF, we have measured signal distribution using a fast memory scope. Both, the PDF of the test load of unmodulated carriers and the real network load have to be verified because a deviation of either both yields an erroneous estimation. In Figure 8, we show the PDF of the specification load of 42 unmodulated carriers generated by a commercial test system as measured with a fast memory scope. The PDF shows a deviation for high signal values; high signal levels occur less frequent then expected for a Gaussian distribution. A second commercial specification system of a different manufacturer was tested as well, with the same result, a too low occurrence of high signal values. Similarly, we have assessed the PDF of the composite signal of 95 digital carriers. Unlike the test loads of unmodulated carriers, this digital signal has a Gaussian shaped PDF.

Next we have made an assessment of the impact of the deviation of the PDF from a Gaussian distribution. In MATHLAB, we have programmed the appropriate algorithms to generate time domain samples of a composite signal of 42 unmodulated carriers (load A) and of 42 digital carriers (load B) with an equal average signal power P_{load}. Correlations were carefully avoided by adding appropriate random phase and frequency offsets and a random phase noise. The PDF were calculated to check whether their shape agrees with a Gaussian distribution. In a second MATHLAB module we have implemented the component model as specified in **Eq. 1** up to the 5th order. Next we simulated the non-linear response for the time domain loads A and B, for different composite signal level P_{load}. The distortion signal was assessed in an 8 MHz channel that contains no unmodulated or digital carrier. Since both signals have an (almost) equal PDF and an equal average signal level P_{load}, the same non-linear distortion signal level was expected. For a low signal level P_{load}, there was a minor difference of 0.5 dB between the distortion signal generated by both loads; however, for a high composite signal level P_{load} a distortion level difference of 2.5 dB was found. This difference shows that the PDF of load A and B are reasonable the same, but not identical. Stated differently, we

managed to program reasonable but not perfect uncorrelated composite loads A and B.

In a following simulation, we have assessed the impact of a too low occurrence of high signal values. The PDF of load A was artificially deformed by numerical clipping of the signal at a level of 5 or 6 times P_{load}. By filtering, all artifacts related to the clipping were removed from the 8 MHz channel used to assess the distortion signal. This clipped signal was subsequently applied to the component and again the non-linear distortion signal was assessed for different signal levels P_{load}. The result of this simulation with reference to load B is shown in Figure 9. Apparently, clipping the signal of load A results in a 9-12 dB lower distortion signal level as compared to load B. This result demonstrates that (small) deviations of the PDF may cause very large deviations of the distortion signal level. Clearly, to ensure a good specification, the test load must be composed of sufficiently uncorrelated carriers. This should be verified by measurement of the PDF.

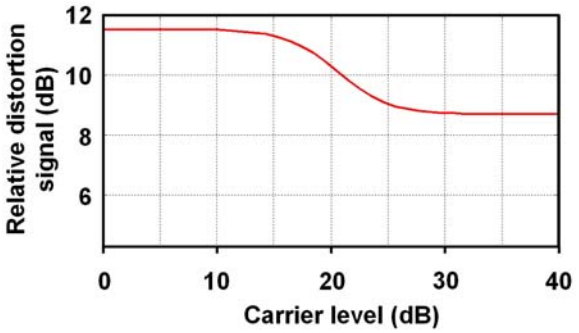


Figure 9 Distortion signal generated by a composite load of 42 unmodulated carriers clipped at 5 P_{load} with reference to the distortion signal generated by 42 digital carriers with an equal composite signal level.

SNR Measurement

To complete the studies, we have measured the SNR curves of several amplifiers when applying a digital load of 95 digital carriers. To avoid a signal overload of the spectrum analyzer, a band pass filter was placed in between of the amplifier output port and the spectrum analyzer. Later we learned that the impedance mismatch of the band pass filter could generate a harmful signal reflection and that an attenuator should be inserted between the output port of the amplifier and the band pass filter. Figure 10 gives a typical result of our studies of an SNR curve for an amplifier with a digital load, for a measurement set up with and without an additional attenuator. The result shows a clear degradation of the SNR curve in case of the set up without the attenuator.

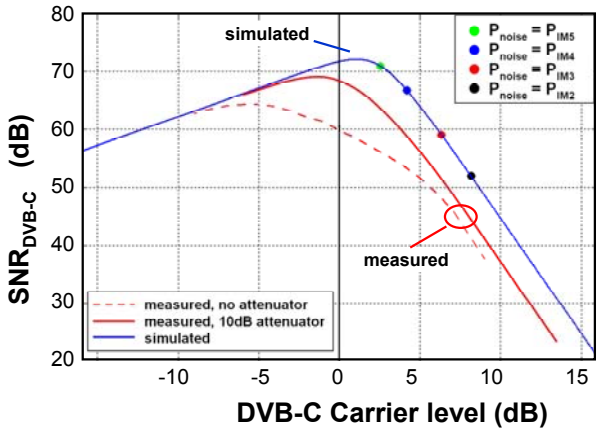


Figure 10 Measured and simulated SNR curves of an amplifier with a load of 95 digital carriers.

Match measured and simulated SNR curve

The challenge of this study is finding a good match between the measured and simulated SNR curve of a component. In the above part we have summarized a number of issues that we have encountered: accurate assessment of all component parameters (a_2, a_3, a_4, and a_5), the probability density function of the test load, and the measurement set up for the performance measurement itself. Currently, we are studying different algorithms to extract the component parameters from a multi-level measurement with a load of unmodulated carriers as shown in Figure 7. Although these studies are not conclusive, a first attempt is made to simulate

the SNR for the component of Figure 10, using component parameters a_2, a_3, a_4, and a_5 extracted from a multi-level specification measurement of Figure 7. Unfortunately, the composite test load of 42 unmodulated carriers had a PDF with an under representation of high signal levels as shown in Figure 8. Because of this, the component parameters a_1, a_2, a_3 and a_4 are all underestimated. Figure 10 provides a comparison of the measured and simulated SNR curves, showing a mismatch of about 2 dB. The simulated curve is shifted to an about 2 dB too high carrier level, which agrees with an 8 dB underestimation of the distortion signal level during the specification. As shown in Figure 9, such an error can be attributed to remnant correlation between the unmodulated carriers of the specification load.

In addition, we have indicated the points of equal n^{th} order intermodulation level and the amplifier thermal noise, for example $P_{noise} = P_{IM5}$. These points confirm that in this particular amplifier, the 2^{nd}, 3^{rd} and 4^{th} order terms can be all neglected, in agreement with the explanation given in Figure 5.

SUMMARY AND CONCLUSION

In this contribution we have demonstrated that in case of digital carriers, the existing $2^{nd}/3^{rd}$ order component model does not provide an adequate quantitative description of the performance degradation due to the non-linear character of amplifiers. Instead 4^{th} and 5^{th} order non-linear terms have to be taken into account.

We have studied the specification of the component parameters a_2, a_3, a_4, and a_5 of a component using a load of unmodulated carriers and measuring the distortion signals for different frequencies over a range of carrier levels. From such a measurement, the model parameters a_2, a_3, a_4, and a_5 can be extracted. In a separate measurement we assessed the SNR curve of the same component for a digital load of 95 DVB-C carriers. Simulation of the SNR curve for this load using a 5^{th} order component model resulted in an 8 dB underestimation of the negative asymptote for high carrier levels, or a 2 dB overestimation of the digital carrier level itself. Likely, this error is associated with some remnant coherency between the unmodulated carriers of the specification load.

Although the match between the measured and simulated performance of an amplifier is not yet exact, the study indicates that the mismatch is caused by the specification of the component. Provided that this problem is appropriately addressed, the simulation of the performance will provide a power full approach for the capacity optimization of the coaxial cascades of HFC networks.

ACKNOWLEDGEMENT

This study was made possible through the funding of DG Information Society of the European Commission.

---

[1] www.ict-redesign.eu

[2] www.ict-redesign.eu, see for example Deliverable 4: "*HFC Channel Model*", Deliverable 10: "*Methodology for Specifying HFC networks and components*" and Deliverable 14: "*A new frequency plan and power deployment rules*"

[3] *Noise loading analysis of a memory-less nonlinearity characterized by a Taylor series of finite order,* Yen-Long Kuo, IEEE Transactions on Instrumentation and Measurement, Vol. IM-22, No. 3, September 1973

[4] IEC 60728-3 Cable networks for television signals, sound signals and interactive services, part 3 Active coaxial wideband distribution equipment.

[5] *UTOPIC, a new RF planning tool for cable networks*, Jeroen Boschma, Broadband, Vol. 30, No 3, December 2008. This paper can be downloaded from www.tno.nl/utopic

NEXT GENERATION - GIGABIT COAXIAL ACCESS NETWORK

Michael Emmendorfer—Sr. Dir., Solution Architecture and Strategy Office of the CSO
Mike.Emmendorfer@arrisi.com
Tom Cloonan, Ph.D.—Chief Strategy Officer
Tom.Cloonan@arrisi.com
Scott Shupe—Chief Technologist Office of CTO
scott.shupe@arrisi.com
Zoran Maricevic, Ph.D.—Sr. Dir., Solution Architecture and Strategy Access and Transport
zoran.maricevic@arrisi.com)

ARRIS

Abstract

The media and telecommunications industry is entering a decade of rapid change. The change will be driven from consumers and competition. The consumers of the next decade will likely be those whom have a desire to have any content made available anytime, anywhere and to any device. The programmers and telecommunication providers are planning how to meet this challenge. Consumers are not just recipients of content they have increasingly become creators and/or distributors of content. We have seen in the last decade the use of peer-to-peer (P2P) and the sudden increase of YouTube and social networking, this has driven how telecommunication providers, like cable operators have become not only content distributors to the home but also increasing "from" the home. A key challenge the cable industry may face in the future is the transition from a largely broadcast service delivery network to a rapidly growing unicast delivery network. This paper is a forward looking study which will examine alternative architectures for the cable industry to address new competitive threats posed by fiber to the premise (FTTP) providers. The paper will examine the business and bandwidth drivers of the coming decade and predict the transition of the cable delivery network to accommodate the future of more unicast video and data services. As the cable industry examines next generation access architectures such as RFoG and EPON for new build residential and commercial deployments, this paper will focus entirely on leveraging their most valuable network asset, the existing coaxial network to the home. This study examines strategies to meet the demands utilizing IP based network technology to and "from" the home. This Next Generation – Gigabit Coaxial Access Network may eventually be capable of delivering multi-gigabit IP services to the home while defining architectures to enable 1 Gbps from the home, all while leveraging the coaxial network to the home.

INTRODUCTION

The study of the Next Generation – Gigabit Coaxial Access Network or simply Gigabit Coax Network (GCN) is an attempt to examine the drivers and possibilities of the coming decade and how the cable network may evolve to support a multi-gigabit IP network to the home while defining architectures to enable 1 Gbps from the home, all while leveraging the coaxial network to the home. The paper is an initial assessment meant to inform members of the cable technical community of some of the network migration drivers and options. This paper is also intended to spark research and debate within the industry surrounding requirements and possible solutions. The paper will document the transition drivers and trends as

well as predictions of how the cable delivery network may evolve to support higher IP based services to and from the home with an emphasis to consider an architecture which may support 1 Gbps symmetrical services. The cable industry seems to be entering a period of unprecedented transition. This transition will have two key threads, 1) the transition to a unicast service delivery platform (service personalization) and 2) the increased spectrum and bandwidth allocation for IP as the delivery technology. The future is always difficult to predict and especially in the area of technology. The demands of the end consumers and the value of the service will also remain a challenge to forecast. We can examine trends from the past which may help shape and guide our predictions of the future; this paper will examine these business drivers. The cable network is incredibly flexible allowing the operators to select migration paths to expand capacity where and when needed, the current approaches are reviewed in this document. The paper examines a key area called upstream augmentation which will serve as the building block for the cable network to expand capacity in the upstream.

BUSINESS DRIVERS

It may be hard to imagine another time in cable's history where the competitive landscape was so fierce. The competition has expanded from DBS providers to Telecos offering triple play services to recently competitive threats from Over The Top (OTT) providers such as Hulu and Netflix. It is important to understand some of the key business drivers which we may face in the future as these will serve as guide for network planning. These drivers as stated above will come from consumers and competitive threats. It remains uncertain which path the MSO may take to meet the consumer's insatiable demand for unicast and personalization of video content or to the degree they will address the competitive threats. We know that the usage of the high speed data network and the allocation of spectrum to support this service will continue to increase. The future delivery of video services may evolve as operators examine the viability of using IP based network technology for the distribution of their video services to consumers. We need to examine these business drivers listed below and others to guide the strategy and planning for the future of the cable network. The cable network end to end is well positioned to meet the demands of the future and leveraging the existing coax to support the transition to greater IP based bandwidth to and from the home is a compelling advantage the cable industry has over alternative technologies.

Internet Bandwidth Trends

As illustrated in the figures below the consumption curve may have began with email, web browsing and newspaper like illustration on user PCs. This evolved to a magazine experience and use of short video clips. Perhaps midway through the decade the network was used for digital music distribution, gaming and P2P. We may have ended the decade with what may become the biggest drivers of internet growth, OTT providers of video and social networking. This next era will see an increase in the usage of full movie downloads from the internet to the home; this will increase the downstream bandwidth as was as upstream bandwidth. Figures 1 and 2 provide illustrations of the Internet bandwidth trends.

In addition to consumption rising as illustrated in the figures above, a key contributor to overall bandwidth drivers is the maximum service tier offered to consumers. The figure below shows a 25 year history of the max bandwidth offered or available to consumers. This figure also attempts to predict the max service tier we may see in the future if the growth trend aligns with the

preceding 25 years. The maximum service tier plays a significant role in the application developed. It would be fair to assume that some of the most recent application such as over the top video was a result of increase in the higher data speeds offered to consumers.[1]

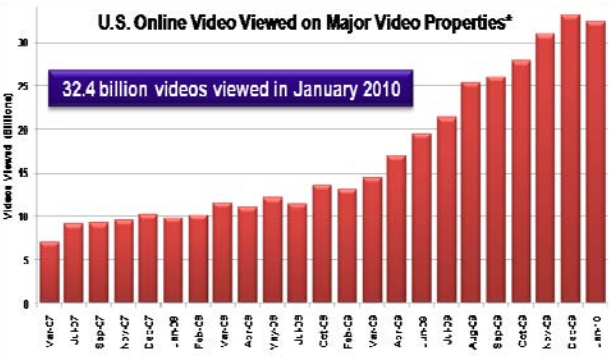


Figure 1— On-line Monthly Video Viewing
Sources: comScore and Internal ARRIS Research

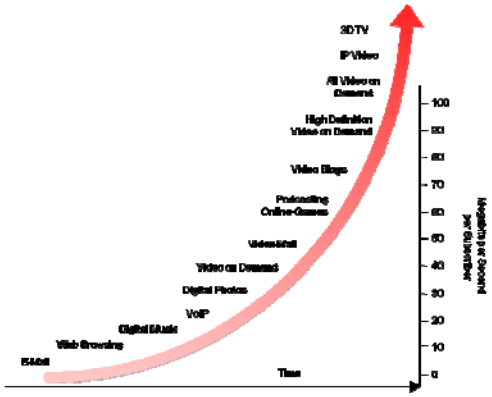


Figure 2—Megabits per Second per Subscriber

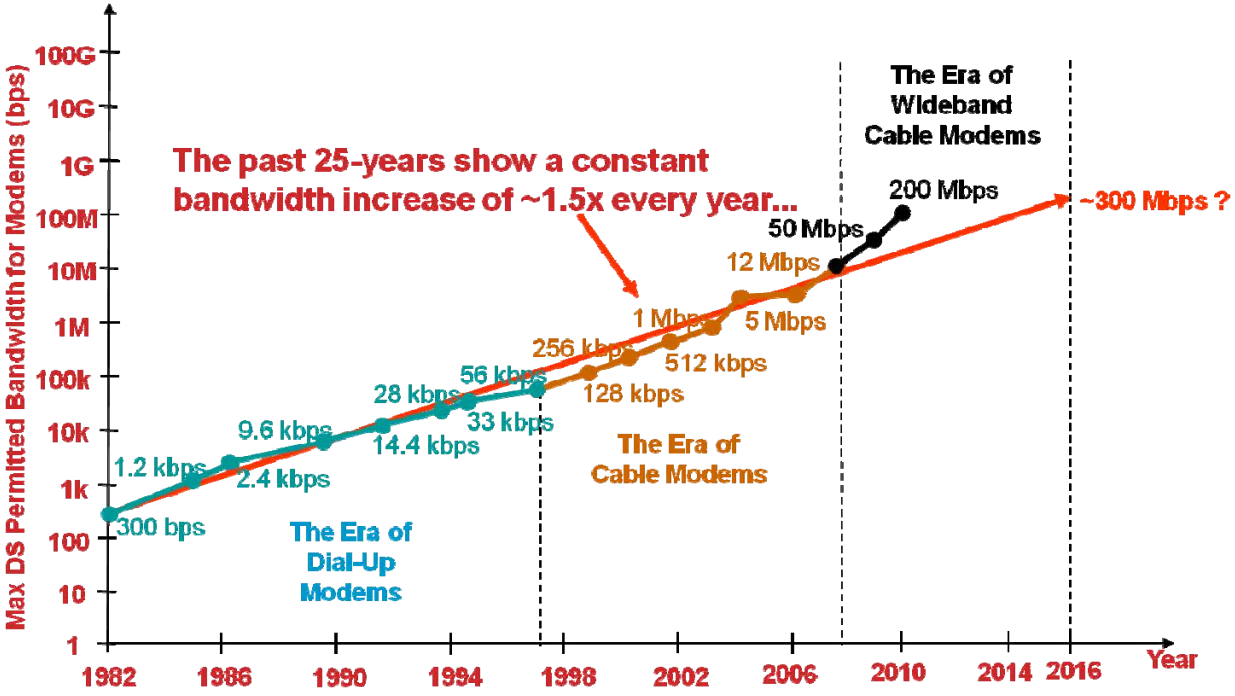


Figure 3—Trends & Predictions of Maximum Offered Modem Bandwidths

Downstream Spectrum

The expansion of downstream spectrum over the last 60 years provides cable a high capacity network to the home. The figure below is an illustration of cable's investment in expanding the downstream network capabilities to meet the service demands of the consumer, where today some systems may have as much as 6 Gbps (assuming a 1 GHz system) of forward capacity. This investment will help the MSO transition from a largely broadcast provider to a high bit rate unicast provider to the home.

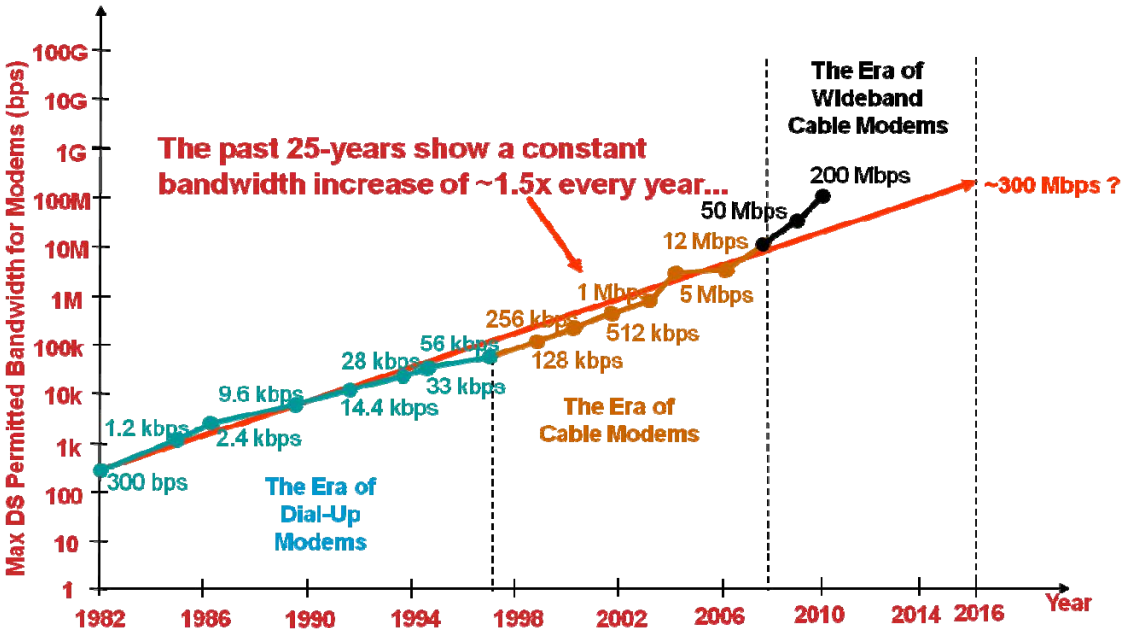


Figure 4—Cable Network Downstream Spectrum Expansion

Downstream Bandwidth Predictions "A Network in Transition"

The amount of spectrum allocated for services will evolve from an entirely broadcast allocation providing analog TV and digital TV distribution network to a more unicast network supporting services like high speed date, telephone, and video on demand. To efficiently offer these unicast services MSOs use a technique called narrowcast whereby the same spectrum allocation may be reused throughout a system because the distribution of these signals is targeted to a predefined service area. The use of narrowcast services allows the MSO to reuse spectrum in a given market. The figures below illustrate the spectrum allocation in the year 2010 and an estimate for year 2015 for the broadcast and narrowcast/unicast services and the number of 6 MHz channels which may be allocated. These figures clearly illustrate the transition the MSOs may make to support the need for greater and greater narrowcast/unicast service. In perhaps as many as five years the MSO transformation may moved from a largely broadcast content distributor to largely a unicast content distributor. The transition of the downstream bandwidth enables the MSO to offer consumers far more personalized services.

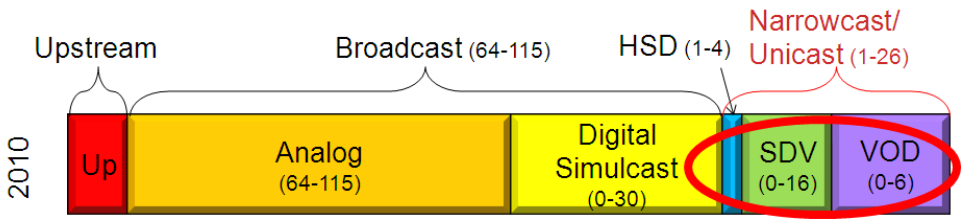


Figure 5

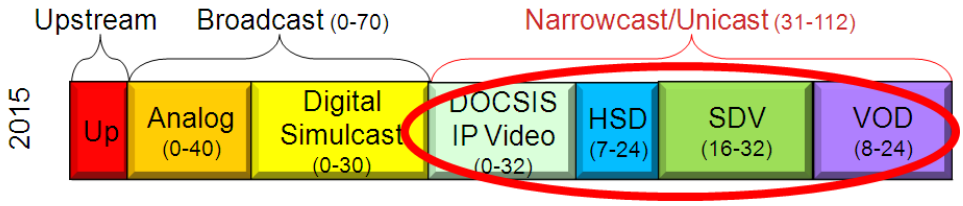


Figure 6—Downstream Bandwidth Predictions

Figure 7 illustrates the spectrum allocation projected in 2010 and 2015 for the combined HSD and DOCSIS IP Video services in the form of megabits per second (mbps). The graph captures the low and high estimates for each year. The high-end allocation of spectrum and bandwidth suggests that cable may allocate over 2 Gbps of downstream capacity for IP based services and technology in the coming decade.

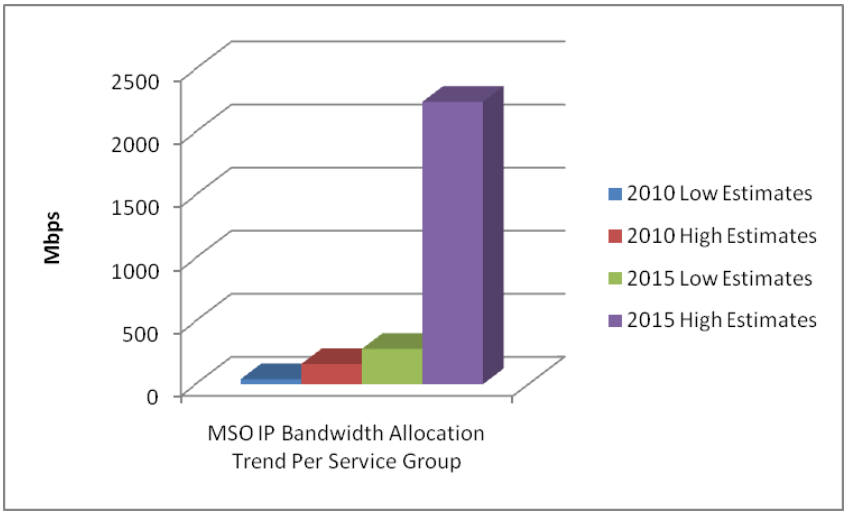


Figure 7—IP Based Bandwidth Supporting HSD and DOCSIS IP Video Predictions

The diagram below suggests the greatest network transition in the MSOs history from a broadcast content delivery network to the home to nearly an entire unicast content delivery network. This transition which began many years ago with the introduction of HSD and VoD services enables the cable operator to continue to expand their service offering incrementally and through targeted capital investment where and when needed. The consumer's ability to obtain any content, anytime, from anywhere and to any device will be fulfilled during this transition. MSOs are well positioned to enable these future services and capabilities. This figure captures the possibilities an MSO may consider for the

next decade and the allocation of their spectrum for more and more unicast services.

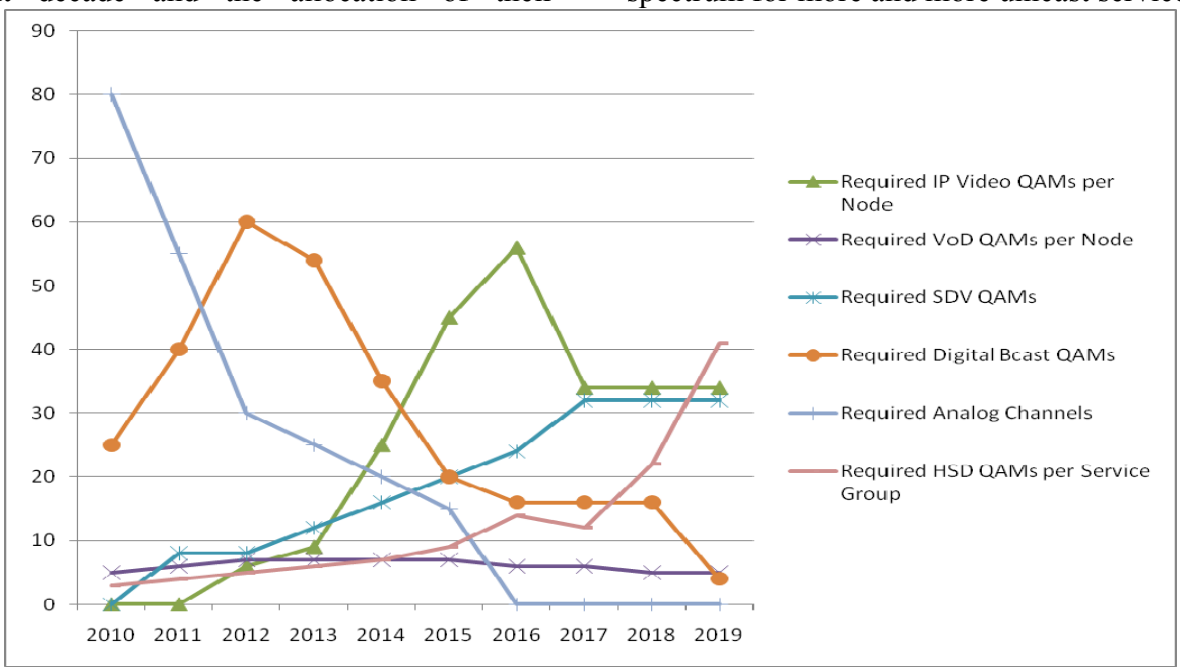


Figure 8—A Network in Transition

Upstream Spectrum

The upstream may also see a transition in the coming decade to accommodate the increase in consumer network usage and the transition of the downstream to more IP/DOCSIS channels. As seen in figure 4 the downstream spectrum allocation has increased steadily over the last 60 years. The upstream, however, has largely been untouched. This is primarily because the current 5-42 MHz spectrum allocation (U.S.) remains lightly loaded and has much as 150 Mbps of capacity.

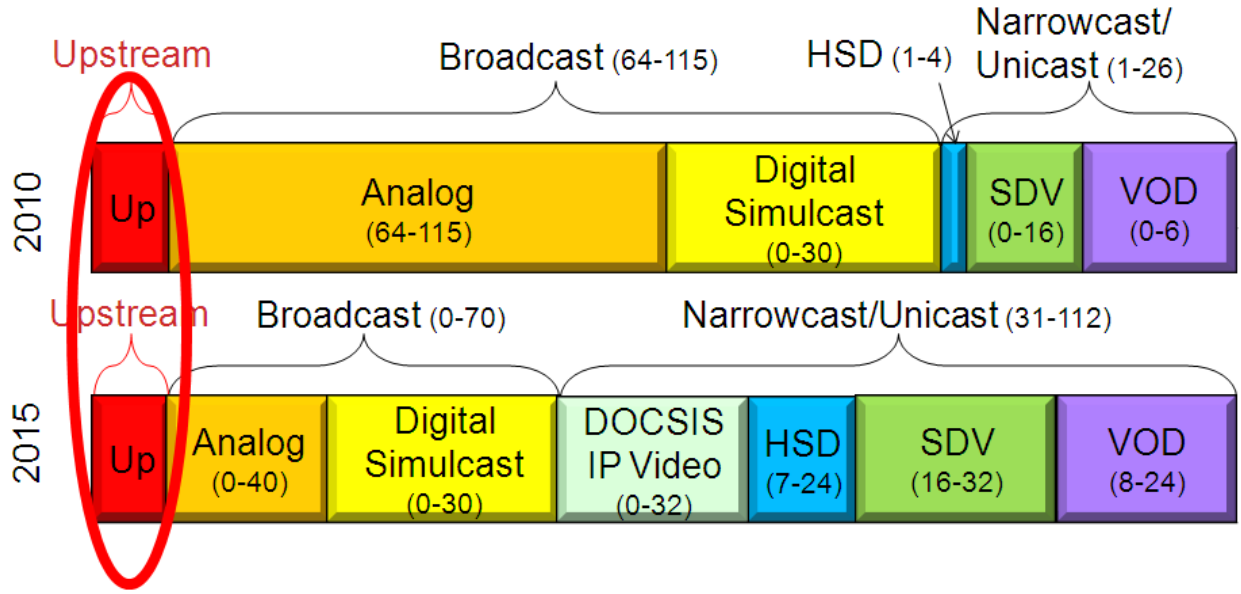


Figure 9—Upstream Spectrum Allocation Remains Unchanged

Upstream Bandwidth Predictions

In the beginning of this section we examined the Internet growth trends and the downstream bandwidth predictions to include the expansion of more IP/DOCSIS bandwidth in the coming decade and the growth of the max service tier offered. The downstream bandwidth usage has risen over time and the upstream continues to increase as well. The upstream traffic load may be represented by a value of about 25% of the downstream HSD traffic load during busy hour. The upstream may grow in percentage terms faster than the downstream in this coming decade as a result of consumer adoption of user generated content such as YouTube, Social Networking, P2P, and Video conferencing. The increased consumption of upstream traffic will also be attributed to an increase in downstream bandwidth allocated to IP/DOCSIS. The reasoning behind the increase in the upstream as a result in the increase in downstream traffic is derived by the transmission technology used, Transmission Control Protocol (TCP) and the acknowledgment packets which are transmitted upstream to the content distribution server. As the downstream expands the increase in the upstream traffic load will increase perhaps as much as 50% CAGR for the next decade. In the figure below the upstream traffic load is examined assuming an expanded downstream traffic allocation for HSD and IPTV and the continued expansion of upstream bandwidth as a result of consumer and application behavior.

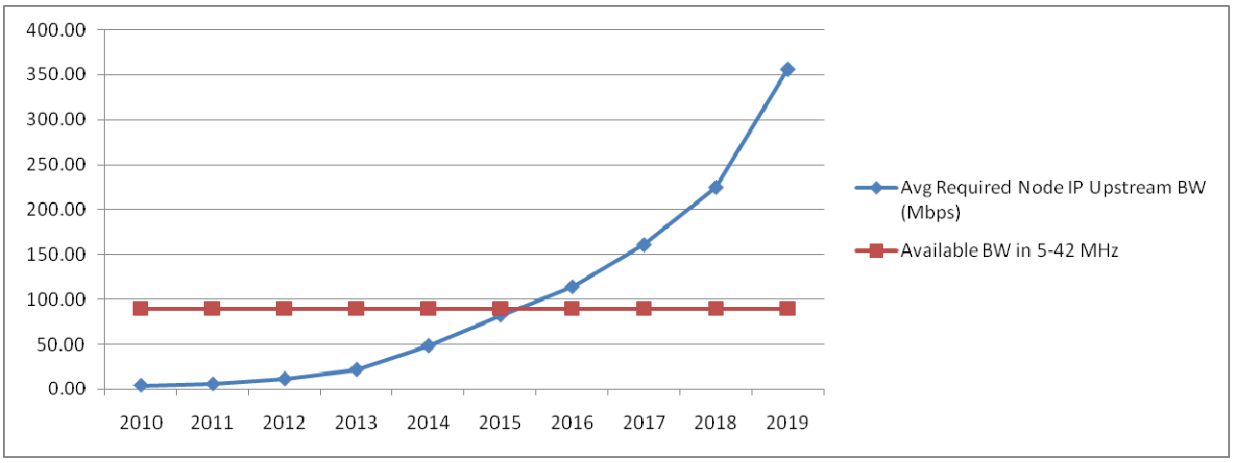


Figure 10—Upstream Bandwidth Predictions

The diagram above assumes a 500 HHP node and the allocation of upstream capacity is 90 mbps. In this diagram the upstream for the node may be exhausted by the year 2015, however, the HFC is remarkably nimble and through targeted investment additional capacity may be made available as described in the diagram below. The use of node segmentation allows the MSO to partition a node and perhaps only the nodes affected to increase capacity by decreasing the HHP served in the upstream. The diagram also suggests that upstream augmentation may be needed if an upstream service offering exceeds the available throughput of the upstream. The diagram below suggests that

investment in upstream capacity in not needed for some time but will likely occur as part of routine traffic planning throughout the decade.

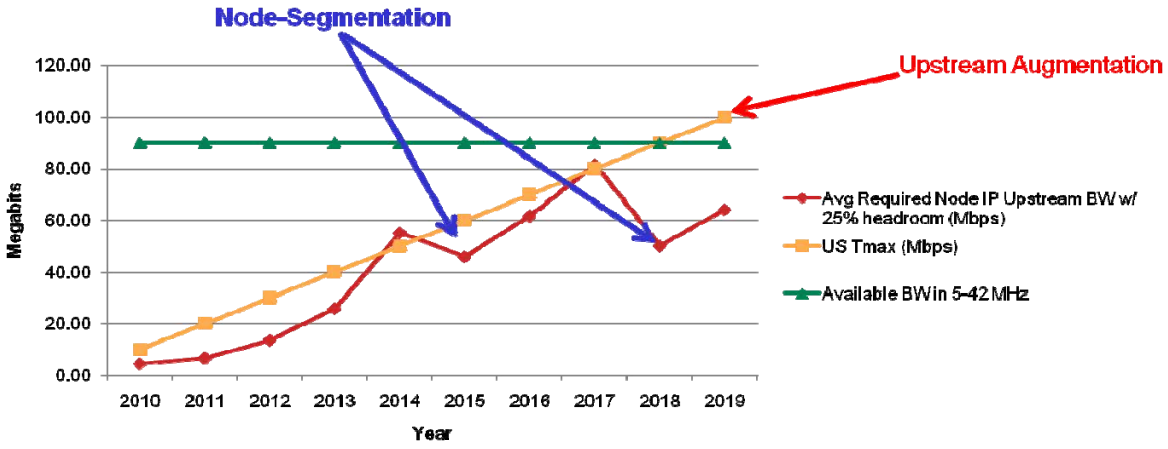


Figure 11—Upstream Prediction Addressed using Node Segmentation and Upstream Augmentation

Competitive Threats

In the beginning of this section we cited some key business drivers for the transition of the network may be driven by competition. The diagram below illustrates the various network technologies available or which may emerge in the coming decade that will compete against cable. In this diagram the technologies maximum bandwidth to and from the home are examined. This illustrates that when compared to alternative technologies cable's massive bandwidth to the home is a key differentiator. Cable has the ability as illustrated in the sections above to allocate more and more of the spectrum bandwidth to IP based technologies which will enable continued evolution of the service offerings as well as take advantage of the efficiencies found with IP network technology.

The upstream bandwidth allocation is very competitive with alternative technologies with the exception of GPON. This paper will examine migration strategies for the upstream to compete with the capacity found using PON technology.

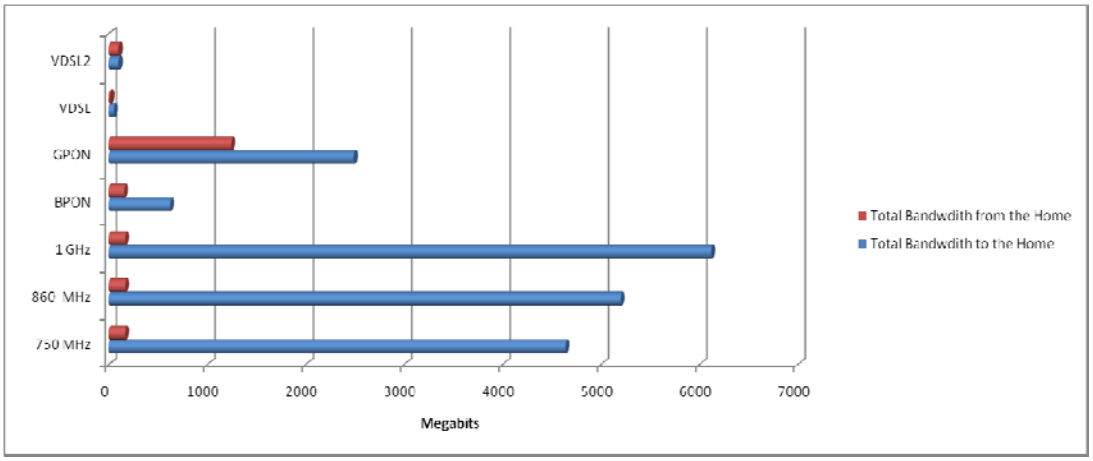


Figure 12—Comparison of Maximum Bandwidth Available To and From the Home

Business Driver Summaries

There are many factors cable operators are considering as they begin planning the network evolution over the next decade. The changes in technology and service expectations of the consumer coupled with competitive threats will redefine the cable network. The following sections will begin to address how cable will respond to these business drivers and document some of the migration options to remain competitive.

CABLE'S CAPACITY EXPANSION METHODS

The modern cable network is incredibly flexible allowing the MSO to make targeted investments where and when needed to either incrementally or in some cases substantially increase network capacity depending on the capacity expansion method selected. The use of capacity expansion methods may be applied across an entire network footprint or with laser beam focus to address capacity challenges. The table below is an attempt to capture the various methods available to increase or improve capacity of the network. The diagram brings together methods and techniques used by various disciplines within the MSO, such as outside/inside plant, IP/Data, SDV, and Video Processing. The techniques will allow the MSO to transform their network from broadcast to unicast and from analog/digital to IP.

Today, in fact MSOs may use techniques to increase capacity without touching the outside plant; this is dramatically different than the approaches that were used for decades. The technique referred to as Bandwidth Reclamation and Efficiencies, as illustrated in the top of figure 13 is becoming the primary method to address system wide capacity challenges. In most cases this technique may be implemented with equipment in the headend and home, thus not requiring conditioning of the outside plant or headend optics. A technique recently put into practice by some cable operators is partial or even full analog reclamation, this enables the operator to transition the channels currently transmitted in analog and to transmit them only in digital format allowing greater bandwidth efficiencies by requiring the use of a digital terminal adapter (DTA) alongside televisions that may have only had analog services. Another technique for Bandwidth Reclamation and Efficiencies is the use of Switch Digital Video (SDV). The use of SDV allows the cable operator to transmit in the network only the video streams that are being viewed by consumers. This allows the operator to increase the number of channels offered to consumers, in fact the actual channels offered to the consumers may exceed the throughput capabilities of the network but through careful traffic engineering and capacity planning this approach is an excellent way of adding additional capacity to the network. This technique is a form of over subscription and has been in practice for decades by the telecommunication industry. The items captured in Bandwidth Reclamation and Efficiencies are the modern methods to expand capacity. In many respects the Bandwidth Expansion "upgrade" approach as illustrated in figure 4 whereby the entire network was upgrade to increase capacity may be seldom used in the future. If used this may be part of a joint plan to increase the spectrum allocation of the return path.

In the future, the use of IP for video delivery will provide even greater bandwidth efficiencies IP used for digital video transmission and will also provide functionality similar to the techniques used in SDV. Another key advantage is that IP allows for the use of variable bitrate (VBR) encoding increasing the capacity of the network [2].

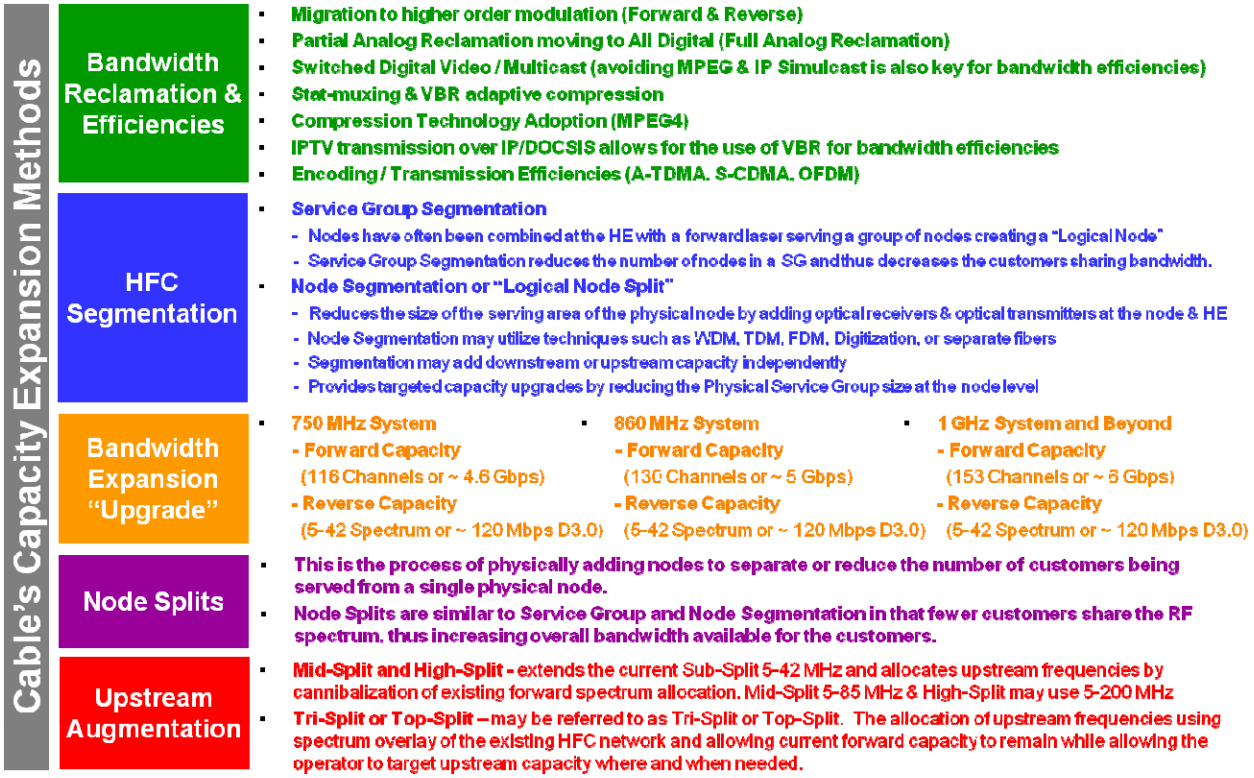


Figure 13—Cable's Capacity Expansion Methods

Summaries of Capacity Expansion Methods

Cable operator's selection priority of the capacity expansion methods has and will continue to vary. The MSOs will eventually use all or near all of the Capacity Expansion Methods in the table above.

Downstream Capacity Expansion

- DTA's & SDV will provide long term downstream plant capacity expansion
- Reduced Service Group Size enabling fewer customers to share bandwidth
- Node Segmentation and Node Splits will continue to be used in a targeted basis
- Possible downstream bandwidth expansion along with upstream augmentation

Upstream Capacity Expansion

- Use of highest order modulation and Channel Bonding to increase throughput [3]
- Progressively smaller upstream service groups
- Ongoing node splits / segmentation
- These incremental steps should last for a majority of the decade
- Upstream Augmentation expands upstream spectrum and bandwidth such as conversion to mid-split, high-split, or tri-split (as described in detail in the section below)

UPSTREAM AUGMENTATION ANALYSIS

The application of Upstream Augmentation is not likely to occur in the near future. The current spectrum allocation should be sufficient considering the bandwidth predictions as cited earlier

in this paper. This paper is again a forward looking study meant to increase awareness of the upstream augmentation options. The hybrid fiber coax (HFC) network has a lot of legs left in both the downstream and upstream direction. This paper considers a future architecture capable of delivering multi-gigabit IP services to the home and perhaps 1 Gbps from the home, all while leveraging the coaxial network to the home. This is referred to as the Next Generation – Gigabit Coaxial Access Network or simply Gigabit Coaxial Network (GCN). Some cable operators are already planning for an expansion of IP based traffic in the downstream and this may reach multi-gigabit speeds within this decade. The cable network has the ability to transition all of the capacity to IP if desired.

Overview

The use of upstream augmentation will have many trade-offs for network planners to consider. This diagram captures the upstream and downstream allocation of bandwidth predicted by 2015. This diagram is meant to illustrate the placement of the upstream spectrum given three (3) Upstream Augmentation Options:

- Mid-Split
- High-Split
- Tri-Split

This use of Mid-Split or High-Split consumes existing downstream capacity. The use of High-Split may compress the year 2015 channel allocation forecast assuming a 750 Mhz system. The use of Tri-Split allows the existing forward capacity to remain as well as the forecasted utilization because additional upstream spectrum is allocated above the current downstream spectrum. The following sections examine all three approaches as well as others.

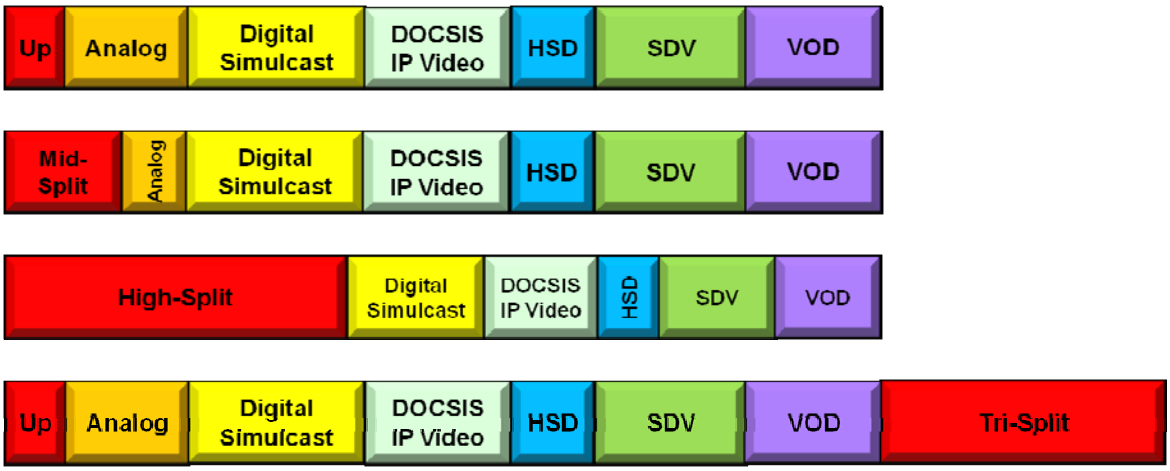


Figure 14—Upstream Augmentation Comparison

Mid-split 5-85 MHz Analysis

The Mid-Split Architecture is defined as 5-85 MHz with the downstream starting at approximately 105-108 MHz; this may also be referred to as the 85/105 split. The mid-split has been discussed for many years; in fact the DOCSIS 3.0 [4] [5] specifications included support for mid-split. The inside and outside plant network element may have support for mid-split however this depends on the year of the deployment, manufacturer used, and type of network element. The mid-split architecture essentially doubles the current upstream spectrum allocation. The tables below capture some the advantages and disadvantages when considering Mid-Split

Mid-Split Advantages	
Area	**Comment**
Bandwidth	Upstream moves to nearly 315 Mbps +
Spectrum Upstream Allocation	5-85 Upstream
Headend Optical Transmitter	Existing Equipment should support
Headend Optical Receivers	Existing analog receivers should support up 200 MHz
Nodes (Optical Side)	DFB 200 MHz Tx should support
Node (RF Side)	Best Case: replace the Diplexer filters with a pluggable filter swap
Amplifiers	Best Case: If pluggable replace the Diplexer filters
Passives	Mid-split leverages existing Passives
DOCSIS 3.0 CMTS and CM/EMTAs	Recent DOCSIS 3.0 products (CMTS and EMTAs) are built to use mid-split spectrum this maybe leveraged for full High-Split

Table 1—Mid-Split Advantages

Mid-Split Disadvantages	
Area	**Comment**
Bandwidth	Does not support PON like speeds and perhaps limited to 315 Mbps + (not 1 Gbps) We will assume: 10-85 MHz is useable 2 MHz set aside for Legacy STBs 2 MHz set aside for Legacy Status Monitoring 3.2 MHz for Legacy DOCSIS Traffic Leaves about 67.8 MHz of usable capacity for DOCSIS 3.0 or Ten 6.4 channels at 30 mbps and One 3.2 channel at 15 mbps Usable DOCSIS bandwidth perhaps 315 Mbps +
Spectrum Guard band (between US/DS)	Guard band 85-105 or 85-108 (about 20 MHz)
Impact to Exiting Forward Capacity	Reduced by about 50 MHz
Spectrum Interference Concerns	Assume 10-85 MHz is useable
Headend Optical Receivers	Digital Receivers would have to be replaced
Nodes (Optical Side)	FP 200 MHz Tx will need to be replaced 42 MHz Digital Return will have to be replaced
Node (RF Side)	Worst Case: Replace the housing because there is no pluggable filter or amp that fits into existing housing
Amplifiers	Worst Case: Replace the housing because there is no pluggable filter or amp that fits into existing housing
House Amplifiers	Mid-split will require change of a home amp
OOB Set-Top Box Communications	Some STBs may be hard coded within the mid-split range (75.5 and 104.25 MHz)

Table 2—Mid-Split Disadvantages

High-split 5-200 MHz Analysis

The High-Split Architecture is defined as 5-200 MHz with the downstream starting at approximately 250-258 MHz; this may also be referred to as the 200/250 split. The High-split is being considered because full or partial analog reclamation is underway or planned by cable operators. This will allow a smoother transition when considering consumption of existing analog spectrum. As with mid-split DOCSIS 3.0 specifications systems may be used, however to take advantage of the full spectrum additional development is required. The tables below capture some advantages

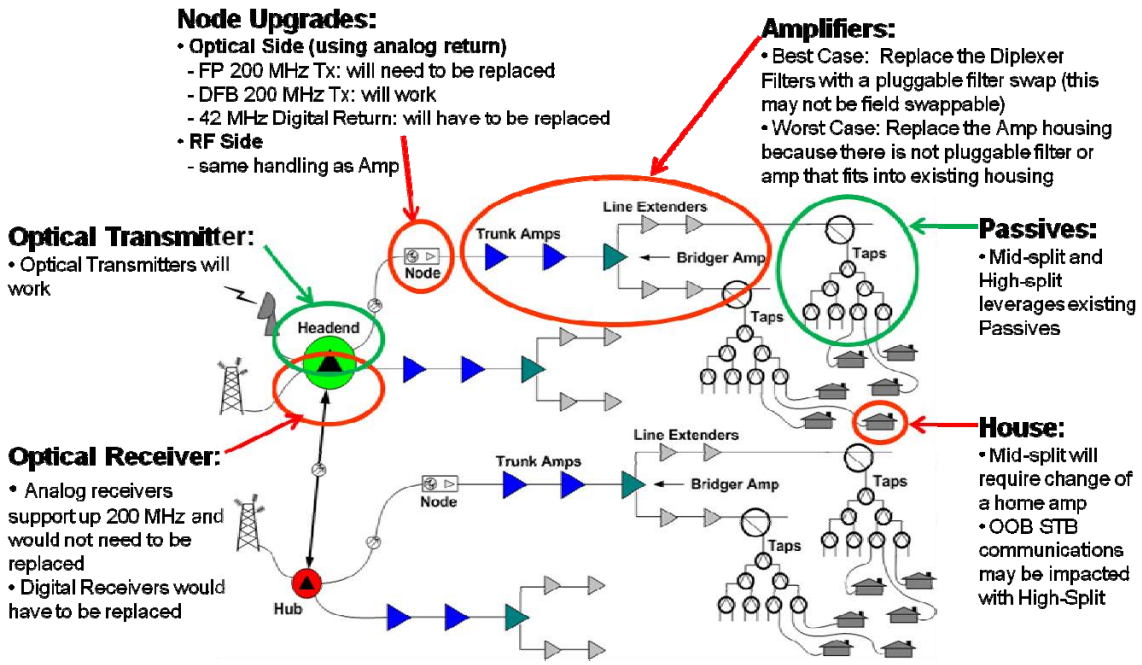
and disadvantages when considering High- Split.

Area	Comment
Bandwidth	Upstream moves to nearly 855 Mbps + with existing DOCSIS technology
Spectrum Upstream Allocation	5-200 Upstream
Headend Optical Transmitter	Existing Equipment should support
Headend Optical Receivers	Existing analog receivers should support up 200 MHz
Nodes (Optical Side)	DFB 200 MHz Tx should support
Node (RF Side)	Best Case: If pluggable is supported replace the Diplexer filters with a pluggable filter swap
Amplifiers	Best Case: If pluggable replace the Diplexer filters
Passives	High-split leverages existing Passives
DOCSIS 3.0 CMTS and CM/EMTAs	Recent DOCSIS 3.0 products (CMTS and EMTAs) are built to use portion of High-split spectrum

Table 3—High-Split Advantages

High-Split Disadvantages	
Area	**Comment**
Bandwidth	• Does not support PON like speeds and perhaps limited to 855 Mbps + (not 1 Gbps) We will assume: 10-200 MHz is useable 2 MHz set aside for Legacy STBs 2 MHz set aside for Legacy Status Monitoring 3.2 MHz for Legacy DOCSIS Traffic Leaves about 182.8 MHz of usable capacity for DOCSIS 3.0 or (28 Channel of 6.4 channels at 30 mbps) and One channel at 3.2 channel at 15 mbps) Usable DOCSIS bandwidth perhaps 855 Mbps Assumes no other interference • 1 Gbps Speeds maybe achieved with changes to DOCSIS 3.0 thus new industry investment in new PHY encoding, MAC/PHY layer technology and legacy investment may be stranded.
Spectrum Guard band (between US/DS)	Guard band 200-258 (58 MHz)
Impact to Exiting Forward Capacity	Reduced by about 200 MHz
Spectrum Interference Concerns	FM Radio Band, DTV and Aeronautical frequencies - avoidances of these bands reduces the overall spectrum bandwidth available for data services
Headend Optical Receivers	Digital Receivers would have to be replaced
Nodes (Optical Side)	• FP 200 MHz Tx will need to be replaced • 42 MHz Digital Return will have to be replaced
Node (RF Side)	Worst Case: Replace the housing because there is no pluggable filter or amp that fits into existing housing
Amplifiers	Worst Case: Replace the housing because there is no pluggable filter or amp that fits into existing housing
House Amplifiers	High-split will require change of a home amp
OOB Set-Top Box Communications	• Some STBs may be hard coded within the mid-split range (75.5 and 104.25 MHz) • ANSI/SCTE 55-2 2008 [6] and ANSI/SCTE 55-1 2009 [7] defines 70 MHz – 130 MHz as usable.

Table 4—High-Split Disadvantages



Red Circle Areas: represent investment or possible investment
Green Circle Areas: represent no investment required

Figure 15—Anatomy of the Mid-Split and High-Split Architecture

Tri-split 1.3 – 1.8 GHz Analysis

The Tri-Split Architecture may be defined as 1.3 – 1.8 GHz this may also be referred to as spectrum overlay. The Tri-split may be considered because this avoids consuming existing downstream in terms of Capacity, Services, OOB STB management and the entire DS architecture does not have to change. As with mid-split and high-split DOCSIS 3.0 specifications systems may be used, however to take advantage of the full spectrum additional development is required. Tri-split is a touch it once architecture in that this clearly competes against PON. The tables below capture some the advantages and disadvantages when considering Tri-Split.

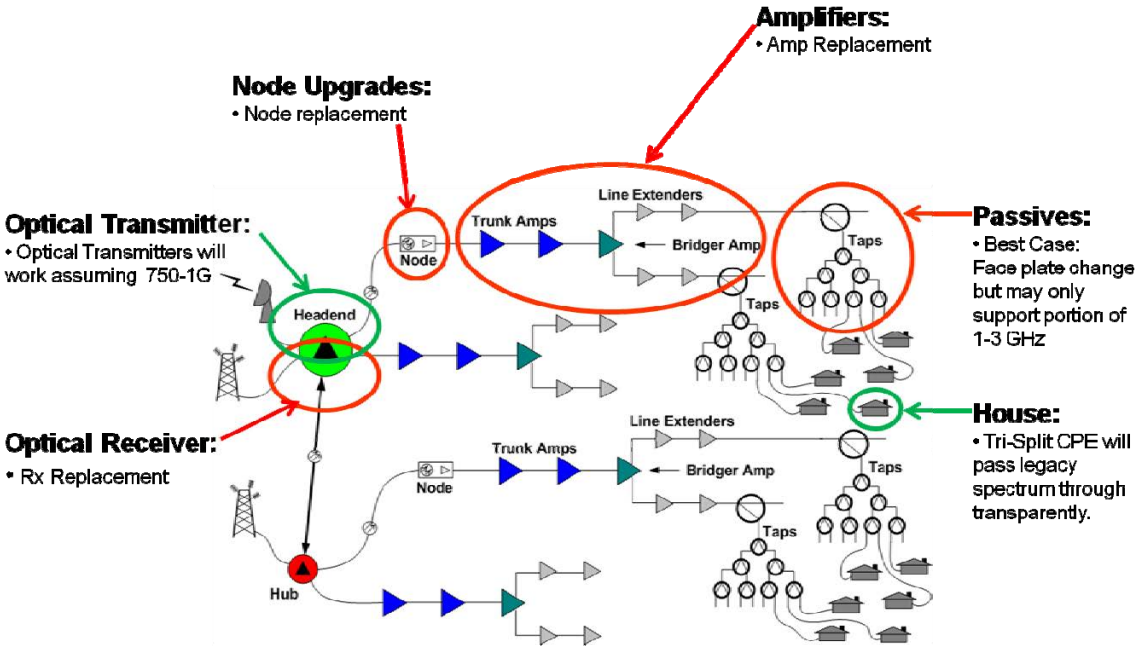
Tri-Split Advantages	
Area	**Comment**
Bandwidth	• 1 – 2 Gigabits with existing DOCSIS technology • 1 Gbps Speeds maybe assumed if we account for MoCA bandwidth which may leak into plant. • Clearly competes against PON in terms of bandwidth over HFC
Spectrum Upstream Allocation	1.3 GHz - 1.8 GHz Upstream In the Tri-Split there is no ceiling (Upstream does

	push up against the Downstream)
Impact to Exiting Forward Capacity	No Change to the Downstream in terms of: Capacity, Services, and the entire DS Architecture does not need to change.
Headend Optical Transmitter	Existing Equipment should support legacy downstream
Passive	Face Plate Change Possible
House Amplifier	Possibly Leveraged for existing spectrum
OOB Set-Top Box Communications	Not Affected
DOCSIS 3.0 CMTS and CM/EMTAs	Recent DOCSIS 3.0 products (CMTS and EMTAs) are built to use mid-split spectrum this maybe leveraged for full Tri-Split

Table 5—Tri-Split Advantages

Tri-Split Disadvantages	
Area	**Comment**
Bandwidth	-
Spectrum Guard band (between DS and new upstream)	Guard band 300 MHz
Spectrum Interference Concerns	MoCA 1.0 ratified in 2007 may operate in 850-1,500 MHz range MoCA 1.1 ratified in 2008 may operate in a 50 MHz band in the range the 850-1,550 MHz range [8]
Headend Optical Receivers	Replace
Node	Replace
Amplifiers	Replace
Home	High Frequency usage results in high power level required from the CM

Table 6—Tri-Split Disadvantages



Red Circle Areas: represent investment or possible investment
Green Circle Areas: represent no investment required

Figure 16—Anatomy of the Tri-Split Architecture

Quad-Split Analysis

The term Quad-Split may be applied to additional forward capacity in the 1 – 3 GHz range. This new downstream spectrum if used would likely be placed on top of the upstream allocation which may occupy 1.3 – 1.8 GHz.

Fiber to the Last Active (FTTLAx)

The term Fiber to the Last Active (FTTLA) refers to extending fiber from the node or overlashing from the headend or hub to each active on the plant. This network transition may be used in conjunction with Mid-Split, High-Split, or Tri-Split technologies. This architecture may be considered if a non DOCSIS MAC/PHY layer technology is used and may not support transmission through actives or only through few actives due to performance and distance limitations. It is unlikely that this approach would be considered because of the massive fiber builds, optical network transition to WDM for fiber conservation, and an increased number of actives in the plant if addition new MAC/PHY or media conversion elements are added.

Summaries of Upstream Augmentation

The use of Upstream Augmentation is not anticipated in the near future because of the massive upstream bandwidth the cable industry has in place today. The use of upstream augmentation is another example that the existing coaxial network to the home may evolve to support the demands of the consumer. These approaches cited above extend the useful life of the existing HFC investment.

EXAMINING THE MAC AND PHY LAYER TECHNOLOGIES

As described in the previous section the underling physical layer spectrum options have been examined and it is planned that the MAC/PHY layer options described in this

section may function in any spectrum allocation.

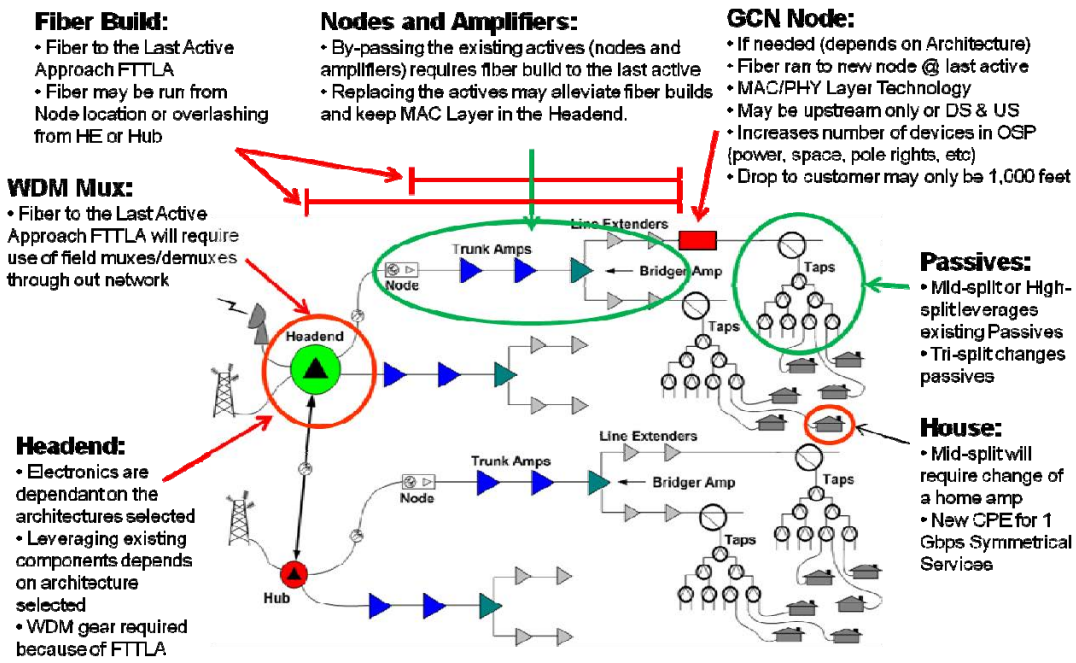
Alternative MAC and PHY

An emerging class of the technology called Ethernet over Coax (EoC) may become a competitive technology to DOCSIS. The use of EoC may be applied to the new spectrum allocation as described in the previous section. The term EoC may reference current standards such as MoCA, HPNA 3.1, G.hn, and perhaps others that may emerge. These technologies were typically used in premise distribution networks and some may find applications in the access network. Some EoC technologies may however have some significant drawbacks when used as an access layer technology such as distance limitations and number of end customers supported in a given MAC/broadcast domain. In addition, consideration of MAC layer limitations in terms of QoS required in a shared media access layer technology to assure QoS for each customer and service type at a scaling level required for the coaxial access network should be considered. The scaling of the MAC layer domain may be a significant consideration. In the cable access network this is a key factor as cable may have 1,000 of customers sharing a MAC domain and as bandwidth increases number of customers served in a MAC domain may rise for economies of scale. The number of unique end points served in a given MAC domain, such as 32, 64, or even 256 subscribers may not be sufficient in a cable access network application.

The use of new PHY layer technology may be desired to increase the bits per second per hertz. This may include the use of orthogonal frequency division multiplexing (OFDM) or similar approaches.

A key architecture consideration about some EoC based technologies is that the distance limitation from the beginning of the coaxial drop to the customer premise may need to be 1,000 feet or less. This is a critical consideration because a drop of not greater than 1,000 feet of coax may require an outside plant infrastructure using Fiber to the Last Active (FTTLA). This may result in a significant cost premium when considering DOCSIS architectures which may travel up to 100 miles or 160 km. These are all key selection criteria when considering alternative MAC and PHY layer technology in access layer deployment architecture.

The architecture illustration considers a fiber to the last active (FTTLA) and the selection of Mid-split, High-split, or Tri-Split. This architecture assumes a drop distance of no greater than 1,000 feet and that no existing actives are leveraged. This architecture places a device called a Gigabit Coaxial Network Node at the last active location and throughout the entire network (note this would not be required if DOCSIS was used).



Red Circle Areas: represent investment or possible investment
Green Circle Areas: represent no investment required

Figure 17—Possible Architecture Using FTTLA and Non-DOCSIS MAC/PHY

Traditional MAC and PHY

The use of DOCSIS MAC and PHY layer technology may be considered for the Gigabit Coaxial Network. The use of existing DOCSIS 3.0 standards may be leveraged as this would already be occupying spectrum and bandwidth. If the existing DOCSIS standards were considered this would allow for large bonding groups to be leveraged as these may terminate on existing or current DOCSIS 3.0 upstream cards. Conversely if an alternative MAC layer technology is used the spectrum DOCSIS occupies may affect the possible throughput of the solutions. There may also be the presence of significant number of existing DOCSIS 3.0 channels occupying downstream capacity and this too may be leveraged for the Gigabit Coax Network. DOCSIS was designed for cable access network distribution and leveraging DOCSIS for the next generation coaxial access network may be considered as this would continue to place the electronics for MAC/PHY processing in the Headend and CPE, thus avoiding placing these active components in the OSP. In addition, the distance capabilities of DOCSIS will continue to allow MAC/PHY processing at the current distances thus not requiring a GCN Node in the OSP plant, this is a significant difference as the current active counts in the OSP would not increase. Moreover leveraging DOCSIS and either Mid-Split, High-Split, or Tri-Split approaches would not requires additional fibers or wavelengths to be deployed.

DOCSIS could evolve adopting new PHY layer encoding technologies like OFDM to improve performance [9] and also improvements to the MAC could be adopted which may strengthen the position to use DOCSIS to support Multi-Gigabit or Gigabit IP Services to compete with PON.

COST ANALYSIS FOR UPSTREAM AUGMENTATION

The following assumptions were used to compile a comparison of Mid-Split conversion, Tri-Split, and EPON overlay:

- 30% of the infrastructure is underground with limited or no conduit access.
- Homes passed density averages 100 homes/mile.
- Expected symmetrical service take rate is 15% of homes passed.
- Enhanced HSD service group size target is 1024 homes passed.

The results of the comparison revealed that the EPON overlay had the highest infrastructure cost, the highest success-based cost, and the longest build out time of the group. The other solutions were then compared to EPON on a percentage basis.

Mid-split

Changing the upstream/downstream split boundary in existing HFC networks is an economical approach to providing additional upstream bandwidth. Only the active elements in the network need to be upgraded to alter the split boundary. The amplifiers will need modified diplex filters to be substituted, and the nodes will likely need upgraded DFB lasers to handle the additional QAM traffic load. Infrastructure costs for this conversion using the stated assumptions were calculated to be on the order of 23% of the EPON overlay for traditional Node + X amps in cascade architectures, and 29% for Node + 0 architectures (Figure 18). This represents the lowest enablement costs and fastest time to market solution out of those considered (Figure 19).

Tri-Split

The overlay approach requires much more material and effort. The upstream traffic will be shifted to a spectral area above 1 GHz which the taps and passives will not cleanly pass. Therefore, all the taps and passives would have to be replaced with upgraded versions. This could possibly be done with faceplate upgrades to minimize cost, time to market and customer disruption. Each amplifier would have to be retrofitted or replaced, in order to filter and amplify the new upstream frequency band. Each node would have to be retrofitted or replaced, in order to filter, amplify and optically transmit the new frequency band. A new receiver arrangement in the headend or hub would be necessary to receive an convert the optical signals to RF in the appropriate frequency range of the termination system. Infrastructure costs for this conversion using the stated assumptions were calculated to be on 54% of EPON overlay for traditional Node + X amps in cascade architectures, and 61% for Node + 0 architectures (Figure 18). This represents mid-level enablement cost solution with a much longer time to market over the mid-split (Figure 19)

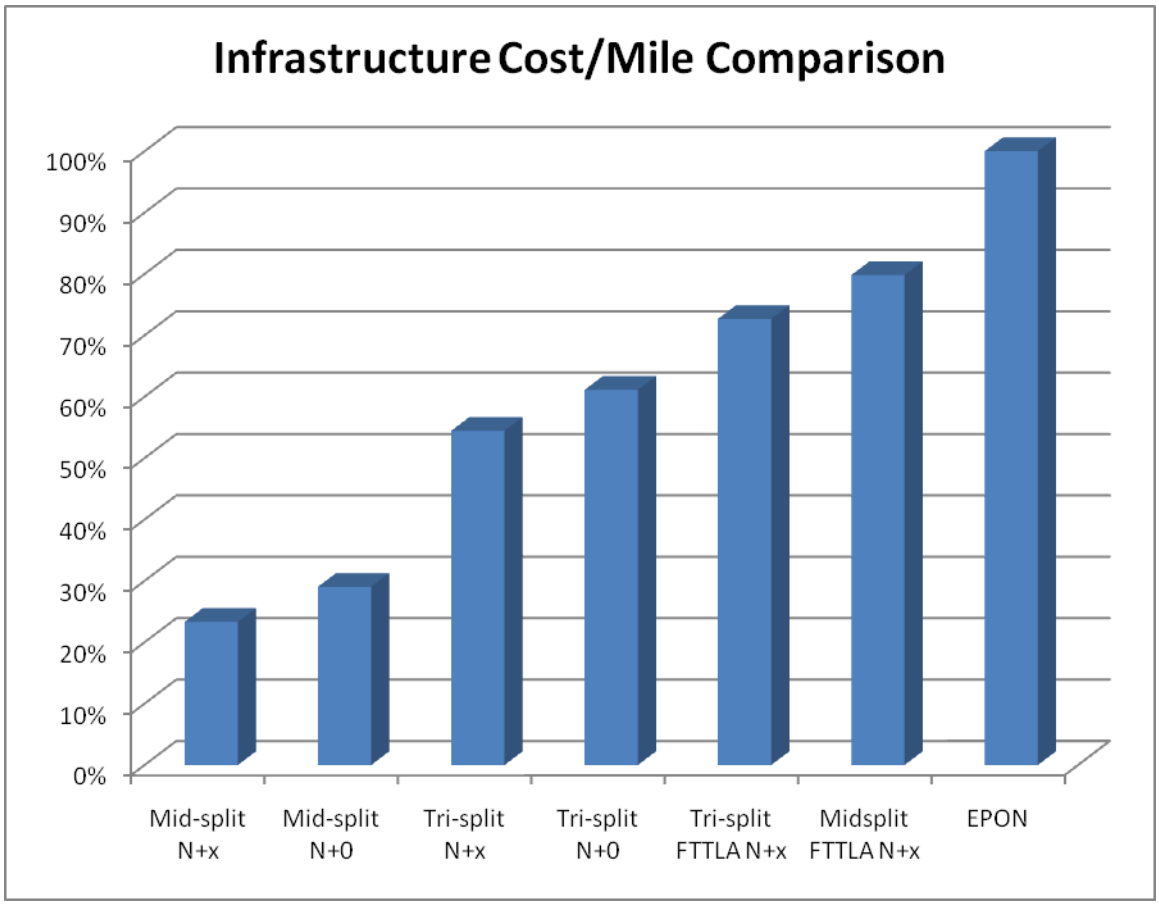


Figure 18—Infrastructure Cost per Mile Compared to FTTx & EPON Overlay

Fiber to the Last Active (FTTLA)

Converting an existing Node + X architecture to a Node +0 architecture is commonly known as a Fiber to the Last Active (FTTLA) upgrade. If this approach is combined with the previous mid-split and overlay approaches, then the additional costs of extending the fiber and converting the current amplifiers to nodes is added. The additional fiber cost isn't affected by which solution the FTTLA is combined with, but the active cost burden for the overlay is shared with the FTTLA conversion, so some economies are realized. When upgrading to mid-split along with FTTLA the cost increases to 80% of EPON overlay. However upgrading to overlay along with FTTLA increases the cost to 73% (Figure 18). These solutions represent the highest cost HFC solution with the longest time to market of the HFC approaches (Figure 19).

EPON Overlay

Extending fiber to every customer premise is a well known solution for high rate symmetrical data delivery. Overlashing in the existing aerial plant, and trenching and boring would likely be required in most underground areas to add the required fiber to overlay an EPON system. Even with very moderate construction labor estimates, this solution is significantly more expensive that the HFC upgrade scenarios.

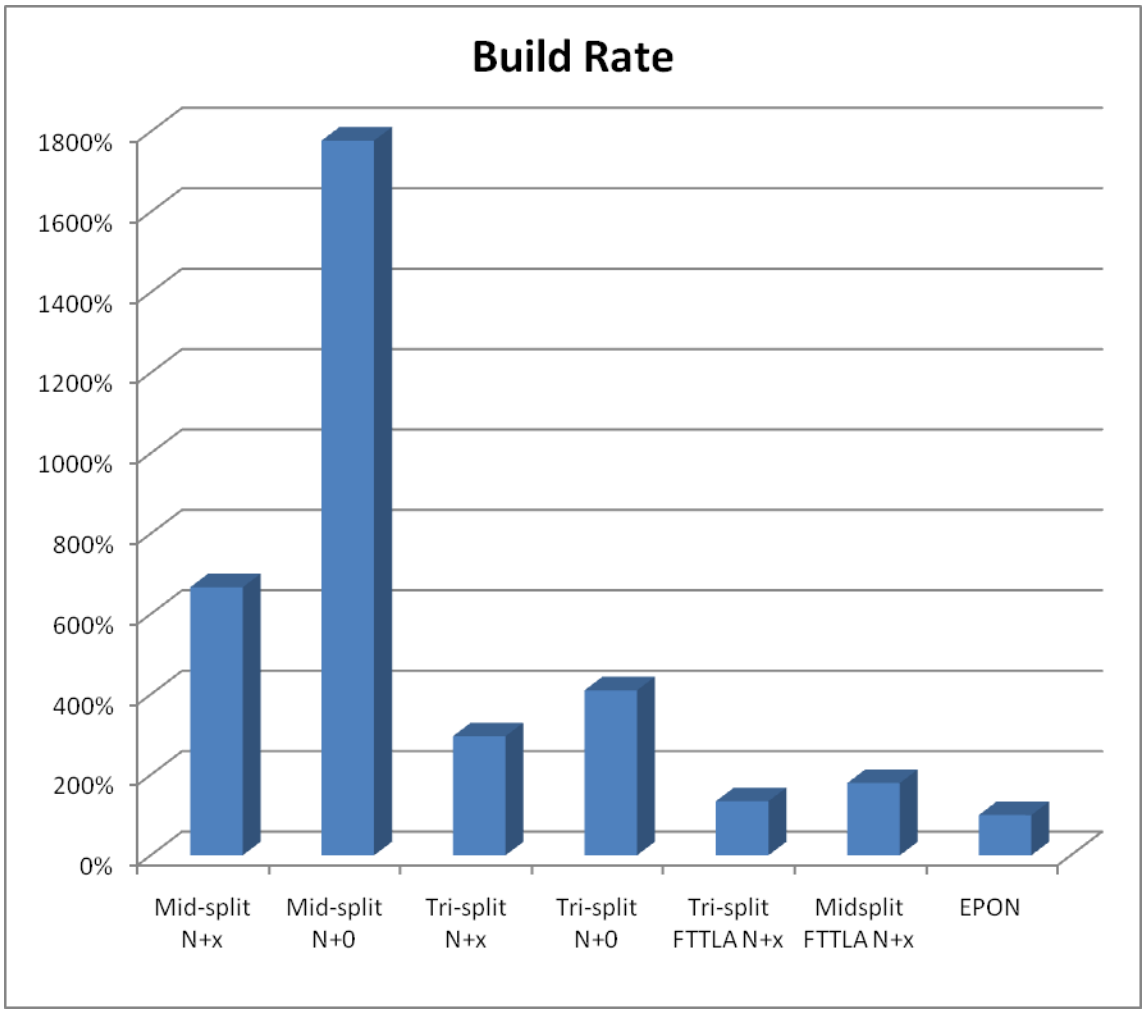


Figure 19—Build Rate Time to Market Advantage When Compared to EPON

Success Based Costs

Once the infrastructure is available, the costs to provide service must also be considered. In this analysis, it was assumed that the HFC plant is mature and drop facilities exist to the typical customer premise. This is not true however, for the EPON solution since it is fiber based all the way to the side of the premise. It is also assumed that the CPE costs for a next gen cable modem and EPON ONU are similar for the same throughput capability. The assumptions listed yield 15 customers per mile. Again using the EPON solution as the benchmark, the HFC solutions were compared with 15% take rate. The relationship is shown in Figure 20.

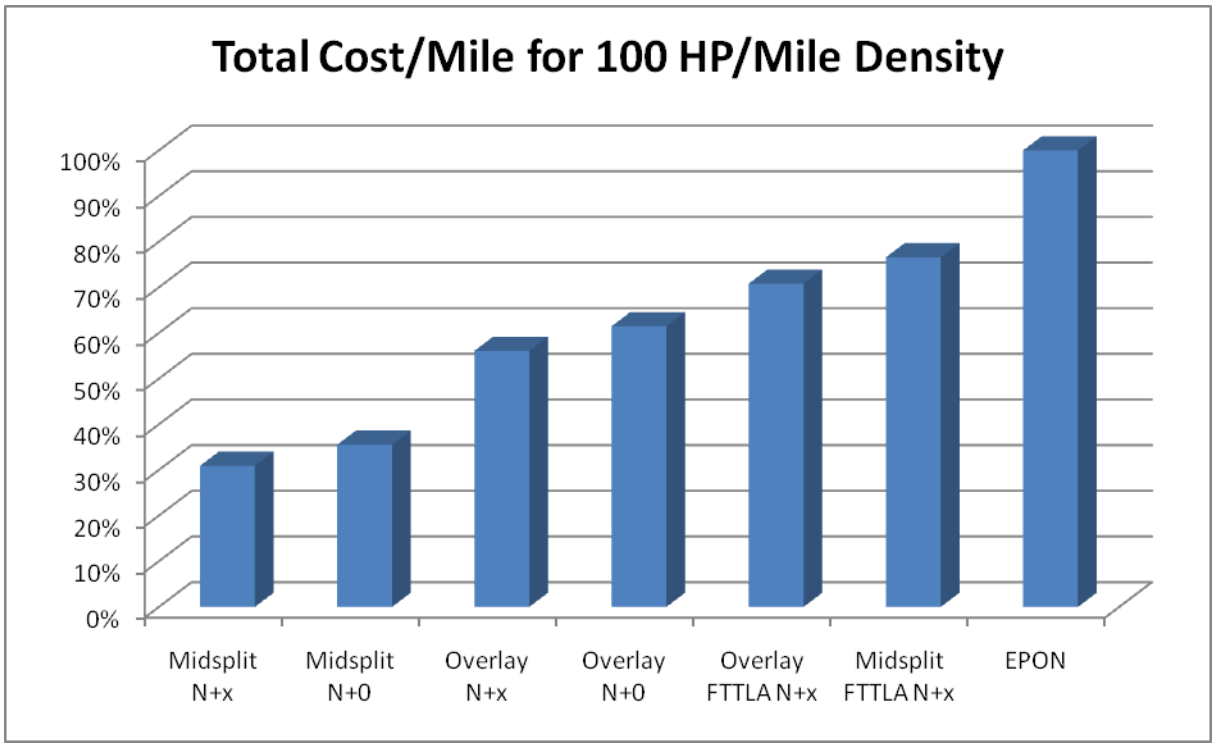


Figure 20—Total Cost per Mile Compared to FTTx & EPON Overlay

CONCLUSIONS

The cable industry may be part of a significant business and technical transition driven from considerable changes in technology, consumer demand and competition. The cable operators will have the ability to transform their business and network from largely a broadcast oriented content delivery service to an increasingly more personalized customer experience. While the demands of the business change, the network, as illustrated in this paper, is incredibly nimble and flexible to accommodate this transition. As the cable industry transforms their network to a unicast service delivery platform it is likely that an increase in spectrum and bandwidth allocation will be allocated to IP based network technologies to address consumer demands and competition. The use of upstream augmentation allows the cable industry to address competitive threats posed by fiber to the premise (FTTP) providers. The Next Generation – Gigabit Coaxial Access Network is capable of delivering multi-gigabit IP services to the home while having the ability to eventually support 1 Gbps from the home, all while leveraging one of cable's most valuable assets the existing coaxial network to the home.

REFERENCES

[1] Tom Cloonan, "On the Evolution of the HFC Network and the DOCSIS® CMTS - A Roadmap for the 2012-2016 Era," SCTE Cable Tech Expo, 2008.

[2] Mark Bugajski," IP adding IPTV over DOCSIS® 3.0 — How will this help?" ANGA Cable Conference, 2009

[3] Ayham Al-Banna and Tom Cloonan, "DOCSIS3.0 US channel bonding: Performance analysis in the presence of HFC noise," SCTE Conference on Emerging Technologies, 2009

[4] Data Over Cable Service Interface Specifications DOCSIS 3.0 Physical Layer Specification, CM-SP-PHYv3.0-I07-080522, CableLabs, 2008.

[5] Data-Over-Cable Service Interface Specifications DOCSIS 3.0, MAC and Upper Layer Protocols Interface Specification, CM-SP-MULPIv3.0-I08-080522, CableLabs, 2008.

[6] ANSI/SCTE 55-2 2008

[7] ANSI/SCTE 55-1 2009

[8] Charles Cerino, "The Standard for Home Entertainment Networks Over Coax™, KLabs Conference, 2008.

[9] Ayham Al-Banna and Tom Cloonan, "Performance Analysis of Multi-Carrier Systems when Applied to HFC Networks" SCTE Conference on Emerging Technologies, 2009.

LIST OF ABBREVIATIONS AND ACRONYMS

BPON	Broadband PON
CAGR	Compound Annual Growth Rate
DBS	Digital Broadcast System
DOCSIS	Data Over Cable Service Interface Specifications
DTA	Digital Terminal Adapter
EoC	Ethernet over Coax
EPON	Ethernet Passive Optical Network
FDM	Frequency Division Multiplexing
FTTH	Fiber To The Home
FTTLA	Fiber to the Last Active
FTTP	fiber to the premise
Gbps	Gigabits per Second
GCN	Gigabit Coax Network
GPON	Gigabit PON
HFC	Hybrid Fiber Coaxial Cable
HHP	Households Passed
HPNA	HomePNA Alliance
HSD	High Speed Data
IP	Internet Protocol
IPTV	TV (video) over IP networks
MAC	Media Access Layer
Mbps	Megabit per Second
MoCA	Multimedia over Coax Alliance
MSO	Multiple Service Operator
OFDM	Orthogonal Frequency Division Multiplexing
OTT	Over The Top
P2P	Peer-to-peer
PHY	Physical Layer
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RFoG	RF Over Glass
SDV	Switch Digital Video
US	Upstream
VBR	Variable bitrate
VDSL	Very High Bitrate DSL
VDSL2	Very High Bitrate DSL2
VoD	Video on Demand

REFUELING THE CABLE PLANT – A NEW ALTERNATIVE TO GAAS

Phil Miguelez, Fred Slowik, Stuart Eastman
Motorola

Abstract

GaAs hybrid and MMIC technology has enabled improved distortion performance and bandwidth expansion capability up to 1 GHz for the past 10 years. Now a new HFC semiconductor technology, Gallium Nitride (GaN), is coming on line with significant improvements in surge voltage ruggedness, better thermal performance, and capability of higher output levels to extend reach and further bandwidth expansion for new and existing cable plants.

Higher output levels achievable with GaN technology enable operators to lower initial capital costs and operational expenses for fiber deep network deployments. This paper will describe the advantages of GaN technology compared to current GaAs devices and provide design examples showing the potential cost savings in high to low density green field applications. Brown field extensions and bandwidth extensions where GaN can help to minimize cost will also be covered.

INTRODUCTION

In the late 1990's Gallium Arsenide (GaAs) MESFET based gain blocks were first introduced into cable plant actives. Within a few years GaAs hybrids and MMIC's completely replaced the silicon devices that had been the mainstay of cable nodes and amplifiers for the previous three decades. The extended gain bandwidth, lower noise, and improved distortion performance advantages of GaAs enabled cable operators to expand BW from 750 MHz to 870 MHz to the 1 GHz systems that are being deployed today.

The benefits of GaAs hybrids and MMIC's have been significant. These devices contributed to a seamless forward path transition from 64 QAM to 256 QAM in the access network. GaAs also allowed the bandwidth extension to 1 GHz of legacy 750 and 870 MHz systems while maintaining > 90% of the current actives locations.

Now, after a successful run of more than 12 years, GaAs gain blocks are about to be succeeded by a new generation of semiconductor device technology that has already proven its capabilities in numerous military, space, and commercial applications.

Gallium Nitride (GaN) is another III –V group direct band gap semiconductor just like Gallium Arsenide but with unique properties that have allowed the development of daylight LED's, Blu-ray lasers, and high power / high frequency RF amplifier devices. Like many semiconductor device technologies, GaN development was initially funded by the government to take advantage of its high power RF amplification and radiation resistant capabilities in space based applications. More recently GaN RF devices have begun to challenge Silicon LDMOS in the 2 GHz WiMax power amplifier base station market.

The commercial applications and availability of GaN semiconductor devices continues to expand. Until relatively recently larger scale devices were primarily processed for high voltage (> 50 Volts) operation. Now the major GaN wafer fabrication vendors have qualified devices that are optimized for lower operating voltages that are compatible with cable node and amplifier electronic circuit packs.

The major impact of GaN on the access network is its capability for significantly

increased output levels without sacrificing distortion performance. Extended reach for nodes and amplifiers means fewer total actives are required for a given system design and therefore a measurable reduction in capex and opex spending. The remainder of this paper will describe the basic technology benefits of GaN semiconductors, the impact to node and amplifier station performance, and the system cost savings implications for different HHP serving size areas.

GaN TECHNOLOGY

Gallium Nitride devices designed for amplifier applications are typically constructed as heterostructure FET's also referred to as High Electron Mobility Transistors (HEMT's). Using a deposited layer of highly doped AlGaN and a non doped GaN channel layer a junction is created with a large band gap. Electrons generated in the AlGaN layer are swept into the quantum well created between the different band gap material layers. The effect is the creation of high mobility electrons. These HEMT devices provide high current gains and power gains at frequency bandwidths not usually attainable with traditional MESFET structures.

Gallium Nitride has a wider band gap (3.4 eV) than either Silicon (1.2 eV) or GaAs (1.4 eV). This property combined with the ability to operate at high voltages results in devices with roughly 10 times higher power density while maintaining wide bandwidths at frequencies up to 4 GHz.

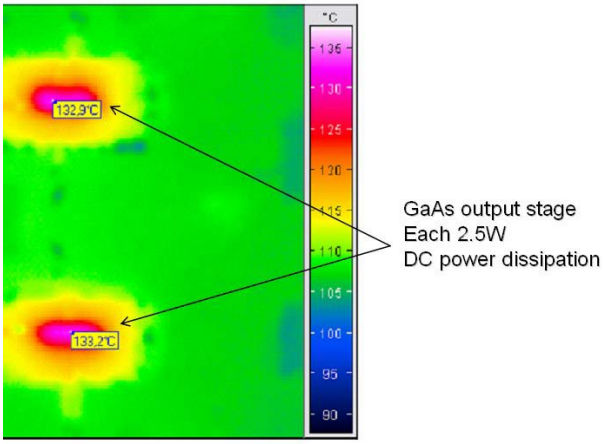
Fabricating bulk pure crystalline GaN wafers proved to be so difficult that early applications for GaN was limited to small area devices such as LED's and specialized lasers or military applications which could absorb the higher costs of small area wafer processing. Advances in chemical vapor deposition and vapor phase epitaxy growth during the 1990's allowed GaN thin films to be deposited on silicon (Si) and silicon carbide (SiC) wafers achieving the possibility of large scale fabrication. Today, Gallium Nitride RF devices are primarily fabricated as discrete die or packaged as part of a multi-chip module to achieve a higher level functionality.

GaN devices processed on Si or SiC wafer substrates have significantly improved thermal performance compared to devices built on GaAs wafers. The thermal conductivity of Gallium Nitride is 2X higher than Gallium Arsenide. Fabricated on SiC wafer substrates (Tc = 4.9 W/cm*K), allows GaN devices to operate at higher output power levels while maintaining the same or lower junction temperatures than equivalent GaAs devices.

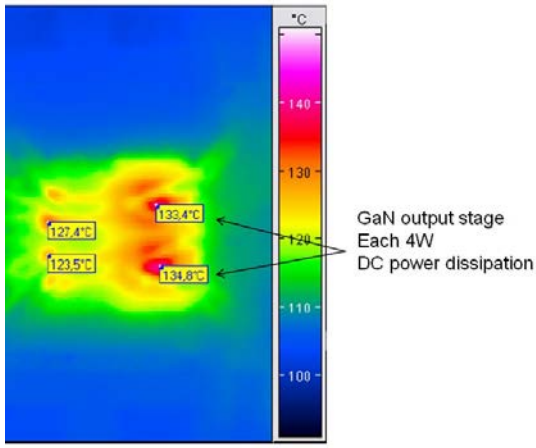
Figure 1 shows the worst case hot spot die temperature of a GaAs power doubled hybrid and a new replacement GaN power doubled hybrid. Like all cable gain block devices these hybrids are typically biased at nearly class A levels with the RF output backed down several dB in order to provide the best intermodulation distortion performance. The GaN output die operate at a significantly higher bias voltage level which in part explains its higher output power capability but also increases power dissipation >1.5X higher than the GaAs equivalent device. Even with this increase in DC power dissipation the higher thermal conductivity of GaN and its SiC substrate allow the GaN hybrid to stay at the same die temperature.

Starting with the initial introduction of GaAs devices in cable plant equipment, ESD and surge voltage ruggedness has always been a concern. The gate structure of typical GaAs FET's can not survive the high current flows usually produced by a transient pulse such as ESD. As a result, each manufacturer has incorporated a number of additional protection circuits and components to increase the raw withstand voltage capability of GaAs

hybrids and MMIC's. The GaAs gain block devices deployed today are very rugged against the transient spike events that can occur in cable plant environments. The typical powered ESD values for today's GaAs



Vcc=24V; IDC=478mA
T(heatsink)=100.0°C



Vcc=24V; IDC=462mA
T(heatsink)=100.2°C

FIGURE 1 – Worst Case Hot spot Temperature Measurements for GaAs and GaN PD Hybrids

gain blocks is ~ 1KV. GaN HEMT devices have a naturally higher ESD withstand threshold of 1600 to 1800 Volts. With additional protection circuitry as used with GaAs devices the ESD ruggedness could potentially increase even further.

GaN DEVICE IMPACT ON CABLE PLANT ACTIVES PERFORMANCE

The most significant performance limitation in access networks today is composite carrier to noise (CCN). Ever increasing digital loading up to 1 GHz creates carrier to intermodulation noise (CIN) which along with thermal noise generated in the various active components combines to produce CCN. While the other analog distortions (CSO, CTB) are still very important to the performance of the network the technology improvements due to GaAs implementation and gain block design over the past 10 years have leveled the playing field among amplifier and node vendors with respect to these 2nd and 3rd order distortions. CCN is now the dominant distortion that determines the maximum output level that nodes and amplifiers can achieve.

Cable gain block hybrids or multi-chip module MMIC's incorporating Gallium Nitride output stages provide higher output level capability while maintaining the same gain, power consumption, and physical dimensions as equivalent GaAs devices. In the following amplifier and node examples the impact of GaN on link and station performance will be clearly evident.

Figure 2 shows comparative data for CCN performance of a 1 GHz Line Extender tested with existing GaAs hybrid gain blocks and with Motorola's new GaN technology hybrid gain blocks. The testing was performed with a full analog + QAM channel load and typical amplifier tilt of 13.5 dB. Although this is single station data the plots illustrate the improved distortion performance of GaN particularly as the output level is increased. This enhanced performance allows the BLE output to be increased an additional 2 to 3 dB

from typical GaAs operating levels without degrading end of line distortion.

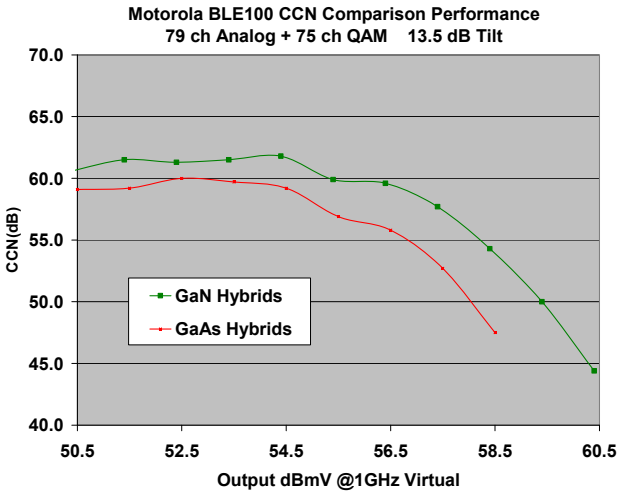


FIGURE 2
BLE Single Station CCN Test Results
(25C Data)

At lower output levels GaN still provides about 1.5 dB of CCN headroom compared to GaAs. This performance improvement with GaN could be used in existing brown field locations where legacy actives are stretched and additional CCN margin is desired.

In this example and the others that follow the CTB and CSO distortion performance with GaN is better or equal to the original GaAs amplifier values.

Figure 3 demonstrates link performance testing using the Motorola SG4 segmentable node in an N+0 configuration. With existing GaAs hybrid gain blocks the SG4 is capable of +58 dBmV (1 GHz virtual) output at 18 dB tilt and a full 1 GHz channel load. The same link utilizing Motorola GaN technology hybrid gain blocks provides 3 dB of additional output level for the same distortion performance. At lower levels the CCN results are dominated by the optical link and therefore the difference between GaAs and GaN is not as dramatic.

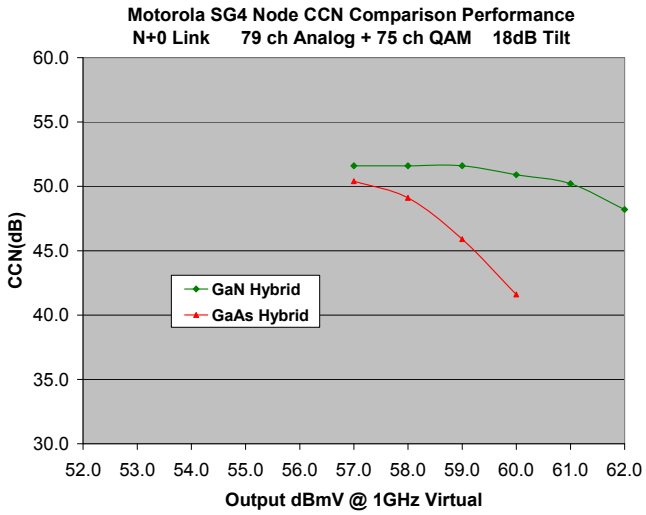


FIGURE 3
SG4 Link Performance, 20km Fiber + Passive Loss (25C Data)

The impact of GaN becomes even more apparent as the channel loading moves to all QAM and tilts are increased to maximize high frequency reach. Figure 4 compares the CCN performance of an N-split (85 / 108 MHz) three output Mini-Bridger amplifier station loaded to 1 GHz with only 256 QAM channel loading. At lower output levels the difference in performance is roughly 2 dB, similar to the results seen in the Line Extender example with analog + QAM loading. The real impact occurs as the output level is increased. Here the higher crash point of GaN allows 3 dB of added link margin.

Specifying the operating point for an all digital link requires careful consideration. As can be seen in Figure 4 the virtual level for either GaAs or GaN based amplifiers could be claimed at values as high as 61 dBmV or 63 dBmV and most likely produce acceptable BER / MER under ideal conditions. The problem is that operating on the right hand side of the graph or “crash region” for these devices means that any variations in gain level, temperature, aging, etc. would cause a large swing in the performance. In the worst case the link CCN could slip below the

minimum acceptable level and severely degrade end of line performance. A conservative approach would be to operate at the peak of the curve or slightly to the right (~ 1 dB) of the peak.

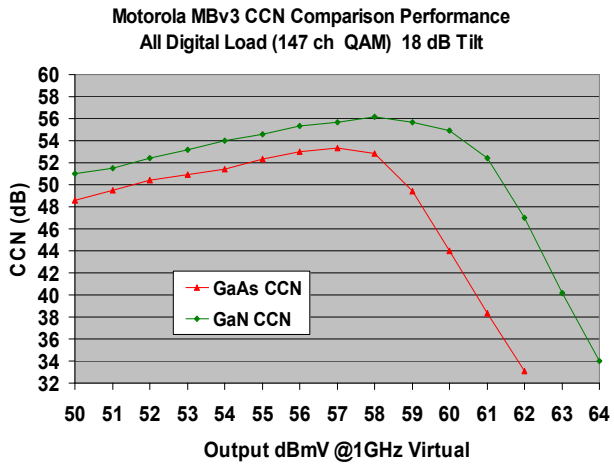


FIGURE 4
MB Single Station CCN Test Results
(25C Data)

GaN BENEFITS ARE IN THE APPLICATIONS

Now that the technical description and performance characteristics of GaN technology have been presented, it is time to discuss the various applications and subsequent results. The obvious method of proving the benefits of GaN is to perform a series of network modeling exercises designed to directly compare new GaN versus existing GaAs technology and let the chips fall where they may. In order to do this we have selected several sample design areas with varying density and topology to focus upon. These sample networks consisted of the following types:

1.) Urban Plus density Greenfield plant which averages 256 Homes per Mile (HPM).

2.) Suburban density Greenfield plant which averages 96 HPM.

3.) High-rise Multiple Dwelling Units (MDU) Greenfield consisting of 133 and 223 units per building.

4.) Rural density Brownfield 550 MHz to 1000 MHz plant up-grade which averages 18 HPM.

We also considered looking at rural Greenfield plant however, that did not make much sense since most rural Greenfield plant construction is moving towards RF over Glass (RFoG) as the predominant lowest cost architecture for sub 50 HPM densities.

NETWORK DESIGN PARAMETERS

All designs adhered to the following requirements:

1.) Technology:
GaN versus GaAs

2.) Architecture:
Traditional HFC with a variety of amplifier cascades deployed as necessary to permit maximum node sizes of 700 HPN in the Urban Plus area and 500 HPN for the Suburban and Rural areas.

3.) Frequency:
54-1000 MHz as the
Downstream (DS)
5-42 MHz as the
Upstream (US).

4.) Channel Loading:
78 analog to 550 MHz,
The remaining is
256 QAM to 1000 MHz.

5.) Cable Type:
Greenfield: P3-750, P3-500
Brownfield: P3-750, P3-500 &
P3-625
MDU: RG-11 and RG-6

6.) Tap Port Level:
19 dBmV @ 1000 MHz Virtual
15 dBmV @ 550 MHz Actual
10 dBmV @ 54 MHz Actual

7.) Network Powering:
90 VAC

8.) Network Performance:
CNR: 49.0 dB
CCN: 48.0 dB
CTB: 57.0 dB
CSO: 56.0 dB

DESIGN OBJECTIVES

In all network design models, the objective was to produce the most efficient design from an equipment usage perspective. The same network designer was used for all sample design models in order to eliminate design talent diversity.

In all sample designs, we maintained the same service area boundaries. In other words, no attempts were made to expand the node serving area reach in order to reduce optoelectronics quantities and related cost. Although this could have been done, since the trend is moving towards smaller node size, we did not factor this into the modeling effort.

The MDU design utilized a tapped riser design approach. A tapped riser design usually has actives placed in a storage closet or stairwell with taps placed on each floor and drops run to each apartment. This is illustrated in the Figure 5 diagram.

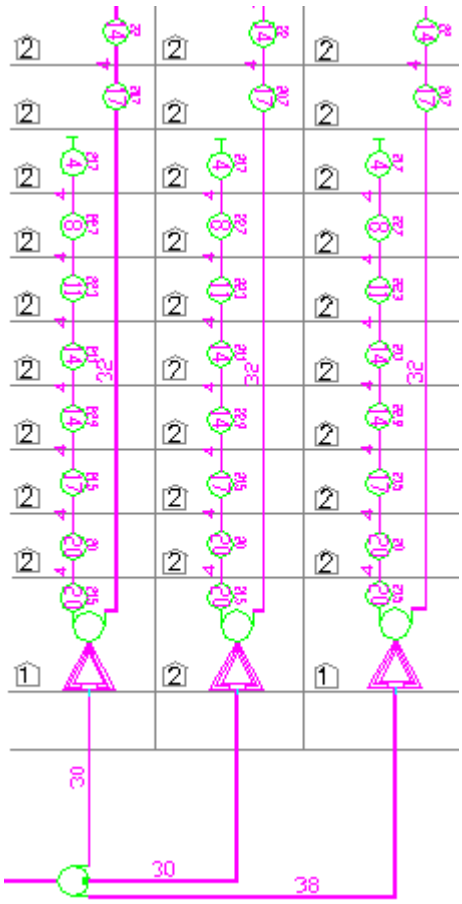


FIGURE 5

NETWORK MODELING RESULTS

Results of the sample design models indicated variations in four major areas: network active counts, network powering, cable usage, CAPEX and OPEX.

Let's discuss these individually prior to presenting Tables A, B and C which offers a detail level review of the results.

1.) Network Active Counts: two significant things occurred in this area. First, the total active counts were reduced in the GaN designs due to increased operating levels.

Typically, GaN enables a 2-3 dBmV higher output capability than the GaAs counterpart. The active count reduction

ranged from 6% to 30% depending upon the design area considered.

Second, the types of active devices used also changed. High density areas used few amps with more outputs, while the suburban area used fewer outputs per active. Both of these elements reduce network CAPEX and OPEX.

2.) Network Powering: the GaN designs yielded a significant reduction in power consumption. This ranged from 12 % to 19 % for the various models. This power reduction can lower CAPEX for initial deployment by reducing P.S type from 15amp to 12 amp for example. Additionally, since plant powering costs can range from $200 to $400 per plant mile per year, a 12% to 19% power reduction may result in an annual OPEX savings of $76 per mile.

3.) Cable Usage: the quantity and mix of cable varied slightly from design to design, ranging from -2.5% to +2.5%

4.) CAPEX: overall network electronics cost were reduced in the GaN designs ranging from 7.3% to 8.7%.

GREENFIELD MODEL

Let us now take a closer look at specific results provided. In Table A we look at the impact that GaN has in a Greenfield environment.

GREENFIELD NODE	Urban Plus GaAs	Urban Plus GaN	Suburban GaAs	Suburban GaN
Plant Mileage	2.72	2.72	9.67	9.67
Aerial	2.72	2.72	7.08	7.08
UG	0	0	2.59	2.59
Total Actives	7	5	31	29
Actives/Mile	2.6	1.8	3.2	5.9
Cascade	N+1	N+1	N+4	N+4
House Count	695	695	924	924
HC/Mile	256	256	96	96
1GHz Design				
Actives Used	**7**	**5**	**31**	**29**
SG4000	1	1	2	2
BLE100			9	10
MB100	1		8	15
MBV3	4		11	3
BTD	2	5	3	1
Powering				
15 Amp Power Supplies	1	1	3	3
Total Power Draw (Amps)	9.27	7.51	30.4	25.97
% Power Savings		19.0%		14.6%
Cable Footage	**22,071**	**22,692**	**76,112**	**74,144**
Total Electronics	$ 10,999.27	$ 10,201.15	$ 26,983.19	$ 24,933.40
Per Mile Electronics	$ 4,043.85	$ 3,750.42	$ 2,790.40	$ 2,578.43
% Change to electronics		-7.3%		-7.6%

TABLE A – Greenfield Urban + and Suburban

In both cases studied there was a reduction in the number of actives used, power used and total equipment cost. Active types also changed in each design. In the high density area we found that due to the increased output capability of GaN devices, more multiple output devices were chosen for increased design efficiency. The lower output GaAs product for like amplifier types did not produce this same advantage. The suburban density showed a decrease in the number of outputs per active device. Without the multiple paths that high density areas provide to take advantage multiple output amplifiers, use of this type of active creates the need for dual cable feeds making for a less efficient design. Instead, the higher output allowed for fewer outputs per active, reducing the amount of dual cable. In all cases it was discovered that an equivalent area can be fed using a fewer output GaN amplifier as was covered with a GaAs active.

MDU MODEL

In Table B we look at the impact that GaN has when used in a Greenfield high rise MDU environment.

GREENFIELD NODE	**MDU A GaAs**	**MDU A GaN**	**MDU B GaAs**	**MDU B GaN**
Total Actives	3	3	9	6
Cascade	N+2	N+2	N+2	N+2
Unit Count	133	133	223	223
Floors	15	15	15	15
1GHz Design				
Actives Used	**3**	**3**	**9**	**6**
BLE100	2	3	8	3
MB100	1		1	3
Power Supplies				
15 Amp	1	1	1	1
Power Draw (Amps)	2.0	1.8	6.1	4.0
% Power Savings		10.0%		34.4%
Cable Footage	**1,568**	**1,568**	**4,893**	**4,893**
Total Electronics	$ 1,644.84	$ 1,502.15	$ 4,172.51	$ 3,816.72
% Change to electronics		-8.7%		-8.5%

TABLE B – Greenfield High-Rise MDU

The higher output achieved using the GaN amplifiers allowed for the use of multi-output devices in fewer locations, reducing the number of active locations needed in each MDU.

An additional advantage of the GaN amplifier not captured in this design is that the GaN amplifier increases the number of floors that can be reached. In both the MDU A and MDU B designs, using an equivalent GaN amplifier in place of the designed GaAs amplifier, an additional three floors could have been reached, increasing the possible number of units passed by 17%.

BROWNFIELD MODEL

One of the more significant advantages that GaN provides is the ability to cost effectively upgrade existing 550 MHz systems out to 1 GHz. Where GaAs allowed a near perfect drop in upgrade from 750 MHz to 1 GHz, GaN provides this same ability for 550 MHz systems.

BROWNFIELD NODE	GaAs	GaN
Plant Mileage	28.25	28.25
Aerial	26.81	26.81
UG	1.44	1.44
Total Actives	128	128
Actives/Mile	4.5	4.5
House Count	509	509
HC/Mile	18	18
EXISTING DESIGN		
Actives Used	**128**	**128**
Cascade	N+9	N+9
1000 MHz Design	**GaAs**	**GaN**
Actives Used	**136**	**129**
Cascade	N+6	N+6
ACTIVES OVERVIEW		
% Actives Held Location	98%	100%
% Actives Held w/ Epak	91%	94%
% New Actives	6.3%	0.8%
TAPS & PASSIVES		
Total Taps	347	347
% Taps Held	91%	99%
Total Passives	93	93
% Passives Held	91%	95%
Poweing		
15 Amp power Supplies	10	10
Power Draw (amps)	98.9	95.8
% Power Savings		3.2%
CABLE		
Existing Cable	198,518'	199,585'
New Cable	1,779'	712'
% New Cable	0.9%	0.4%

TABLE C - Brownfield

Looking at Table C, GaN technology allows the MSO to increase bandwidth of their existing 550 MHz systems to 1GHz, while maintaining their active locations and housings. The MSO can maintain 95% of their current active housings, using an E-pack to upgrade the existing amplifiers.

Assuming that the current taps and passives in the 550 MHz system are capable of passing 1GHz, 99% of the tap faceplates as well as 95% of the passives were maintained.

The GaN amplifiers held 94% of the current amplifier locations with just an e-pack upgrade. While this is just 3% better then an equivalent GaAs design, the GaN amplifiers were able to reach the end of line (EOL) taps with just the addition of one new amplifier. The GaAs design added 6.3% new actives to be able to reach all of the EOL taps.

A WORD ABOUT OPEX

Operational Expenditures (OPEX) are frequently ignored by some operators when deciding upon equipment purchases. The annual recurring expense to maintain and operate the plant can be a significant contributor to profit margins.

Reducing OPEX is a key metric that should never be neglected. Table D below illustrates the OPEX savings for just a few key contributors in a network – powering, battery maintenance and active device maintenance. There are certainly other areas such as plant sweep & balance, but we have focused herein upon the three main categories stated.

Note the overall reduction in OPEX ranging from 13% to 17% due to incorporating the advantages of GaN technology into a Greenfield plant design.

OPEX - Greenfield	Urban Plus GaAs	Urban Plus GaN	Suburban GaAs	Suburban GaN
Mileage	2.72	2.72	9.67	9.67
Powering Cost				
# PS	1	1	3	3
Wattage	981.53	795.18	3218.82	2749.76
Kwh/Year	8598	6966	28197	24088
Cost/Kwh	$0.10	$0.10	$0.10	$0.10
Annual Power Cost	**$859.82**	**$696.57**	**$2,819.69**	**$2,408.79**
P.S. Maintenance				
Batteries	4	4	12	12
Battery Cost	$150.00	$150.00	$150.00	$150.00
% Annual R&R	25%	25%	25%	25%
Annual Battery R&R Cost	**$150.00**	**$150.00**	**$450.00**	**$450.00**
Active Device Maint. Cost				
Actives Count	7	5	31	29
Average Cost/Active	$900.00	$1,085.00	$800.00	$775.00
Annual Replacement %	2%	2%	2%	2%
Annual Replacement Units	0.14	0.1	0.62	0.58
Annual Material Cost	$126.00	$108.50	$496.00	$449.50
MTTR (Hrs)	1	1	1	1
Labor Cost/Hr.	$100.00	$150.00	$100.00	$100.00
Annual Labor Cost	$14.00	$15.00	$62.00	$58.00
Annual Active Device Maint. Cost	**$140.00**	**$123.50**	**$558.00**	**$507.50**
Total OPEX	**$1,149.82**	**$970.07**	**$3,827.69**	**$3,366.29**
OPEX/MILE	**$422.73**	**$356.65**	**$395.83**	**$348.12**
% Difference		**-19%**		**-14%**

TABLE D - OPEX

CONCLUSIONS

Upon review of this paper, it becomes evident that there are numerous advantages of deploying GaN technology. Perhaps the key advantage is that deploying GaN yields a proactive ability to "Go Green" as fallout without really having to try to do so.

Going Green: reducing the Carbon Footprint of networks is a key objective that will continue to become a desirable goal for all things in life over time. It is the right and smart thing to do. Some areas that will benefit are:

- Fewer active devices to purchase and maintain
- Ability to maintain more existing active devices in existing plant upgrades
- Ability to improve system performance
- Reduced deployment cost
- Reduced powering cost for the life of the network
- Reduced transportation costs
- Reduced OPEX costs for the life of the network

References:

Motorola White Paper: “Lowering the Cost of Fiber Deep Networks with Motorola Gallium Nitride (GaN) Technology”. October 2009

LIST OF ACRONYMS

CAPEX – Capital Expenditures
CCN – Composite Carrier to Noise
CIN – Carrier to Intermodulation Noise
CSO - Composite Second Order
CTB - Composite Triple Beat
EOL – End of Line
GaAs – Gallium Arsenide
GaN – Gallium Nitride
HEMT - High Electron Mobility Transistors
HFC – Hybrid Fiber Coax
HPM – Homes per Mile
LDMOS - Laterally Diffused Metal Oxide Semiconductor
MESFET - Metal Semiconductor Field Effect Transistor
MMIC - Monolithic Microwave Integrated Circuits
OPEX – Operating Expenditures
RFoG – RF over Glass
Si - Silicon
SiC – Silicon Carbide

STEREOSCOPIC DELIVERY OF 3D CONTENT TO THE HOME

Walt Husak
Dolby Laboratories, Inc.

Abstract

3D is enjoying a renewed revival in the theatrical market due to the commercial success of 3D films released over the last several years. The technology advances in the cinema coupled with similar advances in consumer electronics promise to provide affordable 3D experiences to the home. Of the many ways to experience 3D, stereoscopic delivery is the most viable method due to the availability of displays and known production techniques. The delivery method described in the following paper addresses a cost effective method to provide stereoscopic content to the home using a tiered approach.

INTRODUCTION

Broadcast Distribution of 3D

The high costs of 3D production favor a distribution model where a premium can be charged to consumers which would offset to the increased production and delivery costs. The most obvious initial 3D providers would be satellite and cable companies where 3D could be packaged as a premium service, sold as pay per view, or delivered as video on demand. Service providers subsidize the customer's set top boxes and recoup those costs through monthly subscription fees. Likewise, network infrastructure costs are absorbed by the providers and recouped through monthly subscription fees.

Given the small amount of content that will be produced in 3D, coupled with the high cost of producing that content, an initially very low cost approach to delivery of 3D would be highly desirable to the service providers. Ideally, the upgrade costs should approach zero while at the same time the operators could collect additional revenue.

There are two methods service providers could use to deliver 3D to the home: a frame compatible method or a 2D compatible method (sometimes referred as a service compatible method). The frame compatible method (e.g. side-by-side) has the advantages of being able to use the current network infrastructure including set top boxes, with an acknowledged penalty of reduced resolution in the initial roll out of 3D services. In the future, the lost resolution can be provided to new decoders by means of a parallel enhancement stream. 2D compatible systems (e.g. MVC[1]) offer the advantage that the transmission consists of a 2D version with an enhancement layer to provide the 2nd eye view. However, to receive and decode 3D, new set top boxes have to be deployed.

Frame compatible 3D video signals closely resemble a normal video signal so few changes are necessary to accept and retransmit the signal from a network perspective. Similarly, current set top boxes can pass the frame compatible signal along to a 3D display for viewing by the subscriber. In the future, the operator can decide to upgrade the plant and set top boxes to pass full resolution signals as part of a larger upgrade cycle.

2D compatible systems offer full resolution upon deployment but require substantial changes to both the network and the set top boxes. Current networks are not designed to accept and process full bandwidth 3D signals so new production and processing equipment is necessary. Likewise, existing set top boxes can only understand the 2D version of the signal and therefore new set top boxes would need to be deployed to receive

3D. Any additional revenue will be absorbed by deploying additional individual customer's set top boxes.

Survey of frame packing methods

The coding performance and image processing considerations of the various decimation and frame packing approaches are important considerations in the selection of the most appropriate method. The order of operations to create a frame compatible image is to take the left and right pair and decimate those images so that each image contains half the samples of the original image. The sub-sampled images are then packed together to form a frame compatible image that is the same size as the original left or right image.

Table 1 lists the various sampling methods commonly used to decimate stereo images in preparation for frame compatible formatting. The first column shows the sampling method and the second column shows the direction the pixels are decimated.

Table 1
Frame Compatible Sampling Methods

Sampling Methods	Decimation Direction
Column decimation	Horizontal
Line decimation	Vertical
Quincunx	Diagonal

Table 2 lists the various packing methods commonly used to create frame compatible images. There are several combinations that are illogical such as column decimation and line interleave packing or line decimation and side-by-side packing. Most frame compatible systems make use of sampling and packing in the same direction. For instance, one could use column decimation with side-by-side packing or line decimation with over/under packing. Quincunx sampling can be used with several packing methods including side-by-side, over/under or checkerboard.

Table 2
Frame Compatible Packing Methods

Packing Methods	Description
Column interleave	Every other pixel
Line interleave	Every other line
Checkerboard	Pixel & line interleave
Side-by-Side	Horizontal half image
Over/Under	Vertical half image

Analysis of frame packing methods

Beyond the obvious combinations of sampling and packing, there are operational and performance issues that need to be considered when deciding which methods and combinations should be used. This section will discuss the various performance and operational considerations. The packing method is most sensitive to operational issues such as video preprocessing in the video encoder and image post-processing in the set top box and display.

Both checkerboard and line interleave frame packing suffer from processing techniques such as filtering and resizing. In both cases, the processing causes inter-pixel contamination resulting in ghosting at best and complete loss of the stereoscopic effect at worst. Since it is difficult to predict and control image processing throughout the video path, these methods are poor choices for the frame packing method. Side-by-side and over/under are less sensitive to these processing techniques.

The next issue to consider is interlaced video and its impact on the sub-sampling process. Interlaced video by its nature is vertically decimated. The two vertical decimation methods applied consecutively (interlacing and line interleave) compound the problem by doubly decimating the video. As an example, a 1080i60 signal has 540 lines per field. Decimating the image further would reduce the vertical resolution to 270 lines – equivalent to QVGA. By its nature, interlacing introduces vertical aliasing making reconstruction from images that have been

vertically decimated much more difficult. Side-by-side is not affected by interlacing and checkerboard falls in between.

Coding performance is another consideration when selecting the most appropriate frame packing and method. Figure 1 shows the relative coding efficiency using MPEG-4 AVC[2] of several sampling systems. Column interleave sampling with side-by-side packing and line interleave sampling with over/under packing have the same coding efficiency relative to each other across a variety of bit rates. However, quincunx decimation with checkerboard packing requires more than twice the bit rate compared to side-by-side or over/under for the same quality (PSNR).

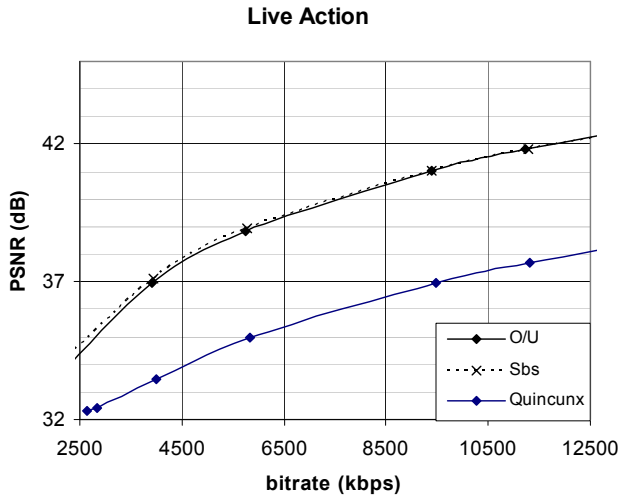


Figure 1
Relative coding efficiency of column interleave, line interleave and quincunx sampling.

It is clear from the analysis that side-by-side packing is the most appropriate base layer packing method for use with both progressive and interlaced formats. This is due to side-by-side being robust to interlacing, image processing and having superior coding performance to quincunx and the same (actually slightly better) performance than over/under.

One can also decimate in one format and pack in another. A popular method is to decimate in quincunx and pack in side-by-side. Figure 2 shows the performance comparison between side-by-side decimation and quincunx (checkerboard) decimation, both packed into the side-by-side format. For reference, normal AVC coding of 2D is also shown. The coding performance of quincunx decimation with side-by-side packing is lower than column decimation with side-by-side packing. At 10 Mbps, side-by-side decimation has a 2.5 dB performance advantage over quincunx decimation. When using the data from Figure 5, one sees that quincunx decimation with side-by-side packing has a coding efficiency that is superior to using quincunx decimation and checkerboard packing but inferior to side-by-side decimation and packing. In short, the coding efficiency of the combined method is roughly in between the efficiency of each method on its own. The mediocre efficiency is due to vertical and horizontal edges are no longer straight and require extra bits to code; also quincunx sampling with any sort of packing is sensitive to vertical resampling and color processing.

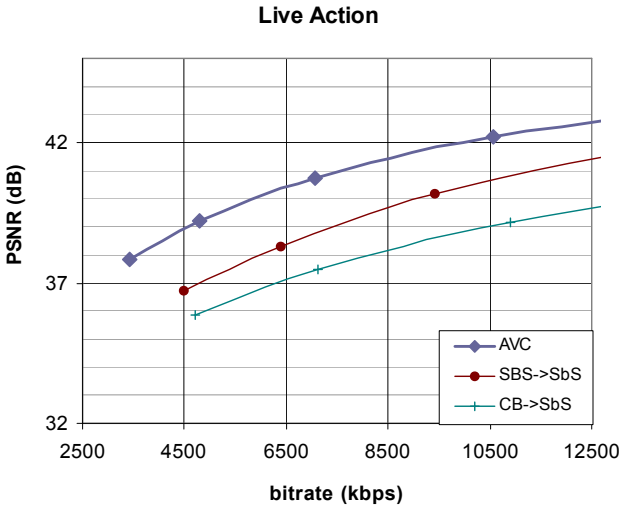


Figure 2
Relative coding efficiency of different sampling and packing formats

It should be noted that any frame compatible format has a decreased coding efficiency when compared to 2D video. This can be seen in Figure 2 where side-by-side sampling and packing (the most efficient frame packing method) and quincunx sampling with side-by-side packing both require a higher bitrate relative to the 2D AVC coding. This is due to the increased high spatial frequency image energy resulting from squeezing two images into the space of one image.

FULL RESOLUTION 3D

Full Resolution Frame Compatible

The ultimate goal is for content distributors to deliver full resolution stereoscopic signals to the home. As stated earlier, one method is the 2D compatible service which requires replacement of set top boxes in the home and an upgrade of network infrastructure. Frame compatible systems can also support full resolution by sending metadata that can recreate the full resolution using common layering techniques. Dolby has introduced such a system to meet the needs of the broadcasters.

Dolby's 3D system is a two tiered 3D delivery system that allows low cost initial deployment using a frame compatible base layer, with an available enhancement layer allowing a path to full resolution. Side-by-side has been chosen for the reasons discussed above, as well as widespread acceptance by 3D display vendors.

Figure 3 shows a functional overview of Dolby's 3D Full Resolution Frame Compatible delivery solution. A stereo pair is multiplexed into two frame compatible images one using one set of pixels and the other using the complementary pixels. The

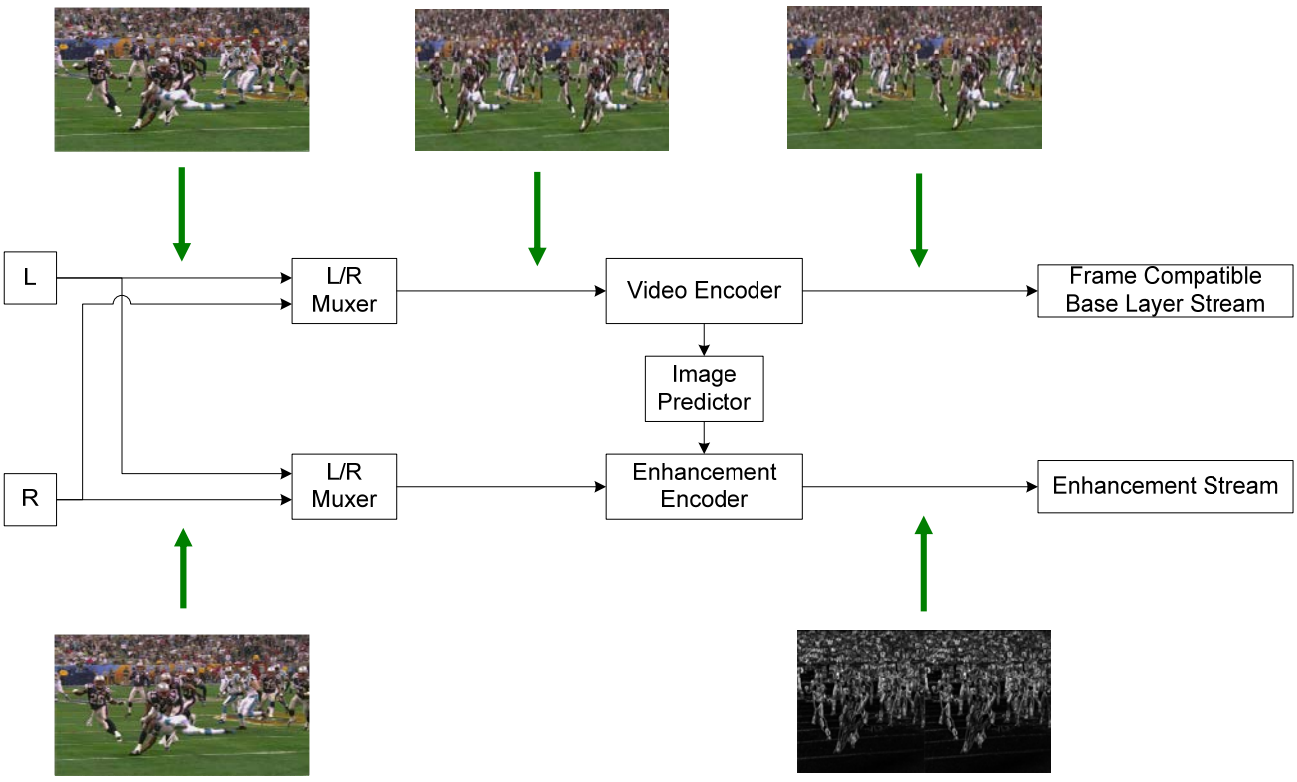


Figure 3
System Overview

first frame compatible image is compressed using MPEG-4 AVC as if it were normal video image set. The complementary image is used as the basis for the enhancement encoder. The enhancement encoder uses information from the base layer encoding process to predict the enhancement layer. By making use of redundant information in the base layer, the total amount of data in the enhancement layer is greatly reduced.

The frame compatible base layer selected by Dolby is the side-by-side method for decimation and the packing. Dolby offers the option to use a variety of pre-decimation low-pass filters in order to provide the optimum performance for a given piece of content and bitrate. The side-by-side packed video stream is compressed and transmitted using the standard service provider work flow. In the case of legacy MPEG-2[3] video delivery, only the base layer would be transmitted allowing the use of legacy MPEG-2 set top boxes. For operators that have enabled MPEG-4 AVC (H.264), the base layer would be encoded using AVC with an option to also encode an enhancement layer.

In a video compression system, most of the coding efficiency is realized by using prediction techniques to recreate pixels. Using a split filter system (complementary filtering of the high frequency component from the base layer) or a difference signal from the base layer and compressing it using standard coding techniques suffers from several fundamental weaknesses. A simple high frequency split system suffers from the two video codecs (high and low frequency respectively) operating open loop relative to each other. Unless a very high bitrate is used for the enhancement layer, the recreated pixels will not be phase coherent with the source nor will the bit depth be adequate for the high frequency information. A simple differencing system - where the enhancement layer is subtracted from the base layer - is limited in performance gain due the simplicity of the prediction. In addition, production tools such as keystone, floating windows and occlusions cause a significant amount of energy to be coded as residual data.

Dolby's full resolution method predicts the enhancement layer from the base layer and makes use of redundant information between

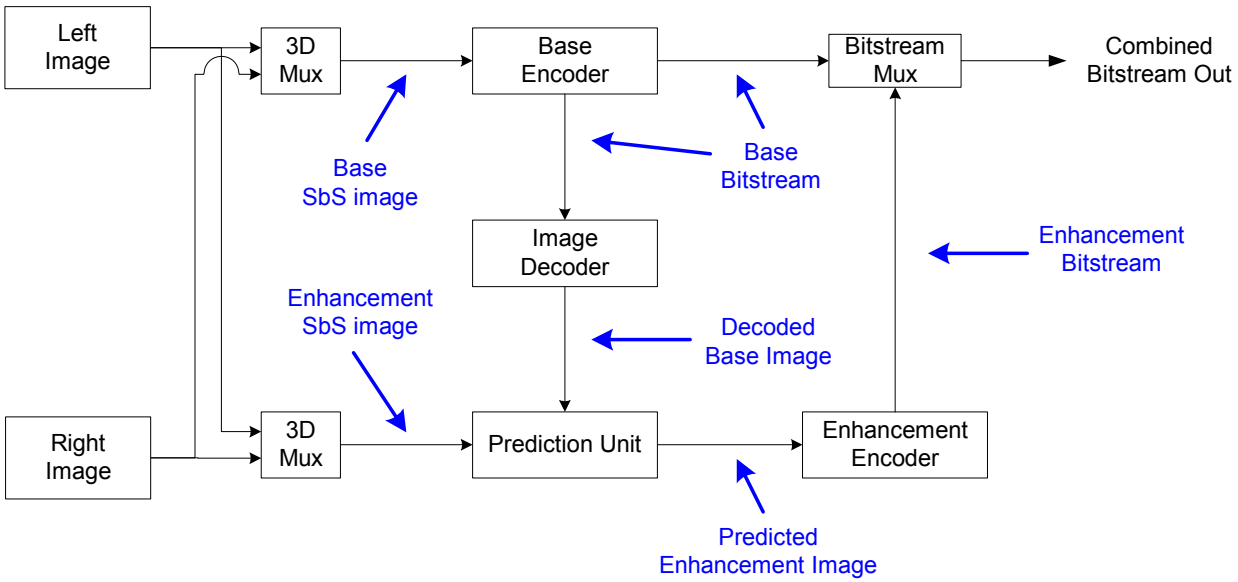


Figure 4
Dolby prediction system

layers to reduce the resulting bit stream. Figure 4 illustrates how the prediction between layers is accomplished. A stereo pair is presented to a pair of 3D multiplexers. The multiplexers filter, decimate and format the stereo images into the side-by-side format. The base layer uses one set of pixels from the original image set and the enhancement layer uses the complementary pixels. The base layer is then encoded using a standard video encoder. The resulting base bit stream is applied to the bit stream multiplexer and is also decoded locally.

The locally decoded base image is then used to predict the enhancement image. The base and the enhancement side-by-side images are very similar to each other due to both being taken from the same original images but merely offset by one pixel. The predicted enhanced image largely contains the differences between the base side-by-side image and the enhanced side-by-side image. The enhanced side-by-side image is then coded and the resulting bit stream is combined with the base layer bit stream for delivery to the decoder.

An important point in using the locally decoded image is the results of coding decisions made by the base encoder are automatically applied to the enhancement layer. This overcomes the weaknesses of using two separate open loop codecs for the base and enhancement layers. Figure 5 shows the performance gain using a predicted resolution enhancement system. At 7.5 Mbps, the enhancement system adds 0.4 Mbps while increasing the quality by 3.25 dB. As a percentage of the original bit rate, 5.4% overhead yields a doubling in video quality.

The enhancement layer can be delivered as a compressed stream along with the base layer by including MPEG-4 structures called Network Abstraction Layer (NAL) units that are specific to the enhancement layer. Legacy decoders should ignore the enhancement layer NAL units and decode the base layer as if it were a standard 2D video stream, and output the side-by-side image. Another means to deliver the enhancement layer is to use a secondary video stream with its own PID within the MPEG-2 transport stream[4]. Decoders that are enabled to decode the Dolby solution will extract the enhancement layer and decode the data to recreate the original full resolution video.

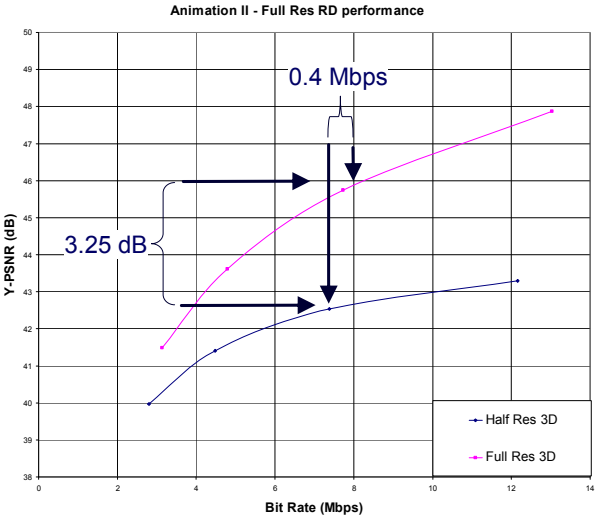


Figure 5
Enhancement layer performance

TESTING

Objective Performance

An important feature of the Dolby 3D system is the compression efficiency. A more efficient video compression system allows content to occupy a smaller part of the service multiplex than a compression system that is less efficient while still maintaining the same quality. Another way to consider the effects of a more efficient encoder is a higher quality image can be transmitted within the same bitrate.

This section will summarize tests that were performed. The following tests were conducted using 23,000 frames from four

sequences. Three sequences were live action and the fourth was an animation. All tests were performed using identical quality for both eyes in order to have an accurate comparison between methods.

Figure 6 shows the relative performance of the full resolution system using Peak Signal to Noise Ratio (PSNR). PSNR is a method of testing widely used for comparisons between different bitrates or toolsets that takes the Mean Squared Error (MSE) of each pixel and averages the information across the image as a root mean square.

The data has been normalized for easy comparison between delivery methods. The first bar is the 2D equivalent delivery bitrate and has been fixed at 100%. Not surprisingly, a 3D simulcast (one channel per eye) is twice the data rate of the 2D signal. The side-by-side coded data is 35% greater than the 2D bit rate due to the increased high frequency content resulting from squeezing two images into the space of one image. The added enhancement data is 6% more bitrate for a total of 41%. The overhead for the MVC signal is 81% more than the 2D signal. The MVC bitrate overhead is highly dependant on content and can range from 40% for animations to as much as 90% for live action material. The ability of MVC to use inter-view prediction is based on how well the views in the stereo image pair are correlated.[5]

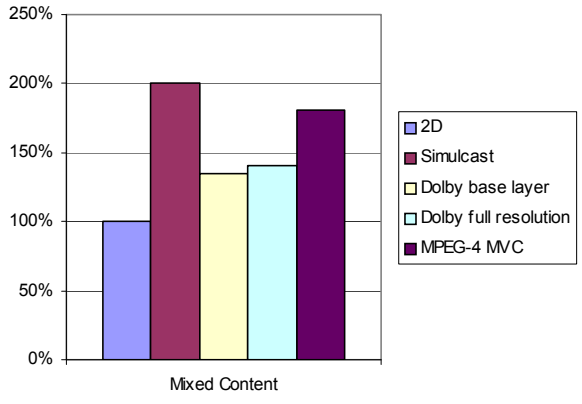


Figure 6
Relative performance of the Dolby full resolution 3D system

Subjective Performance

PSNR provides the engineer a simple and rapid test for comparing similar codecs, tools and content. It is difficult to use PSNR across substantially different content or coding systems due to the different artifacts that may manifest themselves specific to those codecs. For instance, it is valid to use PSNR to compare a number of AVC based codecs but mixing a wavelet codec and a block based codec such as AVC would limit the functionality of the PSNR metric. An additional shortcoming of PSNR is the inability of PSNR to consistently track a real viewer's Mean Opinion Score (MOS) when they are rating the quality of a subjective test across a variety of content. Nevertheless, PSNR is a simple test that is widely understood in the image processing community without having to run complicated subjective tests for each codec, bitrate and piece of content.

Dolby performed a series of subjective tests to understand the real world performance of stereoscopic delivery systems. The test used ITU-R Rec. 500[6] as a reference for designing the test. Modifications were made to the procedures since Rec. 500 did not thoroughly address stereoscopic subjective testing. The tests were conducted as a double blind quality rating test using MOS values obtained from both expert and non-expert viewers. The test was broken down into two stages. The first stage was a ranging exercise that compared 2D broadcast bitrates with 3D broadcast bitrates. The second stage used the results of the first stage as a baseline and compared several different coding techniques to understand what broadcasters may expect when deploying stereoscopic delivery systems.

The first stage was performed by presenting the viewers with a clip coded at multiple bitrates and the viewer was told to select the quality that most closely represented the target quality for their delivery system. The test was conducted twice – once for 2D content and once for 3D content. The data was coded as a standard AVC 3D simulcast (one channel per eye) with the 2D content represented by the left eye view. Each viewer's MOS was tabulated and the results were normalized to the 2D MOS values.

Figure 7 is a chart of the results from the 2D to 3D comparison. Due to the normalization, the 2D data is fixed at 100% of the bitrate. Intuitively, one expects the 3D results to be approximately twice the 2D results. The results from the subjective tests show that viewers did not find coding artifacts in 3D as objectionable as coding artifacts in 2D

In some cases – such as the animation – the 3D simulcast actually required fewer bits for the simulcast transmission than the 2D transmission. The movie sequence and the concert sequence showed slightly higher bitrates for 3D on the order of 40% and 25% respectively. The football sequence is an anomaly but is included in the chart for completeness. The right eye contained a source artifact that was not seen during the 2D presentation due to using the left eye as the reference. Subsequent testing has shown the Football sequence behaving similarly to the other sequences. In addition, several other clips were tested that also showed less than 50% overhead for 3D simulcast over 2D.

The effect of 3D content scoring higher for a given bitrate is commonly referred as "stereo masking". This phenomenon can be seen in Digital Cinema[7] where the maximum bitrate for 2D delivery to the theater is the same as the maximum bitrate for 3D delivery even though the two views are sent as two completely separate image streams.

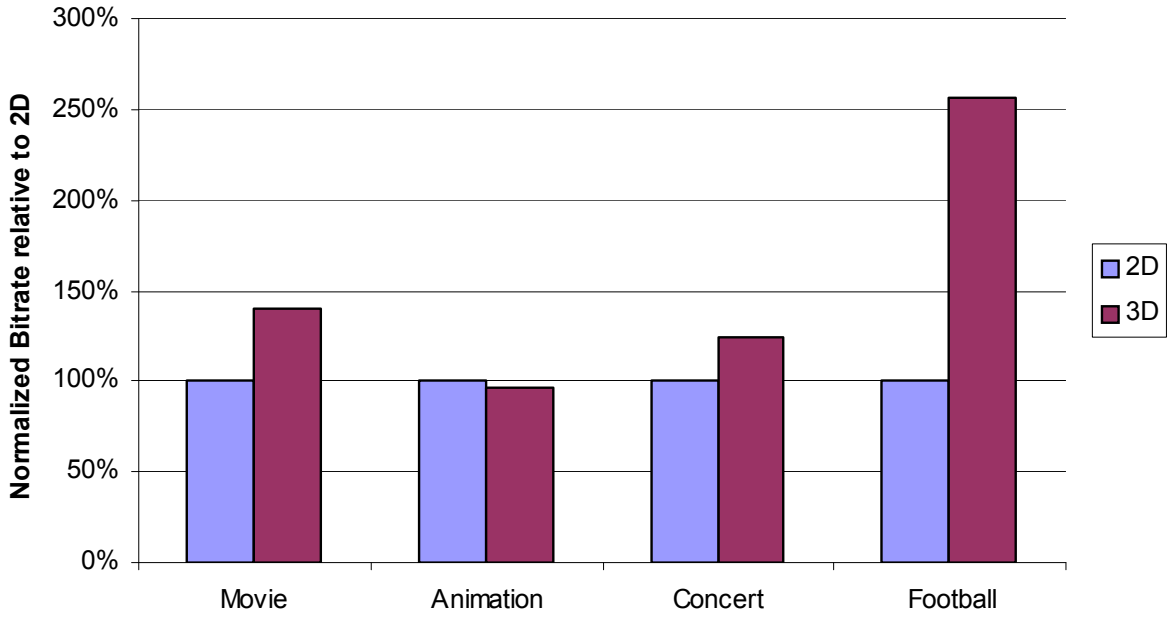


Figure 7
Subjective comparison of 2D and 3D codings

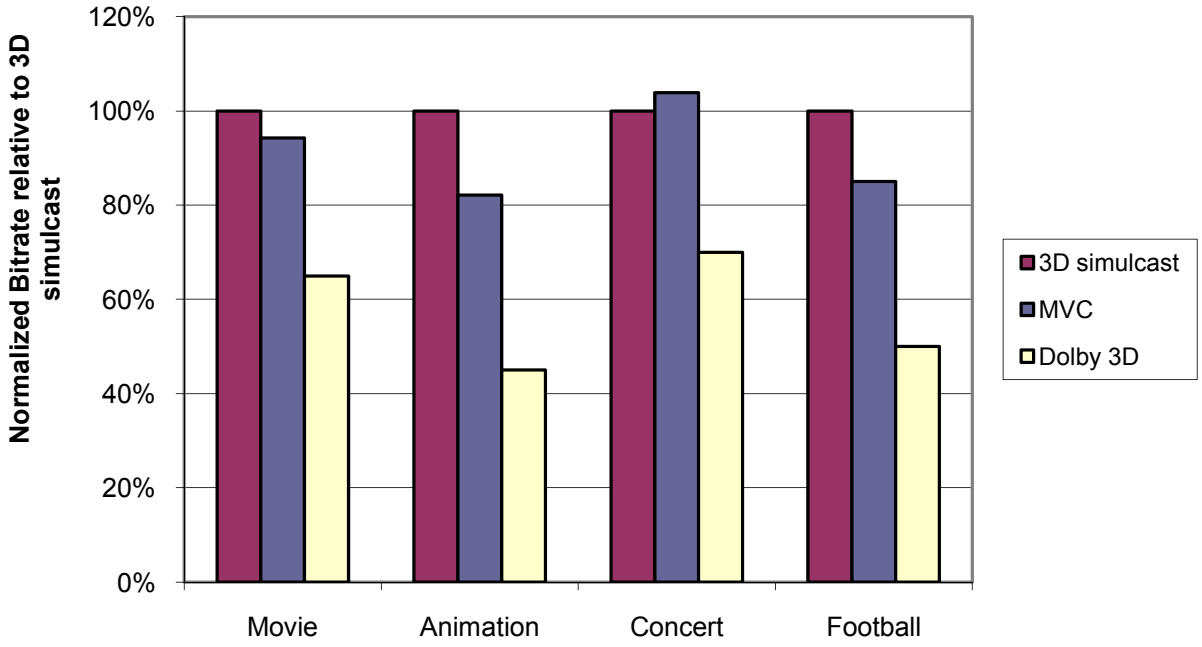


Figure 8
Subjective comparison between 3D coding techniques

Stereo masking creates an interesting dilemma for 2D (or service) compatible systems. The most critical aspect of the 2D compatible systems such as MVC or 2D+Delta is the ability to extract a 2D signal from the service. Forgetting for the moment that most content producers have stated 2D productions will be completely separate from 3D productions; the coding strategy for the entire stereo signal must use a 2D bitrate that meets the quality needs of current broadcast. This means one cannot use stereo masking to tailor their 2D compatible bitrate to minimize the impact to their service channel.

As an example, sports programming would be the most challenging content to deliver. This is because sports programming is by its nature, live content. The second view needs to use as much as 90% of the left eye's data rate to send equal quality video to both eyes. One could send asymmetric quality between the two eyes (e.g. sending higher quality to one eye and lesser quality to the other eye), but the effects of eye dominance between viewers is not well understood. People that are left eye dominant would be well served while people that are right eye dominant would receive a sub-standard image.

Figure 8 shows the results of the second stage of the subjective tests. The second stage was conducted as a Double Stimulus Comparison Scale (DSCS) using the results from the first test as the baseline reference. The three systems compared were a 3D simulcast where each eye is coded separately, a 2D compatible system represented by MVC and the Dolby full resolution frame compatible system. Again, the results were normalized – this time to 3D simulcast.

As expected, the MVC system requires nearly the same bitrate as the simulcast except when coding animations. The concert footage actually scored higher that simulcast due to the content. The lights flashing and the stage background caused significant differences

between the two eyes making the prediction between eyes difficult to achieve. This resulted in most frames being coded as two separate bit streams with little interdependency. While this particular clip was more stressful in that regard than a typical concert, the differences in lighting due to the flashing and spinning lights will limit the amount of prediction between views.

The viewer MOS scores showed the frame compatible system having equivalent quality with substantially lower bitrates than either simulcast or MVC. The bitrates were around 50% of the 3D simulcast which from stage one we know is just slightly higher than bitrates used for 2D services. One point to note is the lower bound of the subjective test did not exercise the video codec. In other words, the values shown in this paper are conservative numbers for delivery of 3D.

SUMMARY

In this paper we examined delivery of 3D content using 2D compatible systems and frame compatible systems. Frame compatible systems allow a broadcaster to deliver 3D using existing set top boxes and network infrastructure. There are several methods of sub-sampling and packing to create the frame compatible image, although several of them suffer from operational issues. Side-by-side offers a simplified approach that codes with the same compression efficiency as over/under albeit without operational limitations imposed by interlaced video. Quincunx sampling offers no additional benefit but adds unneeded complexity.

A means to migrate to full resolution using predicated layering techniques was discussed allowing the operator to deploy a backward compatible system serving existing set top boxes with frame compatible 3D and new set top boxes with full resolution. The method shown allows the operator to upgrade their set top boxes and network infrastructure over time. The use of advanced prediction specialized for frame compatible 3D overcomes weaknesses such as open loop codecs and limitations in complementary filter systems. The relative overhead for the enhancement layer is between 5-10% and also increases the measured by quality over 3 dB.

Finally, objective and subjective test results were discussed. Stereo content was shown to require substantially lower bitrates than intuitively imagined dues to stereo masking. Furthermore, 2D compatible systems were shown to have a significant Achilles Heel in regard to needing to fix one eye to equivalent 2D quality while at the same time requiring high enhancement bitrates for the second eye due to eye dominance. The enhanced frame compatible system required substantially lower bit rates than MVC and 3D simulcast while delivering equivalent quality.

BIBLIOGRAPHY

[1] G.J. Sullivan, et al, "Text of ISO/IEC 14496-10:2009/FDAM 1 Constrained baseline profile, stereo high profile, and frame packing arrangement SEI message," Doc. N10707, London, UK, July 2009

[2] ISO/IEC 14496-10 - MPEG-4 Part 10, "Advanced Video Coding"

[3] ISO/IEC 13818-2 - MPEG-2 Part 2, "Video Coding"

[4] ISO/IEC 13818-1 - MPEG-2 Part 1, "System"

[5] W. Gish & C. Vogt, "MVC compression coding for 3D applications", presented at the 2009 SMPTE technical conference October 28, 2009.

[6] ITU-R Recommendation BT.709-11, "Methodology for the subjective assessment of the quality of television pictures"

[7] SMPTE 429-10-2008, "D-Cinema Packaging — Stereoscopic Picture Track File"

SYSTEM MANAGEMENT IN AN EBIF WORLD

Kevin Taylor, Joshua Seiden, Jeff Calkins
Comcast Cable

Abstract

In this paper we will examine the challenges that face the MSO community in deploying and managing an end-to-end EBIF system. The EBIF system architecture and its challenges will be reviewed. The issues related to the EBIF Video Data Path and the Data Signaling Path will be examined. Also, guidelines for EBIF application developers will be presented for an application to be well-behaved on a MSOs network. Finally, EBIF diagnostic applications will be proposed to aid in the monitoring and management of the EBIF Video Data Path and Data Signaling Path.

INTRODUCTION

The cable industry and their programming partners have a unique opportunity to enhance the video product offered the consumer with EBIF applications. For these enhancements to be successful there needs to be a focused effort to prepare the MSO video path and data signaling path to support the requirements of the EBIF delivery and data return infrastructure. Without careful management of the cable operator's system resources, the customer experience will not be acceptable and the end-to-end system will exhibit instability. In this paper, we will examine the challenges of enabling the MSO EBIF video delivery path, managing the data signaling path in an EBIF enabled system, and propose EBIF applications that will be useful in diagnosing end-to-end system problems.

In discussing this topic, there are two primary areas of the cable operator's infrastructure that will be discussed.

EBIF Enhanced Video Path – The EBIF Enhanced Video Path includes the origination of the EBIF application and its data, the delivery of the EBIF application to MSOs headend and hubs, and the final delivery to the consumer premise equipment (CPE). This path includes all encoders, groomers, multiplexers, ad splicers, encryption devices, QAM modulators, and transmission channels.

Data Signaling Path – The Data Signaling Path includes both the out-of-band (OOB) forward data channel (FDC) and the return data channel (RDC). There are primarily three classes of technology used in the Data Signaling Path: ANSI/SCTE-55-1, ANSI/SCTE-55-2, and DSG/DOCSIS.

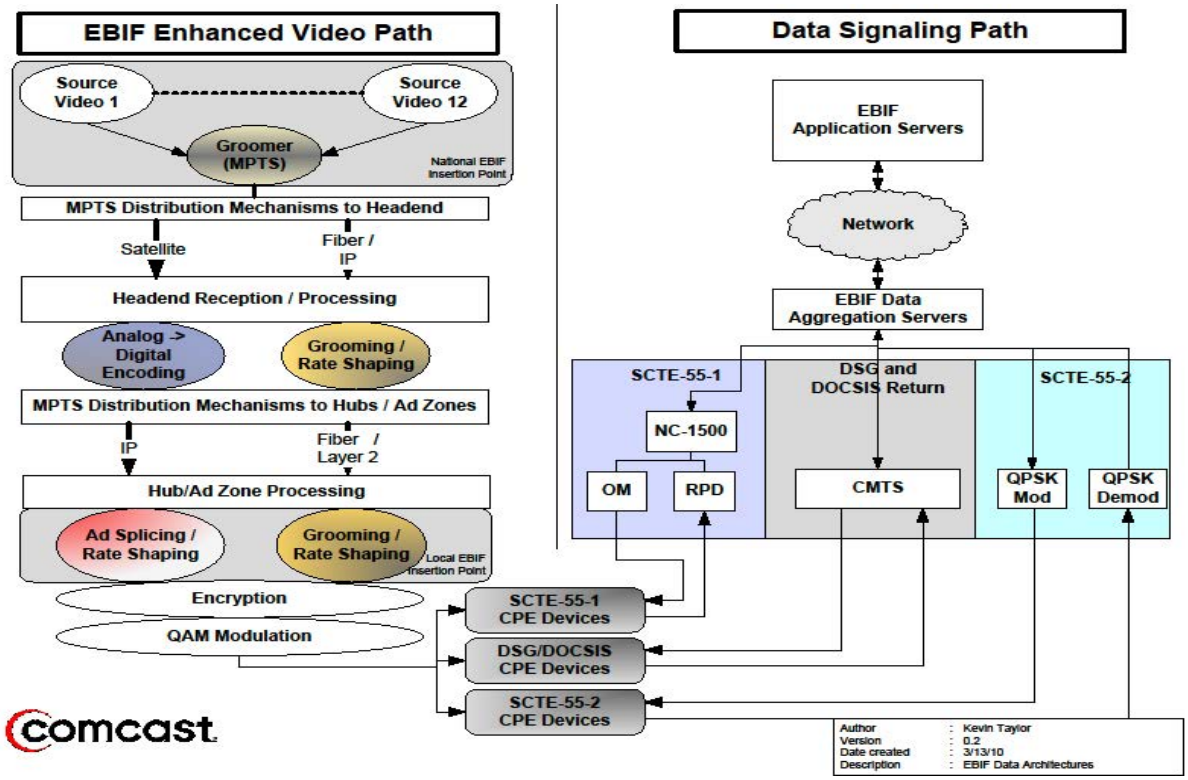


Figure 1 – End-to-End System

EBIF ENHANCED VIDEO DELIVERY SYSTEM ARCHITECTURE AND CHALLENGES

The delivery of EBIF enhanced video can be done by one of two originators – a broadcaster or an MSO. Broadcasters may choose to enhance their national broadcast with EBIF and deliver it with the largest possible footprint, or they may choose to deliver EBIF applications and data separate of the video to achieve a localized feed. MSOs will also originate their own applications by multiplexing EBIF applications and data at the headend to enhance their services. MSOs may choose to use bound applications for content they own, but they may also deploy unbound applications that enhance their products, i.e. Caller ID to the TV.

Challenges exist that may force a broadcaster or MSO to localize the delivery of EBIF applications and data. Current EBIF deployments face the fact that different EBIF User Agents may present EBIF applications differently. This issue has been addressed to some degree by the adoption of EBIF I05 across MSOs, particularly for those MSOs involved in the Canoe project. EBIF I05 does ensure a baseline of functionality that works across all user agents, but beyond that baseline the concept of localizing the EBIF application and data helps to address variances.

One specific concern is that for VOD Telescoping applications, VOD asset IDs are not consistent across servers. This complicates the VOD Telescoping application by forcing the application developers to validate that the asset ID in the application is the correct asset on the VOD server. Both of these issues will be addressed in the upcoming I06 version of the EBIF specification, however, until I06 User Agents are deployed widely, these are issues that must be manually managed by the broadcaster and the MSOs. The best way to manage these issues prior to I06 being

available ubiquitously will be to localize the delivery of EBIF applications and data.

National vs. Local Delivery

The delivery of EBIF applications and data across a large footprint can take two forms – national or localized. A broadcaster who is trying to reach maximum footprint on their delivery may choose to insert the application and data at the origination site or uplink. The architecture to do this is seen in Figure 2.

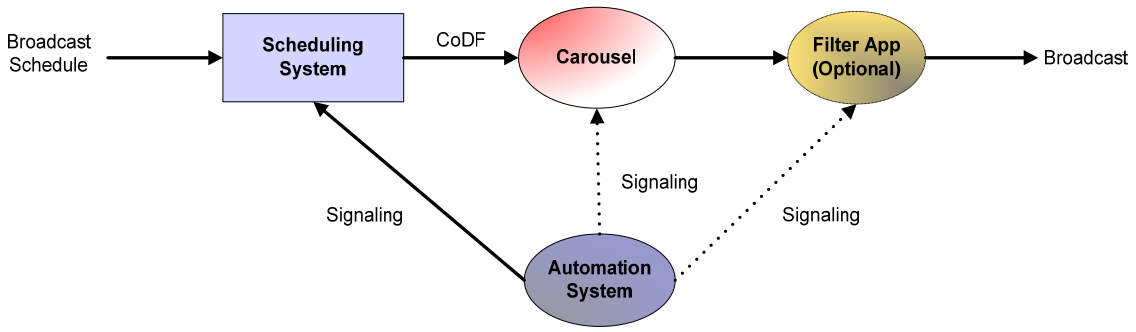


Figure 2 - EBIF Broadcast Center Architecture

Broadcasters need to accomplish two primary objectives when enhancing their feeds with EBIF. First, they must make sure that any applications inserted in the broadcast are not shown during advertisements. While unbound applications may be allowed to overlay any video source, bound applications cannot overlay a third-party's ad spot. To do this, one can take two different approaches. A filter application can be deployed to take signaling from the automation systems and block an application coming from a carousel. Also an interface into the existing scheduling system can be implemented from the automation system to the scheduling system, which in turn can shut off the EBIF application in the carousel.

All broadcasters must also be aware that their control of the signal ends when the broadcast leaves the uplink or origination site. At the headend, MSOs may groom services in a multiplex for more efficient transmission on the cable plant. An MSO may be throttling bandwidth of a service without the broadcaster's knowledge. For example – it is likely that an MSO would take an SPTS from a broadcaster and then groom it into an MPTS for distribution on their network. In this case, a service has the potential to have its overall bandwidth reallocated to achieve the MSOs most efficient use of bandwidth. When this happens, any service with extra bandwidth, EBIF PIDs, other data PIDs, alternative audio, etc. may be subject to this bandwidth reallocation. At this time, the only means to save bandwidth on a service is to take away from the video. Therefore, MSOs may choose to lower the video quality in order to fit the service into an MPTS or channel. For example, EBIF bandwidth is typically limited to less than 200kbs. Lowering the bandwidth on the EBIF application increases the application's launch latency, but currently the bandwidth allocated for the EBIF data cannot be modified. No tools exist today to throttle EBIF data or application PIDS in an MPEG stream while leaving others unchanged. In addition, many EBIF data streams are synchronous to be frame-accurate with the video stream – in this event, even if the data PID could be throttled; doing so would harm the context of the application.

The 200kbs limit may not sound like much, but as more and more services become enhanced with EBIF, multiplexes will have to adjust to carry the additional data. If an MSO is carrying 12 services in a multiplex, and all services are enhanced with EBIF, a potential 2.4Mbps of EBIF data overhead exists. This additional bandwidth may force the cable operator to lower the bandwidth allocated in the video in order to make room, thus degrading picture quality.

As more and more applications are deployed, content providers will begin to see the need to localize their application data on national broadcasts. Differences in EBIF User Agents, VOD systems, and the desire to localize data being presented to the user will cause the content providers to rethink their EBIF distribution model. Implementing different versions of an application or sending localized packages of data over satellite will quickly become unscalable. The only option in this case is to separately deliver the EBIF component of a service terrestrially and then multiplex it back into the video locally. This will not only allow for more efficient utilization of satellite bandwidth, but will also enable EBIF applications to be localized directly to the settop box population.

National Delivery of Localized Data

As EBIF applications become more prevalent, broadcasters and MSOs will need to localize the delivery of applications and data. This will not only lessen the impact of some of EBIF's current limitations, but will also allow the application to become localized. With localized delivery, an application can present data that is meaningful to the consumer at a much granular level. For example, News/Weather/Sports ticker applications can provide local data, advertisements can link or click-to-call to the consumers local store, etc. To accomplish this, it becomes necessary to delivery the EBIF application and data separately from the video and audio. The Broadcaster can continue to deliver the video content via satellite (or terrestrially), but the carouselled EBIF data is sent terrestrially to a remote groomer at the headend. At the headend, the EBIF PID is groomed back into the video source and the interactive service is modulated for the plant. (See Figure 3)

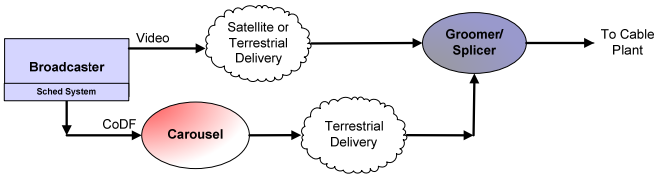


Figure 3 –EBIF Local Data Delivery for National Broadcast

Local Delivery of EBIF

MSOs are increasingly deploying EBIF applications to enhance their service offerings to customers. EBIF can be leveraged to provide unbound applications such as guides, caller ID, news and weather tickers, etc., or can be used to enhance local programming (vote and poll during the local news, etc.). (See Figure 4)

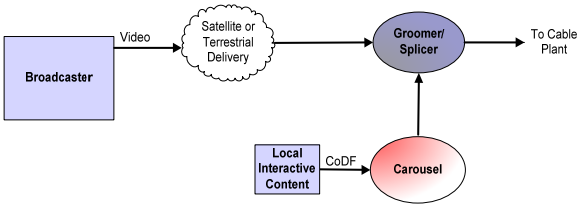


Figure 4 – MSO EBIF Insertion

In this case, the distribution of the EBIF application is more complicated depending on the content that the app is overlaying. The MSO may have a local carousel or may receive an EBIF stream from an aggregator. Content will usually be sent to the carousel via the CoDF format. The carousel will then send the EBIF data to a groomer based on

either a manual schedule or schedules received via CoDF. The data can then be groomed into an existing video feed or the out-of-band channel at the edge.

Many challenges exist when inserting an application locally. The largest challenge that has yet to have been solved is scheduling an application so that it does not overlay on National Ad spots. In a case where local spots are being inserted, Q-tones or digital ad insertion signals (SCTE-35) can signal a carousel to suspend an application. However, from the perspective of a local broadcast, a national spot looks as if it is part of the video, i.e. there is no signaling to determine when the national ad starts and stops. This challenge will have to be solved to automatically prevent the overlay of national spots, but until then, MSOs must schedule their carousels appropriately to prevent this from occurring.

The issue of a MSOs rights to overlay existing content with a third-party application is in a nascent stage. Content providers, MSOs, and Application Developers will have to solve these issues in the near future.

DATA SIGNALING IN AN EBIF ENABLED SYSTEM

Description of the OOB and Return Channel

In the North American MSOs cable systems, there are three main types of out-of-band (OOB) signaling and return path. These three types are SCTE-55-1, SCTE-55-2, and DSG/DOCSIS. In preparation for our discussion of the data signaling problems in an EBIF enabled environment, we will do a quick summary of these technologies and a review of the main characteristics of these OOB and Return Path Signaling technologies.

ANSI/SCTE-55-1 – SCTE-55-1 is used in cable systems supporting the DigiCipher settops and host devices. This approach uses an out-of-band (OOB) channel that has a data rate of 2.048 and return path employing Aloha. It is based on technology developed by General Instruments (Motorola).

ANSI/SCTE-55 -2 – SCTE-55-2 is used in cable system supporting the PowerKey settops and host devices. This approach uses a DAVIC OOB and return. It is based on technology developed by Scientific Atlanta (Cisco).

DSG and DOCSIS – Data Signaling Gateway (DSG) is a protocol for sending one-way message through the DOCSIS channel. DSG/DOCSIS is increasingly used for CPE device signaling and the footprint of this technology is expected to continue to grow.

Another of the primary challenges of launching EBIF in the existing legacy systems is the use of the legacy data signaling paths for the EBIF data return. The current data signaling paths carry guide data, VOD data, polling, code download, and other application data. The downstream signaling paths are already run near capacity, and the return signaling paths are also reaching their saturation point with the addition of new interactive applications. This is the environment into which the EBIF applications are being introduced. This requires a careful engineering and management of the data signaling path to enable successful launch of EBIF applications.

EBIF applications bring with them a set of network loading characteristics that have not been seen previously in cable operator networks. These network loading models include time synchronized events and channel synchronized events, as well as the

existing network loading models. The large scale of planned EBIF deployment and the ability to create highly synchronized events has the ability to create very large scale network events that can overwhelm the MSOs network without careful engineering of both the EBIF applications as well as careful engineering of the MSO network. Another important fact is that the same EBIF application may be running on devices that are part of each of the three types of data signaling path simultaneously requireing engineering to the lowest common dominator network characteristics. This implies that these EBIF applications need to be designed to work on the most constrained of the networks. It should be remembered that the two legacy networks (SCTE-55-1 and SCTE-55-2) have been deployed for more than 12 years. The process of laying down the EBIF infrastructure on top of the existing legacy data signaling path should be viewed as a retrofitting of the legacy data signaling path requiring a disciplined engineering approach to layering EBIF applications onto the existing data signaling paths. This retrofitting should include the two following areas of focus:

- EBIF Application Developers Guidelines
- Signaling Path Management and Monitoring

EBIF Application Developers Guidelines

There are several guidelines that application developers should follow in their design of applications that are destined to run on the MSO Data Signaling network.

Data Signaling Path Capacity Constraints - EBIF Applications are run on MSO networks that are constrained in capacity and already carry a significant network load. Therefore the EBIF developer needs to take great care in designing the data signaling path loading characteristics of the EBIF application to carefully use the data signaling path resources.

Data Signaling Path is a Shared Resource – The MSOs data signaling path is shared amongst several applications that need low latency responses. These applications include VOD and SDV. It is important to the customer experience that there is always capacity to service these types of requests.

Event Timing - One of the most important design points of the EBIF application will be what kind of synchronization is created as EBIF applications respond. MSO data signaling paths are engineered expecting randomized return events from individual devices. If large numbers of CPE devices are synchronized to respond at the same time due to either application characteristic or synchronization with the EBIF enhanced programming, the data peak will be more than the MSO network can accommodate. Therefore the designers of EBIF applications need to take into account the network characteristics of the MSO network to create EBIF applications that are well behaved. EBIF applications need to be designed to randomize the response timing as much as possible to smooth peaks.

Protocol Design - The legacy signaling paths are well-behaved when the offered load consists of smaller return messages spread over time. The EBIF application provider needs to keep the return message size as small as possible and randomized over time.

Network Cell Boundaries - EBIF application developers need to be aware of the underlying data cell size for the networks that the EBIF application will be running on. The EBIF developer should take cell boundaries into account to minimize the number of upstream cells required to deliver return data back to the application server. In the SCTE-55-1 protocol, each upstream cell requires a downstream acknowledgement. For example, an application going from requiring one cell upstream to requiring two cell upstream not only doubles the amount of upstream

bandwidth, but it also doubles the downstream bandwidth to acknowledge the upstream cells. There are step functions in how the application uses network resources, and application developers need to know where these step functions are and work to minimize resource utilization and not cross the step boundaries unless necessary.

To some developers, these recommendations may seem extreme. However, a single poorly design EBIF application can significantly impact customer experience and MSO revenue associated with VOD, Caller Id, Guide response times, and other applications running on the subscriber's CPE. A suite of well-behaved EBIF applications can greatly extend the data signaling path to handle additional applications adding greater value to the cable product offering.

Signaling Path Management and Monitoring

The current SCTE-55-1 and SCTE-55-2 signaling paths are now more than 12 years old which is a lifetime in terms of technology lifecycles. Also, these channels are significantly constrained as compared to the DSG/DOCSIS channels. The legacy signaling paths require greater care and management. However, the majority of cable CPE devices are managed from these legacy signaling paths. Therefore it is imperative that the cable operator carefully manage these signaling paths to allow a positive customer experience for the new EBIF applications that need to use these communications channels.

Forward Data Channel Bandwidth Management

Over the past two years there has been on-going work in Comcast markets to better understand and manage the data signaling path. One of the most surprising results of this work was the discovery of the importance of management of the forward data channel and the fact that the forward data channel was one of the first bottlenecks that needed to be remediated. The results of the changes were quite surprising.

- Code down load time was cut in a third due to eliminating the data peaks that were overrunning the network
- Warehouse staging time was significantly cut. Prior to the network changes about 20% of the settops needed to be re-hit from the billing system. After the network change, the number of re-hits was less than 2%.

From the work done in the Comcast markets, several guidelines have been developed to guide the management of the forward data signaling path. The cable operator needs to manage their forward signaling path by:

- Have a forward signaling path bandwidth budget and manage to the budget. The bandwidth budget needs to take into account both average and peak values. In the Comcast example from above the significant improvements were generated by improved management of the data peaks.
- Work with equipment and application vendors to make sure data sources conform to average and peak data budgets. Prior to the advent of EBIF applications there was enough head room in the data stream to absorb data peaks. However, with the increased load on the network this head room is no longer present and these peaks cause data loss impacting the customer experience.
- Data packing on the Ethernet interfaces needs to be as efficient as it can be. Another problem that was discovered was that when many UDP packets containing a single MPEG packet are processed, the network processing equipment can be overrun. The downstream QPSK runs much more efficiently when the UDP packets are full of MPEG transport packets.

- Monitor the forward path for data loss on both the network and cable interfaces. Much of the data delivered to the data signaling path is delivered in UDP packets which do not guarantee delivery. Also, any network design issues, network changes, or increased traffic can result in data loss on the network feeding the forward data signaling path causing customer impacting events. There are many events that can impact the data signaling path and the only way for the cable operator to pro-actively address these issues is to invest in the appropriate monitoring infrastructure.

The following graphs represent data captures from a Comcast market on the forward data signaling path. In this example, the forward data signaling path bandwidth budget is currently allocated to the ALOHA network proxy (1000kbps), the controller (750kbps) and the guide stream (150kbps). The system streams are dynamic and peak usage can exceed the QPSK downstream modulator's capacity, causing it to drop packets. Because the three systems share the downstream QPSK bandwidth, budget exceeding transients by any of the systems can be operationally or customer affecting. Managing the ALOHA downstream bandwidth is critical for EBIF applications that primarily send upstream messaging; because each upstream cell has a downstream bandwidth cost tied to "acking" each upstream cell.

A mid-week downstream graph of the interactive network proxy that is allocated 1000kbps is captured below. (See Figure 5)

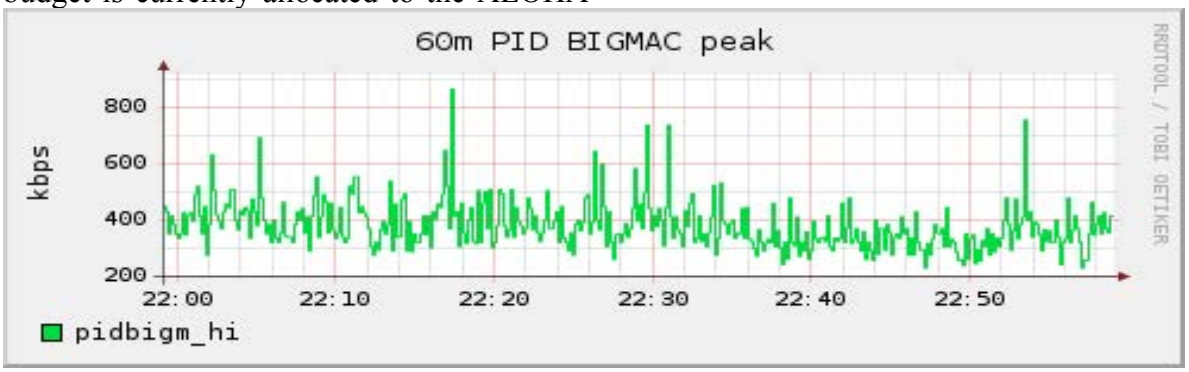


Figure 5 – ALOHA Network Proxy Downstream Data

The controller bandwidth is currently allocated 750kbps of the downstream QPSK bandwidth. The controller streams are moderately bursty. The green waveform (upper waveform) identifies transient 100msec peaks whereas the black waveform (lower waveform) shows an average one second load. (See Figure 6)

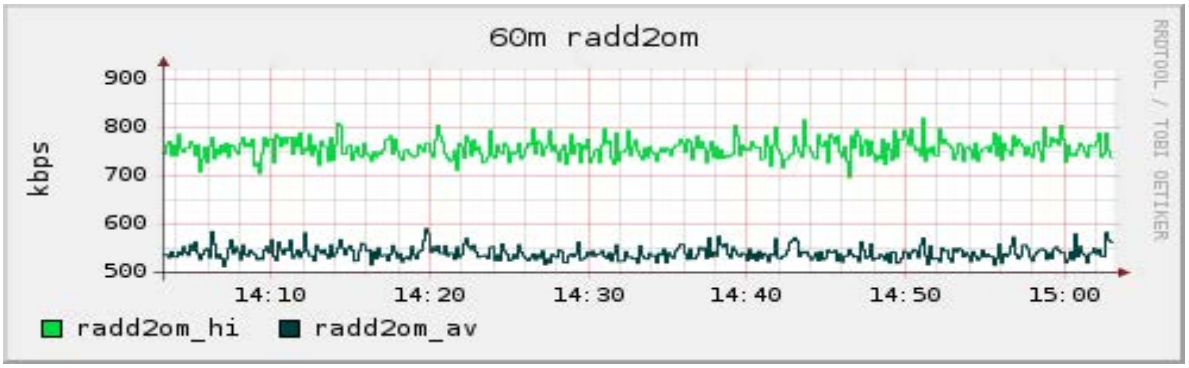


Figure 6 – Control System Downstream Traffic

The controller bandwidth consists of the system control PIDS and code download (CDL) object carousels. The system control PIDS are made up of PAT (PID 0), CAT (PID 1), the NETWORK PID (PID 777) carousel for transmitting channel maps and EMM (PID 1503 and 1504) used to transmit settop control messaging. The following 60 minute production graphs the controller system PID peak transients. (See Figure 7)

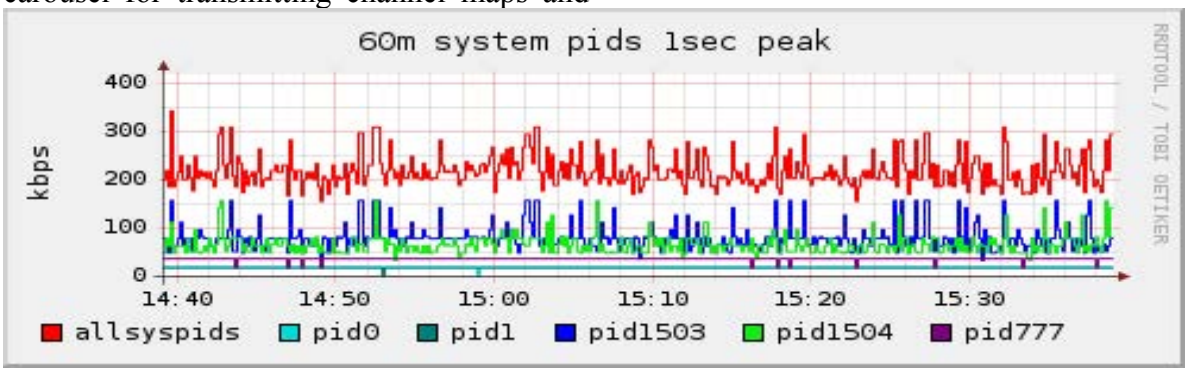


Figure 7 - Downstream Traffic on all System PID's

The largest downstream QPSK bandwidth component of the controller is the OOB settop CDL settop objects and their associated PMT's. (See Figure 8) Management of the OOB CDL carousels provides the greatest opportunity to re-allocate downstream bandwidth to EBIF applications and other ALOHA interactive applications. In some production environments the CDL carousels can be turned-down or turned-off during peak usage times.

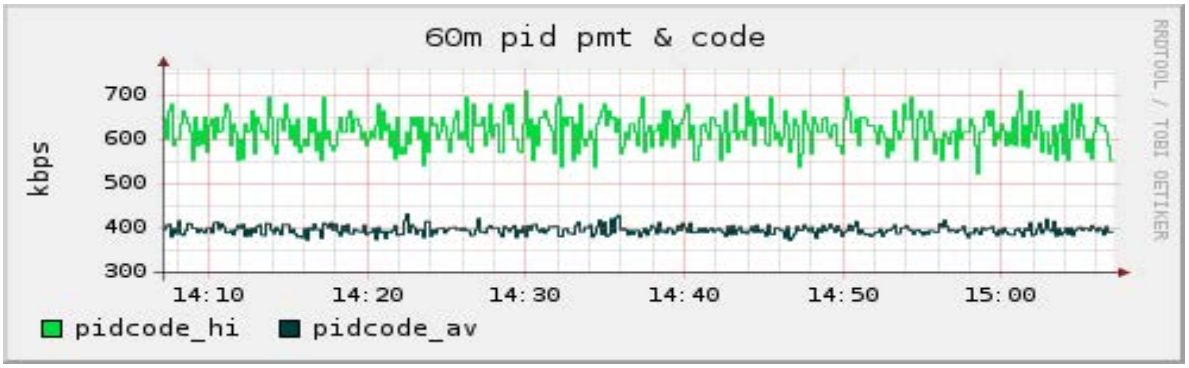


Figure 8 – Code Download Downstream Data

The guide stream presents another opportunity to manage the transient downstream QPSK OOB usage in favor of interactive applications. The guide stream does not transmit a significant amount of data, but the guide stream has a tendency to generate large bandwidth transients considerably larger than the steady state bandwidth. The following graphs shows the average bandwidth load in black and the transient bursts that hit the downstream QPSK and compete for limited downstream QPSK resources. (See Figure 9)

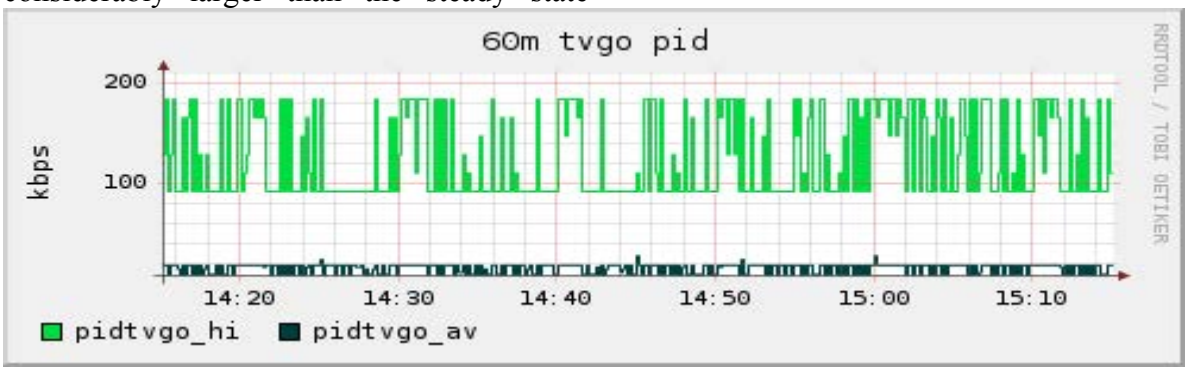


Figure 9 – Guide Stream Downstream Data

The instantaneous sum of all of the data sources on the network feeding the down stream signaling path need to be less than or equal to the capacity of the forward data channel. If this limit is not enforced then simultaneous peak usage can exceed the downstream QPSK's capacity causing data loss, network instability, and customer impacting events.

Return Data Channel Bandwidth Management

The management of the Return Data Channel is always a challenge. To help understand some of the complexities of the return path, we will examine some of the characteristics of the ANSI/SCTE-55-1 return path. The ANSI/SCTE-55-1 return uses the Aloha protocol.

The ALOHA protocol requires an ALOHA network proxy device to acknowledge (ack) upstream cells and the settop interprets a missing acknowledgement as a collision, requiring the settop to "retry" the cell retransmission. In the absence of receiving an "ack" the settop client will retry sending the cell as many as six times (6x). It

is important to note every upstream cell has an associated downstream "ack" bandwidth cost. If the ALOHA network is running properly, the odds of a settop successfully transmitting a cell upstream is approximately 99.9%. The ALOHA network proxy accounts for "successful" and an estimate of "retry" cells. The following graph from a production upstream path hosting approximately 1000 settops shows 24 hours of "good" (acknowledged cells) and "retry" cells, providing an opportunity to monitor the efficiency of each upstream path or each settop. (See Figure 10)

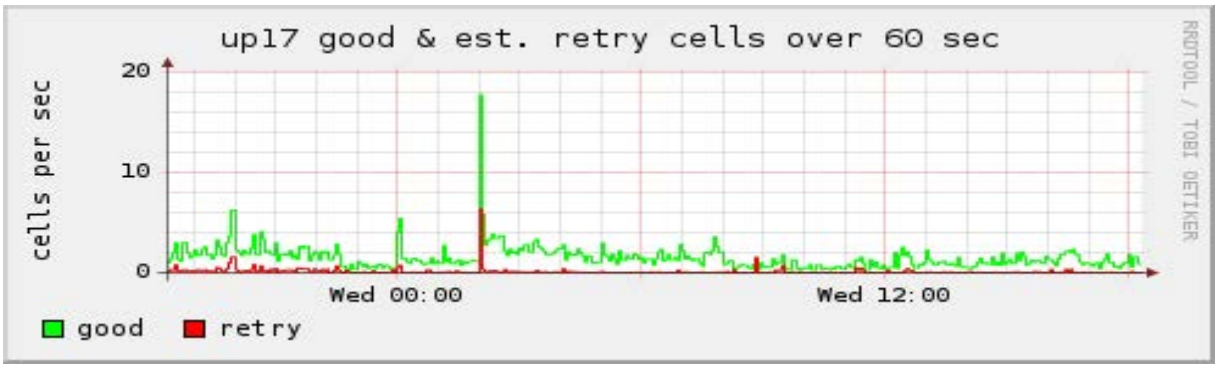


Figure 10 – Upstream Traffic Retries

The SCTE-55-1 return path demodulator is capable of demodulating approximately 92 cells per second (cps) whereas an ALOHA network reaching 50 cps begins to spend more time sending "retry" cells than transmitting successful cells in a timely manner. Operating the ALOHA network above 50 cps equates to operating the network inefficiently or delaying the successful transmission of settop messaging.

The following graphs each 1 second peak identified in each 60 second window. (See Figure 11)

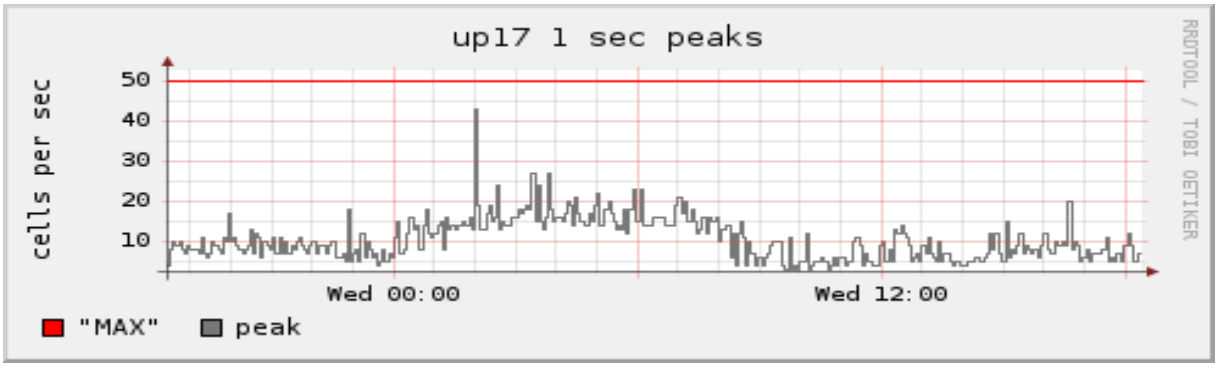


Figure 11 – Upstream Traffic Peaks

ALOHA capacity on the Comcast networks is currently constrained by the downstream QPSK pipe. The following section identifies opportunities to manage

and reallocate downstream bandwidth resources towards for EBIF and interactive application usage.

ALOHA Traffic Models

The initial ALOHA traffic model was based upon randomly arriving single cell messages. Early VOD ALOHA traffic patterns were well represented by modeling this behavior and the resulting upstream settop capacities have worked adequately. However, new feature rich VOD applications and EBIF application driven traffic loads doesn't conform to the original basic ALOHA model assumptions with synchronized messages and longer settop messages which tend to collide. The new network bandwidth loading model that includes EBIF applications and other enhanced application needs to be developed to help predict network performance, settop capacity constraints, and application limitations.

The forward data channel traffic has to be managed as the peak loads are a function of the applications and three independent systems using the downstream (controller, ALOHA network proxy, and guide stream). Maximum node settop counts and downstream settop counts are a function of peak application usage.

The ALOHA network is a scarce resource and all aspects of the network will need to be carefully engineered and managed to support all of the applications that need to use its resources. It will be necessary to compare ALOHA interactive applications and determine the best use of scarce ALOHA bandwidth. As new applications use the ALOHA network, it will be necessary to manage both ALOHA network efficiency and usage as well as to manage node / upstream settop counts.

EBIF ENHANCED VIDEO PATH SYSTEM MONITORING AND DIAGNOSTICS APPLICATION

The MSO digital video infrastructure has grown organically over the past 12 years. This growth has been driven by the following factors:

Individual cable operator priorities - Each cable operator manages their business and operation according to their priorities. This causes a significant difference in timing of technology deployments and how they are implemented.

Individual market and headend sizes – The size and scale of a market and its associated headends has a significant influence on the types of technologies and the timing of those technologies.

Technology Timing – Technology is always in a state of evolution. So the timing of a technology deployment into a market has a significant impact on its current and latent capabilities.

Configuration Options – The technology that is deployed is always optimize for the current set of priorities. A good example of this is in the quest for optimum picture quality, operators will remove unused data from a multiplex. When EBIF data is added in the multiplex at the programmers' uplink, the EBIF data may not pass onto the cable operator's plant due to configuration choices made in the headend.

These four factors greatly influence the capabilities of any particular market and headend deployment. As the EBIF infrastructure and EBIF applications are rolled out, each of the existing markets will have a unique combination of technologies that may or may not be compatible with the EBIF deployment.

One of the primary challenges of fully enabling the EBIF ecosystem will be proving out that each headend and EBIF enable service is actually passing the required data.

As noted above, there are several reasons that a particular headend can be configured in ways that prevent the EBIF data to be passed. In the current MSO community there are many services, in many QAM's, in many ad zones, in many markets. Just as an example of the sheer number "pipes" (an individual service on a QAM located on the edge of the MSO network), if we assume a 1000 ad zones, 70 QAM's in each ad zone, and 10 services per multiplex. This would equate to 700,000 individual "pipes" between the EBIF application provider and the CPE population.

There is a very large number of "pipes" between the EBIF application source and the CPE equipment spread across the North American cable footprint. The only way to build confidence that this many "pipes" actually pass EBIF data is to create an EBIF application and data server that will function as an aggregation point for diagnostics information on the proper passage of EBIF signaling through the many "pipes" of the MSO networks. There are two variations of the approach that should be taken. One approach is from the MSO view of the network, and the second approach is from the programmer's view of the network.

MSO EBIF Diagnostics Application

Goal: The MSO EBIF Diagnostic Application will be used by the MSO to prove out the delivery of EBIF applications from the local EBIF insertion point to the User Agent on the CPE's across all of the headends in the cable operator's markets.

Insertion Point: The MSO EBIF Diagnostic Application will be inserted at the local EBIF insertion point.

Scope: The MSO EBIF Application will be used to prove that EBIF application and data will pass from the local insertion point through the operator's network and all the associated headend equipment to the User Agent running on the CPE.

EBIF Application Functionality: When the subscriber tunes to a service with the MSO EBIF Diagnostics Application, it will signal the diagnostic server that the settop has run the EBIF application. Service and CPE data will be returned to the diagnostics server. The EBIF diagnostics application will have neither a user interface nor any interaction with the user.

Data Collection: The EBIF Diagnostics Server will schedule the insertion of the EBIF Diagnostic Application and gather the response data. The diagnostics server will map the response to the services and nodes within the cable operator infrastructure to build a map of services in nodes that are not responding to EBIF applications.

Programmer EBIF Diagnostics Application

Goal: The Programmer EBIF Diagnostics Application will be used to confirm proper delivery of the EBIF applications from the programmer's insertion point to the participating MSOs CPE's.

Insertion Point: The Programmer EBIF Diagnostics will be inserted at the programmer's EBIF insertion point.

Scope: The Programmer EBIF Diagnostics Application will be used to prove proper delivery of the programmer's EBIF applications to the target population on the MSOs network. It will be used to help identify "pipes" that are not properly configured to pass EBIF applications and data from the programmer. This will need to be a collaborative activity between the programmer and MSO for this activity to be successful.

EBIF Application Functionality: When the subscriber tunes to a service with the Programmer EBIF Diagnostics Application, it will signal the diagnostic server that the settop has run the EBIF application. Service and CPE data will be returned to the diagnostics server. The EBIF diagnostics application will have neither a user interface nor any interaction with the user.

Data Collection: The EBIF Diagnostics Server will schedule the insertion of the EBIF Diagnostic Application and gather the

response data. The diagnostics server will map the response to the services and nodes within the cable operator infrastructure to build a map of services in nodes that are not responding to EBIF applications.

CONCLUSIONS

The advent of EBIF enable networks is an exciting opportunity for cable operators to add a richer customer experience and enable greater advertising revenue. However, there is a cost that comes with this opportunity which is the cost of re-engineering and enabling the EBIF Video Data Path and actively managing the Data Signaling Path.

Areas for more research

- SCTE-55-1 - Current underlying assumption around random data distribution is changing. The data distribution is not random and, even worse, is moving towards higher level of synchronization. New models need to be put together that predict network performance and capacities.
- How do SCTE-55-2 and DSG/DOCSIS stand up under the changing network loads which are moving towards highly synchronized events and non-random distributions.

TAKING THE DOCSIS UPSTREAM TO A GIGABIT PER SECOND

John T. Chapman, jchapman@cisco.com
Cisco

Abstract

Fiber is here. The competition is deploying it. The cable operators are field trialing it. Is this the end for the HFC Plant? Is this the end for DOCSIS?

The on-going challenge for the HFC plant has always been its upstream bandwidth. This white paper focuses on the upstream and how technology can drive the upstream direction to a data capacity of 1 Gbps and beyond.

Included is an optimized upstream solution for existing plants consisting of six carriers for North American 5 – 42 MHz HFC plants and a ten carrier solution for European 5 to 65 MHz HFC plants.

INTRODUCTION

Imagine the access plant of the future 50 or even 20 years from now. What would it look like? Did the following thoughts come to mind?

- all fiber
- all IP
- 1 to 10 Gbps per subscriber
- symmetrical or close to symmetrical bandwidth
- multiple 100 GE connections to the backbone

Much of this technology exists today or will very soon. So why not skip all the intermediate steps and just build the network of the future today?

Here is where the technology vision must intersect with the business vision. Ultimately, what does it cost? And if it costs too much, is there an interim strategy that gets close to the performance needed but at a much lower price point?

This white paper addresses this issue with the specific focus on the upstream. The downstream direction is important as well, but that will be the subject of another white paper.

The first section of this white paper defines four upstream spectrum options for increasing upstream bandwidth. They are:

- Low-split
- Mid-split
- High-split
- Top-split

The next section of this white paper looks at the operational and technical challenges with these options. Topics include:

- Deep fiber
- Spectrum allocation
- Transition Plans
- Analog TV
- Legacy OOB
- Legacy tuners
- CM upstream power amp
- Aeronautical interference
- Optical node technology
- Amplifier technology
- In-line equalizers
- RF transmission path
- Plant power
- Professional installation

The final section then covers the cost of each of the four techniques and compares it to a fiber build.

UPSTREAM SPECTRUM OPTIONS

There are at least four general approaches for providing upstream bandwidth with respect to the RF spectrum. These are shown in Table 1 and explained below. Note that these are variations on the classic mid- and high-split frequency plans long used by cable operators. The data capacity is calculated in this white paper based upon assumptions stated in this white paper.

This section focuses on the definition of these options. The pros and cons of each approach are covered in later sections.

Name	Upstream Frequency Range	RF BW	Data Capacity
Low-Split	5-42 MHz	37 MHz	120 Mbps
	5–65 MHz	60 MHz	210 Mbps
Mid-Split	5–85 MHz	80 MHz	300 Mbps
High-Split	5–200 MHz	195 MHz	770 Mbps 1 Gbps
Top-Split	> 1 GHz	1 GHz	2.2 Gbps 3.6 Gbps

Table 1 – Upstream Spectrum Options

Low-Split

A low-split system is what is in use today. In the USA, low-split refers to 5 MHz to 42 MHz with downstream spectrum beginning at 54 MHz. In most of Europe, low-split refers to 5 MHz to 65 MHz with downstream spectrum beginning at 85 MHz.

Table 2 below shows the upper bound of the upstream spectrum and the lower bound of the downstream spectrum for various countries.

The spectrum below 20 MHz is quite noisy and is generally not used by DOCSIS. 26-28 MHz is the CB (Citizen's Band radio) and is usually avoided.

Region or Country	Band Split
North, Central and South America	42/54 40/54 30/52
China, Korea, Philippines, Thailand, Singapore, Australia	65/85
Japan, New Zealand	55/70
India, Malta, Eastern Europe	30/48
Western Europe, Ireland, United Kingdom	65/86

Table 2 – HFC Band Splits

There are many ways that this spectrum is used. The upstream spectrum may be shared between:

- DOCSIS 1.0, 1.1, 2.0, and 3.0 CMs,
- legacy set-top box (STB) out-of-band (OOB) return path signaling channel,
- legacy TDM voice
- Telemetry (power supplies, nodes, amps, sweeps).

It is reasonable to assume that the legacy TDM voice either no longer exists or will not exist in a full DOCSIS scenario. The legacy OOB channel and telemetry usually can be hidden in spectrum below 20 MHz that is not used normally by DOCSIS.

So how much data capacity can DOCSIS extract out of the current upstream path? To answer that, it is useful to look at a max-fit six carrier (for North America) scenario that is emerging. The European scenario could contain up to 10 carriers. This is shown in Table 3.

#	From (MHz)	To (MHz)	BW (MHz)	Modulation	Style	Primary Usage
10	61.4	64.6	3.2	64-QAM	ATDMA	D3.0 (Europe only)
9	54.8	61.2	6.4	64-QAM	ATDMA	D3.0 (Europe only)
8	48.2	54.6	6.4	64-QAM	ATDMA	D3.0 (Europe only)
7	41.6	48.0	6.4	64-QAM	ATDMA	D3.0 (Europe only)
6	35.0	41.4	6.4	64-QAM	ATDMA	D3.0, D2.0
5	28.4	34.8	6.4	64-QAM	ATDMA	D3.0, D2.0
4	23.6	26.8	3.2	16-QAM	TDMA	D1.1, D1.0
3	20.2	23.4	3.2	QPSK	TDMA	D1.0, DSG
2	13.6	20.0	6.4	64-QAM	SCDMA	D3.0
1	7.0	13.4	6.4	64-QAM	SCDMA	D3.0

Table 3 – Maximum Upstream Spectrum Usage

DOCSIS 3.0 CMs that are being deployed are capable to transmitting four return path carriers. Those carriers are being pushed to their maximum modulation of 64-QAM with a 6.4 MHz RF bandwidth. The data capacity of each carrier, assuming 10% overhead, is approximately 27 Mbps. So, in theory, four carriers will allow slightly more than 100 Mbps upstream performance.

DOCSIS 2.0 CMs do not understand bonding, but they do understand the higher order modulation that DOCSIS 3.0 CMs use. Thus, DOCSIS 2.0 CMs can share the same carriers that DOCSIS 3.0 uses.

Table 3 suggests that two ATDMA carriers are enough for DOCSIS 2.0. This is based upon a limited deployment of DOCSIS 2.0 CMs. This also limits load balancing algorithms and allows the SCDMA channels to focus on optimizing performance with only DOCSIS 3.0 CMs.

DOCSIS 1.1 CMs only support carriers up to 16-QAM modulation and 3.2 MHz bandwidth. In order to allow DOCSIS 2.0 and 3.0 CMS to run at optimum speeds, a completely separate carrier is required for legacy DOCSIS 1.1 CMs. Assuming about a 10% overhead, the maximum data capacity of a DOCSIS 1.1 carrier is 9 Mbps. This would be a 5th carrier.

Then there are STBs with embedded cable modems (eCM) in them. It turns out that these STBs do not get the same preferred home wiring treatment. The STBs are often located behind several layers of splitters. The result is that the attenuation of the reverse path in the home is too much for the eCM in the STB to transmit on.

Thus, when DSG is enabled, it does not always work. DSG (DOCSIS Set-top Gateway) is a protocol that places the STB downstream OOB signaling protocol into an IP tunnel that is managed by DOCSIS, and uses the upstream for signaling advanced services such as video on demand (VOD) and pay per view (PPV).

The practical solution has been to use a dedicated carrier that runs at 1.6 MHz bandwidth and with a modulation of QPSK. The result is an upstream carrier that is more tolerant of noise and can work when the CM is operating a maximum power. The throughput of this carrier is about 2 Mbps. This would be a 6th carrier.

DOCSIS 1.0 CMs support the same modulation as DOCSIS 1.1 so they can share

the same carrier. (5th carrier) However, since DOCSIS 1.0 was the first version of DOCSIS available, there are DOCSIS 1.0 CMs out there that do not behave well on a plant. Many of them are from manufacturers who are no longer in business and the software is out of date.

In general, cable operators have tried to eliminate the DOCSIS 1.0 CMs from their network. When they still exist, and they cause problems, the easiest solution is to put them on a dedicated upstream so they do not interfere with the other CMs. This would be a 6th carrier.

Summing these values together, the upstream capacity of a 5-42 MHz spectrum is (4 * 27 Mbps) + 9 Mbps + 2 Mbps ~= 120 Mbps.

The European upstream has an extra 22 MHz of bandwidth. The most aggress use of upstream spectrum would be to assume three 6.4 MHz carriers and one 3.2 MHz carrier. That would allow two sets of bonding groups at 4 channels per bonding group. The total data capacity would be 119 + 94.5 ~= 210 Mbps

Mid-Split

There have been various definitions of mid-split. Some earlier mid-split networks had an upstream frequency range of 5 MHz to 108~116 MHz. This white paper is going to use the newer definition of mid-split with a frequency range defined by DOCSIS 3.0.

DOCSIS 3.0 has an upstream frequency range of 5 MHz to 85 MHz and comes up just below the FM band of 88 MHz to 108 MHz. Downstream spectrum starts at 108 MHz.

To compare data capacity, lets assume the 5 MHz to 42 MHz spectrums remains the same. 44 MHz +/- 3 MHz should be avoided as explained later (section ref TBD) to avoid adjacent tuner issues. That would define the additional spectrum as 47 MHz to 88 MHz. This is an additional 41 MHz of spectrum. (oddly enough, this is the square of 6.4).

41 MHz could handle an additional six 6.4 MHz carriers (6 * 6.4 MHz = 38.4 MHz). If one pushed it, a 3.2 MHz carrier might also fit, although it could not be used by legacy 1.1 CMs because the operating frequency would be too high. The resulting additional data capacity would be 175 Mbps (6.5 * 27 Mbps). The total upstream data capacity, when including the baseline of 120 Mbps, would then be approximately 300 Mbps.

High-Split

There have been various definitions of high-split. Some earlier high-split networks had an upstream frequency range of 5 MHz to 162~174 MHz [FSN]. This white paper is going to use a different definition of high-split that is motivated by picking a frequency range that can support 1 Gbps of data payload.

High-split is not currently defined in DOCSIS. The proposed frequency range for high-split in this white paper is 5 MHz to 200 MHz. Downstream spectrum would start at 258 MHz. The split of 258/200 gives the same diplex filter shape factor as a 54/42 split (54 divided by 42 = 258 divided by 200 = 1.29). This means the filters in the CPE will have the same complexity as existing designs.

An alternate definition would be to define 5 MHz to 20 MHz as a legacy band for non-DOCSIS use (OOB, telemetry), and 20 MHz to 200 MHz for DOCSIS use.

Since there are no CMs that can drive this new frequency range, new one will have to be built. That means that different modulations could be used.

To estimate the upstream data capacity, two scenarios are suggested.

- ATDMA, 6.4 MHz carriers, 64-QAM, 47 MHz to 200 MHz

- OFDM, 50 MHz FFT blocks, 256-QAM, more advanced FEC, 33% overhead for cycle prefix, pilot tones, and FEC.

For ATDMA, the spectrum would support 24 carriers ((200 – 47) / 6.4) for an additional data capacity of approximately 650 Mbps (24 carriers * 27 Mbps/carrier). When added to the 120 Mbps baseline, the result is 770 Mbps.

For OFDM, the presumption is that with more work and with an improved FEC, that the modulation could be improved enough to support a level of 256-QAM. The data capacity of three OFDM FFT (Fast Fourier Transform) blocks would be approximately (3 * 50 MHz * 8 bits/hertz * 66%) is 750 Mbps.

The total upstream data capacity, when including the baseline of 120 Mbps, would then be approximately 870 Mbps. If the OFDM was applied to the 5 to 50 MHz range, the bit rate would be 1 Gbps.

Top-Split

Top-split refers to placing the upstream spectrum above 1 GHz. There are no real standards or proposals here. One concept is to carve out 1 GHz to 1.2 GHz for MoCA (MoCA D channels start at a center frequency of 1150 MHz and go to 1500 MHz), use 1.3 GHz to 1.8 GHz for upstream spectrum, and reserve 2 GHz to 3 GHz for downstream spectrum.

If a HGW strategy is used where MoCA is only on the home network and top-split is only on the access network, then there is no need to set aside bandwidth for MoCA as the two will never exist on the same media.

Table 4 shows the potential data capacity of a top split system. A 500 MHz RF bandwidth is presumed with an OFDM modulation and 75% bandwidth efficiency.

Modulation	Bits Per Hz	Data Capacity
64-QAM	6	2.2 Gbps
256-QAM	8	3 Gbps
1024-QAM	10	3.75 Gbps

Table 4 – Top-Split Data Capacity

The upstream transmitter is generally more expensive than a classic DOCSIS upstream transmitter. Because DOCSIS uses lower frequencies, the upstream spectrum can be directly generated with a DAC (digital to analog converter).

Top-split requires the classic approach of generating an I.F. (intermediate frequency) and then using an upconverter and power amp to get to the target spectrum with the correct power level.

This usually involves adding tripexors to the network. If there is additional downstream spectrum placed up above the new upstream, it could even require quadplexors.

The top-split could be built as a separate overlay network with the existing HFC plant, or as a last mile extension on a node plus zero (N+0) architecture.

For long distances, the modulation technique might be a lower order spread across frequency. For a short distance, a high order modulation could be used.

If the last mile was passive, perhaps even a TDD (time division duplex) could be used to achieve bi-directional bandwidth expansion instead of FDD (frequency domain duplex).

OPERATIONAL CHALLENGES

This section deals with the challenges and potential solutions involved with implementing a new return path. It should come as no surprise that none of the four techniques are without controversy, complexity, compromise, or cost.

Deep Fiber

In order to expand the bandwidth of a low-split solution, service groups (SGs) have to be split. This could occur within existing fiber nodes but often requires that fiber has to be pushed deeper. By splitting service groups, more physically separated upstreams exist.

For example, to increase the overall upstream capacity of a single 500 HHP SG from 100 Mbps to 1 Gbps would require the node to be split into ten 50 HP SGs. Usually splits are done in powers of two, so this might really be 8 SG of 62 HHP per SG (on average).

Splitting SGs creates more SGs with less subscribers per SG. While this mathematically provides the same data capacity as a faster upstream, in practice it is not as good. The peak rate will always be limited to the media speed. With a max shared DOCSIS capacity of 100 Mbps, the offered rate per user may only be 50 Mbps which will eventually not meet the needs of the market.

Also, when SGs get smaller, statistical multiplexing is lost. In practice, SG sizes will still vary based upon geography, and subscriber density will be uneven.

Splitting SGs translates to cost. Upgrading the fiber node, adding deeper fiber, adding additional fiber or wavelengths on the existing fiber runs to get 10x the backhaul capacity to the hub costs money. And with 10x the number of returns paths, 10x the number of upstream ports on the CMTS will also be needed. That is also a cost issue.

However, the methodology and technology for doing service group segmentation is simple, well-known, and somewhat optimized. Technologies such as RFOG (RF over Glass) provide a method of achieving a deeper fiber architectures while maintaining the existing data and video infrastructure.

The other three solutions gain more data capacity through spectrum expansion. So, in theory, mid-split and high-split require less or no changes to SG sizes or deeper penetration of fiber, unless the data capacity of the new spectrum is insufficient.

Impact to Spectrum Allocations

Low-split maintains the existing spectrum plan. No change to the upstream spectrum also means no change to the downstream spectrum and no change to the customer CPE.

Mid-split requires the removal of channels 2 through 6 and channels 95 though 97 (and channel 1 if present). This is 9 channels or 54 MHz of spectrum removed out of the downstream path. The channel count for a 750 MHz system would be reduced from 116 to 107 which is a 7.7% reduction in capacity. If that 750 MHz system was upgraded to a 1002 MHz system with 148 channels (due to mid-split), then that system will see a net gain of 28% in downstream capacity. This is shown in Table 5. [CT-1] [CATV-NA] [CEA-1]

DS spectrum	Current Channel Count	New Channel Count	Relative Impact	Net Impact
750 MHz	116	106	- 7.7%	+28%
862 MHz	135	125	- 7.4%	+10%
1002 MHz	158	148	-6.3%	-6.3%

Table 5 - Impact of a 88/108 MHz Mid-Split and 1002 MHz upgrade on DS Spectrum

High-split requires the removal of channels 2 through 29 (and channel 1 if used) as well as channels 95 though 99. This is 34 channels or 204 MHz of spectrum removed out of the downstream path. The loss of DS channels from high-split and the improvement from upgrading the downstream to 1002 MHz is shown in Table 6.

The removal of downstream spectrum generally will make it imperative for the downstream spectrum to upgrade to 1 GHz if it has not already. The good news is that the upstream and downstream spectrum upgrades share a lot of components and labor costs. Thus, it would be a waste of capital to not include a downstream upgrade at the same time as the upstream upgrade.

DS spectrum	Current Channel Count	New Channel Count	Relative Impact	Net Impact
750 MHz	116	82	- 29%	+ 7%
862 MHz	135	101	- 25%	- 8%
1002 MHz	158	124	- 22%	- 22%

Table 6 - Impact of a 200/258 High-Split and 1002 MHz upgrade on DS Spectrum

Top-split also maintains the existing spectrum. However, it caps the downstream spectrum to a particular frequency (say 1 GHz) that could limit future downstream growth. It builds new spectrum above 1 GHz. This means additional cost in plant operations, maintenance, and customer CPE.

Transition Planning

Mid-split and high-split require the removal of downstream spectrum. This has to happen prior to the start of the plant upgrade. The new downstream spectrum from a 1002 MHz upgrade may not be available until after the upgrade.

That may leave a window during the upgrade cycle where there is a spectrum shortage. That might not be acceptable. Clever planning with duplication of channels in the low spectrum and high spectrum will be needed along with pre-positioning of CPE equipment to ensure a smooth cut-over.

Analog TV

Low-split and top-split do not impact analog TV deployment. While mid-split delivers a severe blow to analog TV services with the removal of channels 2 through 6, high-split effectively backs up the truck and runs analog TV over.

The loss of the lower analog channels 2 through 6 are politically the hardest to get rid of. Getting rid of analog TV also may commit the operator to deploying millions of low-end DTAs (Digital Terminal Adaptors) that would be need to support all the remaining analog TVs.

To seriously deploy a mid-split or high-split system, it is probably time to get rid of all analog TV. While this may be inevitable for some cable operators, for others it would be a tough trade-off

Legacy OOB

Legacy STBs that do not use DSG (DOCSIS Set-top Gateway) instead use a discrete downstream and upstream carrier for communications with the headend. The downstream carrier is 1 MHz wide for SCTE 55-2 (Cisco) and approx 1.7 MHz wide for SCTE 55-1 (Motorola). Typical placement of center frequency is between 73.25 and 75.25 MHz as there is a gap between channels 4 and 5. The older "Jerrold" pilot (prior to Motorola/GI) was at 114 MHz. By spec, the STB must be able to tune up to 130 MHz.

There are no compatibility issues with the STB OOB channel and low-split or top-split.

For mid-split, the OOB channel can be placed above 108 MHz in the downstream spectrum. This should work unless there are some old STB that either can't find the new frequency or don't really support it.

For high-split, this is probably the biggest issue. The 200 MHz cutoff for high-split is well above the 130 MHz upper end of the OOB tuner range. So what to do?

This is primarily a North American issue. In the rest of the world where legacy STB penetration is much lower or non-existent, it may not be a significant issue.

The first approach would be to completely replace legacy STB with DSG capable STB or with STB that can tune the OOB channel above 258 MHz. The challenge with this approach is that 100% of the legacy STB have to be removed before high-split can be enabled.

The second approach would be to send an adaptor to any household that has a STB and has not upgraded to a high-split system. That adaptor might receive DSG and put out a legacy OOB signal. This issue with this approach is that that OOB adaptor design has to be agreed to, manufactured, tested, paid for, shipped, and installed. And once again, 100% of the STB population needs this adaptor before any high-split can be deployed.

The third solution is to provide some sort of OOB source or path deep in the HFC network that can inject the OOB signal into the upstream path. The installation of this solution would be done at the time of the plant upgrade. This is an area for ingenuity.

Legacy Tuners

Tuners in STBs and TVs in North American receive above 54 MHz with an expected maximum input power of +17 dBmV. Low-split and top-split co-exist with legacy tuners. Mid-split and high-split systems output RF energy in the upstream that is within the downstream operating range of the legacy STB and TVs.

If those devices are located near a CM that is blasting out energy above 54 MHz at levels approaching +57 dBmV (DOCSIS 3.0 max power for single 64-QAM), the power levels could saturate the input amplifiers of the legacy tuners, thus preventing the device from receiving a signal at any frequency.

So, what to do? There are several options.

First, it is worth noting that the typical tuner has an output intermediate frequency centered at 44 MHz. If 44 MHz was applied to the input of the tuner, it might pass directly through to the tuner output. Thus, it would be prudent to avoid transmitting any upstream frequencies from 41 to 47 MHz. That is easily done as the DOCSIS 3.0 North American spectrum stops at 42 MHz, so new carriers can be placed above this band.

The general problem is best split up into two smaller scenarios:

- Impact within the same home
- Impact to adjacent homes

The signal levels will be much higher when the CM and legacy tuners are within the same home.

One solution is to put a bandstop filter that would prevent frequencies above 42 MHz and below 85 MHz (for mid-split) or 200 MHz (for high-split) from reaching the legacy device. These bandstop filters need to leave the legacy upstream operating.

Bandstop filters would work for legacy TVs, but may not work for legacy STBs as the bandstop filter would block the downstream OOB signal (typically at 75 MHz). To let the

75 MHz carrier from the headend through but to block the local 75 MHz would require a directional coupler and very specific wiring. This is possible but very error prone.

Should a filtering solution be pursued, having the new upstream spectrum avoid using the downstream OOB frequency range might help the situation. (for example, 74-76 MHz).

Another solution is to avoid the problem altogether. The premise would be that the home gateway (HGW) would be deployed along with a video over IP strategy. The new CM is deployed as a HGW at the edge of the home.

The HGW becomes a demarcation point between DOCSIS and the cable plant on one side, and MoCA and the home network on the other side. This does imply a professional installation (to be discussed in a subsequent section).

DOCSIS could be terminated at the HGW and the HGW would drive the coax in the house with MoCA. Video and data would be deployed with IP STBs that interfaced to the MoCA network.

This is an interesting proposal in several ways. First, it solves the in home legacy tuner interference problem. Second, it isolates all the noise generated by the home network and prevents it from the HFC plant. Third, the HGW can get by with a lower transmit output power level.

It should be noted that there are other HGW upgrade scenarios that pass the downstream spectrum through to legacy STB rather than replacing the STBs.

The other half of the problem was the impact to adjacent homes. The interfering signal would have to travel up the drop from the home, travel between the output ports on the splitter, back down the drop to the next house, and then into the home network of the next house.

The easiest solution would be to set the new upstream power budget such that the signal would be sufficiently attenuated by the path described above so that it would not be a problem. This solution become harder when the customers are in a mutli-dwelling unit (MDU) where the coax drops are short.

There is an additional problem in MDU's where the outlets are cascaded or daisy-chained, where one drop is run to the next, then the next, and so on. This means you cannot easily use MoCA to provide the video and data channels to an individual customer without all customers on that run using the same MoCA channels. MoCA can address this problem through the use of provisioned VLANs.

Worst case, bandstop filters would have to be applied to the drop lines as they leave the in-line tap.

CM Upstream Power AMP

A DOCSIS CM must be capable of transmitting four 64-QAM carriers at +51 dBmV per carrier. [PHY] It should be noted that this is the output transmit power of the upstream amplifier.

The amplifier itself dissipates much more power than this because it tend to be a class A amp which have strong bias currents, and it must operate off of voltage rails large enough to support the dynamic range of the output signal.

An example is the ADA4320-1 from Analog Devices. It has drive current levels that can be programmed to match the number of carriers. For four carriers, it uses about 1.2 W; for one carrier, it uses about 1.0 W. Note

it uses a supply voltage of 5 V, D2.0 amps used 3.3 V supplies.

If the new CM was expected to maintain the same Power Spectral Density (PSD) or received power per DOCSIS upstream channel, it the additional output required can be calculated from follows the equation: 10*log10(BW_ratio). [CT-2]

For DOCSIS 3.0, the bandwidth of the four carriers is 4 * 6.4 MHz = 25.6 MHz.

For mid-split, the new spectrum from 46 MHz to 85 MHz which is 39 MHz. The additional power required for mid-split would be

10*log10(39/25.6) = 1.8 dB

For high-split, the new spectrum from 46 MHz to 200 MHz which is 154 MHz. The additional power required for mid-split would be

10*log10(154/25.6) = 6 dB

Note that doubling the bandwidth is double the power or +3 dB.

This additional power may create cooling problems in the CM upstream power amplifier. There are several solutions.

The first is to use the HGW architecture and to lower the required power level by at least the additional power calculated above.

A second solution is to turn the power amp off in between transmission bursts. This would allow the CM to burst a full rate for a while, but ultimately the rate would have to be lowered so accommodate the on/off duty cycle required for the amp. (note that this is done today on DOCSIS 3.0 CMs).

Aeronautical Interference

The frequencies from 108 MHz to 138 MHz are used for Maritime Mobile and Radio Navigation. This is shown in Figure 1 [SPECTRUM].

The new CM may be transmitting the frequencies from 108 to 138 MHz at a higher power level than the frequencies where transmitted when they were part of the downstream spectrum. The inherent leakage in the plant might be sufficient enough to cause interference.

Research would have to be done to validate this concern. If it is a problem, then the plant will have to be cleaned up to reduce this leakage.

This concern also existed 15 years ago prior to the deployment of DOCSIS. The plant did require cleaning up in many cases. It was done and the result was an HFC plant a more reliable plant. So, it is doable, but must be planned and budgeted for.

.

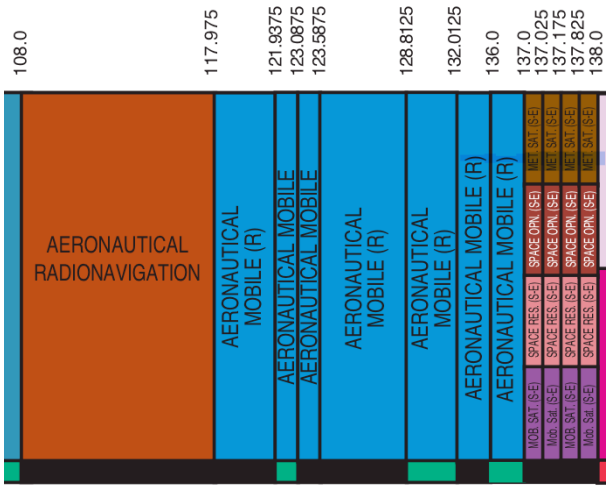


Figure 1 – Government Spectrum Allocation from 20 MHz to 200 MHz [CATV-NA]

Optical Node Technology

Optical nodes have two common choices for return path lasers. They are Fabry-Perot

(FP) lasers and distributed feedback (DFB) lasers. The optical return path can either be analog or digital (such as Cisco's Baseband Digital Reverse [BDR])

The FP lasers are lower performance and less expensive, but will work for one to two DOCSIS carriers. In order to carry a full four or six carriers, or to handle either mid-split or high-split, the optical node has to be upgraded to a DFB return path laser.

An alternative approach is to put the upstream demodulation electronics in the optical node instead of the hub site. This could be done at the existing optical node location, thus preserving the N+5 (or whatever) cable plant.

This shortens the DOCSIS return path by eliminating the optical segment. Now, instead of having a return path extending from inside the home all the way to the hub, it can be from the edge of the home up to the optical node. Rather than a 100 mile radius, it might be more like a one to two mile radius.

With the upstream QAM demodulators in the optical node and a shorter return path, the transmission impairments normally introduced by the electrical to optical and back to electrical are gone. This allows the transmission path to operate at higher rates.

Also, the optical path leaving the optical node can now be digital. This allows for a less expensive laser to be used. In fact, QAM demods could be placed on each of the up to four physical ports of the optical node. This would segment the optical node without the need of running fiber to the next active. Yet, there would be enough digital bandwidth on upstream fiber that only one wavelength and laser would be required.

The one big caveat on this approach is the handling of non-DOCSIS upstream carriers such as legacy OOB and plant telemetry (monitoring of power supplies, amps, nodes). That might require separate demods or just a digitization of a limited amount of spectrum that can be packetized and sent up to the hub.

For mid-split, the return path amplifier in the optical mode may have the required bandwidth depending upon the age of the optical node. For high-split, there is a higher likelihood the return path amplifier will need to be upgraded.

For top-split, an entirely new optical node is often used that is in parallel with the current optical node or deeper in the fiber network. It manages the top-split as an overlay network. This new optical node would either use separate fibers or separate wavelengths to connect to the hub site.

Mid-split and high-split would require that the diplexers in the optical node be changed. In some optical nodes these are pluggable while in other optical nodes they are soldered in. This depends upon the manufacturer of the optical node and the customer requirements.

For all of these reasons, worst case, a mid-split or high-split upgrade will require a swap of the optical node.

Amplifier Technology

For mid-split, the return path amp may work and if the downstream is not upgraded, the downstream amp may be sufficient. The diplexors will need upgrading.

For high-split, there is a small probability that the return path amp may not work at 200 MHz and will need upgrading. If the downstream is to be upgraded to 1002 MHz, the downstream amp will need upgrading. The diplexors will need upgrading.

For top-split, there has typically been a bypass amp that is placed in parallel with the current amps. It has its own triplexors and

two-way amps. A more practical solution would be swap out the amplifier with a new triplex amp.

Since the attenuation of the coax is higher above 1 GHz than below 1 GHz, top-split needs closer amplifier spacing. An HFC plant that is built using maximum distance between conventional amplifiers may need additional amplifiers added to the network.

For all of these reasons, worst case, a mid-split, high-split, or top-split upgrade will require a swap of the amplifier.

In-Line Equalizers

Some HFC plants have passive in-line equalizers. These equalizers use diplexors to isolate the downstream from the upstream so that the passive equalizer can be inserted.

For mid-split and high-split, the in-line equalizers would have to have their diplexors upgraded.

For top-split, the equalizer has to be overlaid with a new device that supports the top-split upstream.

Upstream RF Transmission Path

Low-split extends from 5-42 MHz. The spectrum from 5 MHz to 22 MHz is often special cased due to the presence of noise. If 5-20 is ignored, 20 to 42 MHz is one octave (a doubling of frequency).

Mid-split would contain approximately 2 octaves and high-split would contain approximately 3.3 octaves.

What type of transmission parameters could shift over the span of 3.3 octaves that would become noticeable? Tilt? Group delay? Ideally, the plant would spec its worst case transmission performance, and it would be the job of the new electronics in the new mid-split and high-split devices to compensate.

Plant Power

Mid-split and high split may add new electronics such as QAM modulators or demodulators to the optical node that may increase its power dissipation.

High-split adds an overlay network of optical nodes and amplifiers that will increase overall power requirements for the HFC plant. The HFC plant is powered. That powering typically comes from a mains power at the hub site with backup generators in case of a mains failure. Upgrading these facilities is a cost that should be factored in. Even an increase the power draw on existing facilities is an increase in cost.

Professional Installation

Installation practices vary across cable operators. Some cable operators have separate truck rolls for data, voice, video, cable card, and then a final truck roll to fix everything that did not work.

Other Cable Operators are able to sell DOCSIS service through web signup and mailing a CM. A truck roll is only done when things do not work out.

This latter scenario might not be possible if a CM upgrade is combined with an IP STB and HGW upgrade, along with a check of signal levels, emission levels, and impact on adjacent dwellings.

Table 7 summarizes these operational issues with respect to the 4 solutions under consideration.

Approach	Pros	Cons
Low-Split	• All equipment already exists • No disturbance to spectrum • Simple	• Cost: Requires deeper fiber. • Cost: Requires more CMTS ports • Cannot hit peak rates over 100 Mbps of return path throughput
Mid-Split	• Supported by DOCSIS 3.0 equipment • Works with DS OOB	• All actives in HFC plant need to be upgraded • Cost about the same as high-split and only doubles the US throughput
High-Split	• Supports 1 Gbps throughput • Can co-exist with earlier versions of DOCSIS.	• All actives in HFC plant need to be upgraded • Does not work with DS OOB • New CM and CMTS components
Top-Split	• Leaves existing plant in place. • No Impact to existing legacy customer CPE • Only customer taking new tiers would require new HGW CPE	• Requires triplexors • New active return path has to be build on top • Inefficient use of spectrum • High attenuation requires high power. Existing amplifier spacings may not be sufficient • Blocks expansion of downtream bandwidth directly above 1 GHz

Table 7 – Summary of Operational Issues

COST

For all these solutions, the bottom line is:

- Does it work?
- What does it cost?

There are 6 baseline cost scenarios to be considered.

1) The cost of doing nothing
2) The cost of all fiber
3) The cost of low-split
4) The cost of mid-split
5) The cost of high-split
6) The cost of top-split

There are more variations, but this is a good baseline. Now, some caveats.

- Any analysis has lots of assumptions. This one does as well. Therefore, your mileage will vary.
- This is not a price quote. I'm winging it here. Don't take this to the bank.
- My main interest is comparisons. Thus absolute accuracy level could be off by 50% to 100%.
- Costs change with time, technology, and vendor.

There seems to be multiple way of quoting costs for plant upgrades. Be careful when comparing numbers. The ways are:

- $ per mile
- $ per subscriber
- $ per home passed

I am going to use per home passed as that metric is the most common usage in the DOCSIS world.

HFC Plants are often described a N+5 or N+0, etc. The N means node and represents the optical node. The number following this is the number of max number of amplifiers in a row to get to the last mile. The common design point for current HFC networks is N+5. A deep fiber network with no actives in the coax plant is considered an N+0 architecture.

The Cost of Doing Nothing

Doing nothing can be a viable alternative. It is the lowest cost option, but may not be the best revenue option. If there is competition, inaction can lead to a loss of customers and a loss of revenue.

The other risk is not spending enough when there is an upgrade. For example, if there is an upgrade to IP video which leads to large equipment swaps, performing additional upgrades at the same time may lower cost of both upgrades.

The Cost of All Fiber

This is the opposite of doing nothing. It is doing everything.

With the advent of new RFOG PON and upcoming DOCSIS EPON technologies, building a fiber to the home network is technically possible. It is also the most expensive solution.

One place to get a reference is the published costs by other of FTTH solutions. Bear in mind, these are usually second hand information, so accuracy may vary.

Bell Canada [FASTNET]
- $650/HP to pass a home

France [FASTNET]
- $650/HP to pass a home

Google: [FASTNET]
- $700/HP to pass a home

Verizon [VERIZON]

- $750/HP to pass a home
- $600/HP additional to connect a home

These prices tend to be for aerial plants. If the fiber is underground, cost could be typically 75% higher.

Telcos have some fiber in the local loop, but not as much as the cable operators. Thus, the cable operator should be able to leverage the fiber already run out to the optical node. It will probably require WDM equipment. Also, the cost of the fiber runs increases as you get closer to the home as there are more runs. All in all, this could translate to a 25% discount, give or take.

Passing a home refers to getting the fiber to the curb. Connecting a home refers to the drop cable plus in-home CPE. The cable scenarios already have a drop cable. So to arrive at a comparable estimate, I will take $700 as the average from above for passing a home and assign $200 in cost for installing a drop cable to the edge of the home (labor, fiber, termination box). I am ignoring the CPE cost. That results in an estimate of $900/HP. This is considered an aggressive number.

This new fiber network would require a new OLT (Optical Line Termination) Edge device at the hub. For comparison to the other cable scenarios that follow, this is a 2x equipment increase at the hub (one CMTS, one OLT vs. just one CMTS)

The Cost of Low-Split

For comparison, to get a 100 Mbps upstream up to 1 Gbps, the plant would have to be split by 10x. That means that optical nodes of 500 HP would have to be split down to 50 HP. This is really a smaller node size than practical. It implies a N+0 architecture.

This would also require 10x the number of CMTS ports.

Throughput would be limited to 100 Mbps aggregate.

The Cost of Mid-Split

For comparison, to increase to 1 Gbps throughput would require a 4x optical node split. This is presuming 300 Mbps per upstream as calculated in Table 1.

In theory, a 4x node split could be accomplished within the existing optical node housing. In practice, new fiber is likely to get added to push four new optical nodes deeper. For analysis, we will assume this is a N+3 or N+4 architecture.

If upstream QAM demodulators are pushed down into the optical node, that circuitry will likely have four ports which will effectively provide the 4x sub-split needed all within the existing node housing.

This would require 4x the number of CMTS ports.

All optical nodes, amplifiers, and in-line equalizers need to be upgraded or swapped out.

This upgrade is very similar (or slightly higher), if not identical to a high-split upgrade. Thus, the cost of the mid-split plant upgrade will be presumed to be the same.

The Cost of High-Split

Since high-split has a Gbps return path, by definition, no node split is required. For this analysis, we will assume that this stays at the N+5 reference.

As with mid-split, all optical nodes, amplifiers, and in-line equalizers need to be upgraded or swapped out.

This would require a new CMTS line card and is the equivalent of 2x the number of CMTS ports.

It turns out that this is roughly the same scenario for a 1002 MHz downstream upgrade. When upgrading the downstream path, all amplifiers and optical nodes need replacing. Often the downstream lasers at the hub are replaced as well. This is why it makes so much economical and technical sense to upgrade both directions at the same time.

In talking with industry experts, a 1 GHz upgrade can range from $55/HP to $85/HP. This partly is influenced if the plant is aerial or underground. Adding an upstream upgrade may add a 30% premium to this, driving it to $70/HP to $110/HP. I am going to take the average of these numbers as $90/HP.

The Cost of Top-Split

Early top-split networks were built as an overlay network that involves additional optical nodes, amplifiers, equalizers, taps, etc. The current thinking for a top-split network is to drive to a deep fiber architecture, potentially as an overlay to the existing coax, and connect into the coax after the last amp as a N+0 (or before as a N+1).

The drop in point would have a circuit that terminated the top-split return path and coupled it to the fiber.

The backbone coax is generally capable of supporting up to 3 GHz and maybe more. The network splitters/taps also may need replacement, depending on where the top-split network is connected into the HFC plant.

The RG6 drop cable may need replacement if the length exceeds 200 feet. If the length is less than 100 feet, it should be fine. If it is between 100 feet and 200 feet, it is a maybe.

The cost estimate would be more than a low-split as the fiber cost would be the same but there are added electronics.

Cost Summary

Table 8 shows the relative pricing for plant upgrades normalized to homes passed. This does not include any CAPEX for CPE equipment of any OPEX.

The right hand column indicates the relative hit to the density of the access edge device (CMTS or OLT). The cost impact is actually a fraction of the plant upgrade cost. However, as the multiple increases, the cost of the edge equipment becomes noticeable. As a base measure, this table suggests that a high-split system for a typical 40K homes passed hub is:

40K HP/CMTS * $90/HP = $3.6M

40K HHP with 1000 HP/SG (Service Group) is 40 SGs which is approximately one CMTS.

A useful metric that occurred in this analysis is truck rolls per active. Every active in the plant needs swapping out in many of these scenarios. Then there are follow-up checks. The average truck roll per active is 1.3 to 2.

This analysis left out all the cost associated with the home including professional install costs, CPE costs, etc. These are valid cost that would play into the final choice

It should be no surprise that an all fiber network is the most expensive solution. A

Approach	SG Size	Cost/HP	Relative	Edge Impact
FTTH	n/a	$900	100%	2x
Top-Split	N + 1	$225	25%	2x
Low-split	N + 1	$180	20%	10x
Mid-split	N + 3	$135	15%	4x
High-split	N + 5	$90	10%	2x

Table 8 – Relative Costs for Plant Upgrade

complete new build should always cost more than a retrofit. As such, fiber may be interesting for new builds.

Low-split ended up being the most expensive way to get 1 Gbps of total upstream throughput, and yet each end-point is really limited to 100 Mbps. The counter argument is that it is the simplest solution, the entire solution is available today, there are no compatibility problems, and you can pay as you go.

Mid-split is equal to or maybe even more expensive than high-split. The main disadvantage of mid-split is that for the same plant investment, you can get 4x the bandwidth with high-split and the ability to burst to 1 Gbps. The main advantage is that all the equipment is available today (DOCSIS 3.0) and it allows legacy STB to stay on the plant because is compatible with the downstream OOB.

High-split offers fiber like performance, yet at one-tenth the price of fiber. What a deal! The disadvantage is that it will take 3-5 years to bring the technology to market and the upgrade plan is the most challenging of all the options. Further, the cost of eliminating all STBs and the need for DTAs may reduce the cost benefit of this option.

Top-split offers potentially additional performance compared to high-split but at a higher cost. The upgrade plan is simpler than high-split. Top-split does leave existing customer premise equipment in place and only impacts those taking the newer tiers requiring the higher speeds.

CONCLUSIONS

The HFC plant has plenty of life left in it. There are several ways to drive the upstream data capacity to 1 Gbps. The downstream can be driven to 5 to 10 Gbps (a topic for another paper).

These approaches offer fiber-like performance but on an HFC plant as low as one-tenth the price of a fiber installation.

REFERENCES

[BDR]
Baseband Digital Reverse, Cisco, http://www.aboutcisco.biz/en/US/products/ps8965/index.html

[CATV-NA]
North American Cable Channel Spectrum Plan, Wikipedia, http://en.wikipedia.org/wiki/North_American_cable_television_frequencies#North_America_cable_television_frequencies

[CEA-1]
Cable Television Channel Identification Plan, CEA-542-C, Feb 2009, R8 Cable Compatibility Committee, http://www.ce.org/Standards/browseByCommittee_2549.asp

[CT-1]
Broadband: Counting Channels, by Ron Hranac, Dec 1, 2008, http://www.cable360.net/ct/operations/bestpractices/32720.html

[CT-2]
Broadband: Total Power and Channel Bonding, by Ron Hranac, April 1, 2010
http://www.cable360.net/ct/sections/columns/bullpen/40532.html

[FASTNET]
Google's Gigabit Could Cost $700/home, Fast Net News, Feb 15, 2010,
http://fastnetnews.com/fiber-news/175-d/2575-googles-gigabit-could-cost-700home

[FSN]
Full Service Network, Wikipedia,
http://en.wikipedia.org/wiki/Full_Service_Network

[PHY]
DOCSIS 3.0 PHY Specification, Jan 21, 2009, Section 6.2.24. CM-SP-PHYv3.0-090121.pdf,
http://www.cablelabs.com

[SPECTRUM]
United States Radio Spectrum Frequency Allocations Chart, 2003, United States Department of Commerce,
http://www.ntia.doc.gov/osmhome/allochrt.pdf

[VERIZON]
Verizon's FIOS Expansion Comes to an End, American Consumer News, March 29, 2010,
http://www.americanconsumernews.com/2010/03/verizons-fios-expansion-comes-to-an-end-nyse-vz.html

THE CHALLENGES OF 3D SUPPORT IN THE STB

Kevin Murray, Simon Parnall, Ray Taylor
NDS

Abstract

Distribution of stereoscopic 3DTV has been demonstrated using existing, deployed HD technology providing excellent quality pictures. This shows how easily 3DTV can be made available through broadcast channels as consumer displays reach the market and 3D content becomes readily available. However, a full broadcast service consists of more than just the video—the STB and the features it provides are key parts of the user experience.

This paper looks at several areas of key functionality that the STB provides. It discusses the changes required both in the STB software and the transmissions on which it relies, all whilst utilizing existing HD hardware. Through these discussions we explore how STBs can be updated or extended to support a seamless, high quality 3D aware service.

INTRODUCTION

The set top box (STB) is one part of the distribution path from the broadcaster to the display in the home. It is perfectly possible to utilize this path for stereoscopic 3DTV (S3DTV) without making any alterations to the STB software, by utilizing frame compatible broadcasts – formats that fit a stereoscopic pair into a normal frame. Indeed, broadcasts have been made that do just that. Initially we touch on certain aspects of broadcast formats, and how to minimize stereoscopic specific transmission problems.

In the frame compatible scenario, if the STB is not aware of the format or nature of the video it is carrying, then several functions of the STB will result in very unpleasant visual effects. The main discussion in this paper looks at two key areas of this functionality, and discusses the steps that are needed to make the STB software 3D aware, and the additional signaling or metadata that prevents, or at least minimizes, such unpleasant effects.

The first area we discuss is that of manipulating S3DTV video. Examples of manipulations include supporting such features as picture-in-picture and picture-in-guide. These operations are more complex in S3DTV than in 2D, and in some cases the use of such functionality may be less sensible or visually acceptable. The complexity can vary with format and operation, so we will discuss the key popular formats: side-by-side and top-and-bottom (also known as above-below).

The second key area where updates are needed is in the handling of closed captions. However, the required changes extend to almost all graphical overlays performed by the STB. These updates cover not just the format used to draw the graphics, but also issues in design to reduce eye-strain and in the importance of correct depth placement. Depth placement, and potentially accurate matching of graphics to video, rely on correct signaling and metadata provision in the broadcast. We identify a small set of extensible signaling that assists these key areas.

Although most of this paper is concerned with systems operating with frame-compatible video formats, many of the points apply to non-frame compatible distribution mechanisms and the support required in the STB for such formats. In several cases, the solutions can be identical for frame compatible and non-frame compatible modes. Likewise, many of the examples in this paper

are based on side-by-side formats, but the ideas extend to top-and-bottom formats as well.

STEREOSCOPIC VIDEO FORMATS

There are several formats for stereoscopic video in a frame compatible mode, some of the more common ones (that are also part of the mandatory HDMI 3D formats) are shown below in figures 1 and 2. In the simplest mode where the STB is not aware of the presence of S3DTV, the STB just decodes the received video from this format and outputs it in the same format over the HDMI connection to the display.

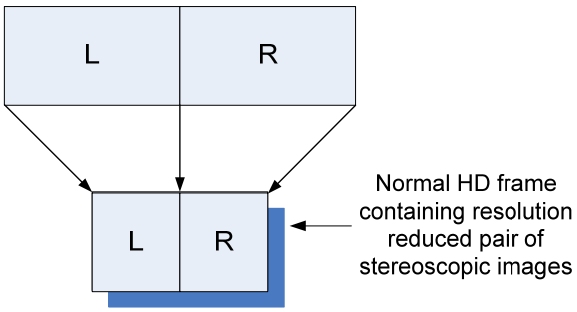


Figure 1: Side-by-Side Format

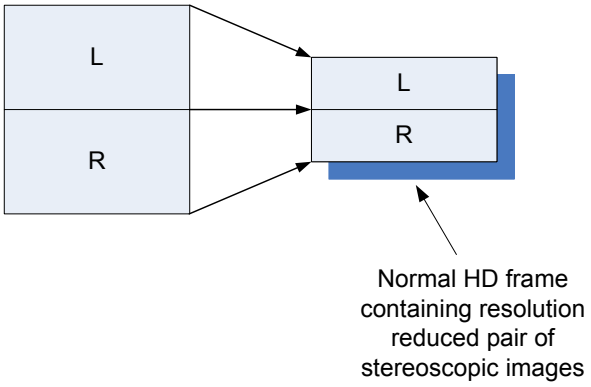


Figure 2: Top-and-Bottom Format

There are two main display technologies available for S3DTV—shutter based and passively polarized—and there is no requirement for a specific format for a specific display technology. These displays use HDMI signaling[1] to identify the input format. The display will then perform the necessary conversion to enable it to display the stereoscopic images. Thus a broadcast in one of the mandatory formats can be supported regardless of the display technology. In the absence of STB supported signaling over HDMI the viewer is required to use a remote control to set the display to the correct input format.

Side-by-side and top-and-bottom represent different trade-offs for resolution reduction. It is worth noting that these different formats result in different effective resolutions on differing display technologies.

A Note on Bitrates

When compressing stereoscopic video it is very important to operate the compression at a level above that where artifacts can occur. If artifacts do occur, there is no guarantee that the left and right eye images will match, and the viewer will experience discomfort.

Format Translations

Just as the displays are able to perform format conversions, so can the STB. Whilst most devices are easily able to scale and resize 2D video, conversion between side-by-side and top-and-bottom formats is probably not within the capability of most deployed devices. In comparison some other conversions, such as from top-and-bottom to line interleaved, are simple to perform.

Further, format conversions can result in a loss of resolution. For instance, conversions between side-by-side and top-and-bottom will normally result in an image where the L and R images are effectively one quarter the resolution of a full HD image. It is better to avoid conversions, or ensure that it is performed with awareness of the native requirements of the display. The chosen broadcast format should also reflect these limitations.

As a simple example, a line-interleaved passively polarized display will normally have a resolution equivalent to that of the top-and-bottom format. Thus if an image is received in top-and-bottom format, converting it unnecessarily to side-by-side would reduce the effective resolution and therefore the quality seen by the viewer.

When considering future STB devices and chipsets, it will be important that they are able to convert whatever formats they receive into the most appropriate output format for the their display. For example, a new device that can support 1080p60 per eye must be able to output at least one mandatory format[1]. Ideally, it should be able to select from more than one to maintain the best quality for the display. This is especially important as the early 3D displays will represent the legacy formats of the future, and it should be possible to drive them with the best quality signal they can accept.

Synchronization

The frame compatible formats have the advantage that they place both the left and right in the same image. This guarantees that the left and right eye images are not swapped in the delivery and display process, and ensures that the left and right images are always perfectly time synchronized.

Approaches for transmitting S3DTV that are not frame compatible can involve twin logical streams that may, or may not, be represented as separate flows within the appropriate transport (e.g. different PIDs within an MPEG-2 transport stream). Whilst both time synchronization and left-right synchronization can clearly be preserved through such twin stream systems, they do represent a point at which errors can occur. The failure of either (or both) synchronization(s) renders the content effectively un-viewable. Alternative approaches such as using a high frame rate stream where alternate frames are for alternate eyes introduces the risk of a left right swap, especially where any processing or frame re-synchronization is performed.

It is tempting to consider encoding left-right images without a reduction in resolution by using either 3840x1080 (for side-by-side) or 1920x2160 (for top-and-bottom). Unfortunately, these sizes of frames fall outside the maximum defined by H.264 for level 4.2 codecs. It seems likely that at least some chipsets will be able to operate with such resolutions, and this may be a desirable avenue of exploration.

Manipulation

There are numerous occasions where video is scaled by an STB, and one such case is picture-in-guide. Figure 3(a) shows how this should occur for side-by-side formats. This approach introduces two main problems: the difficulty of performing the video manipulation, and the potential for an unpleasant visual impact.

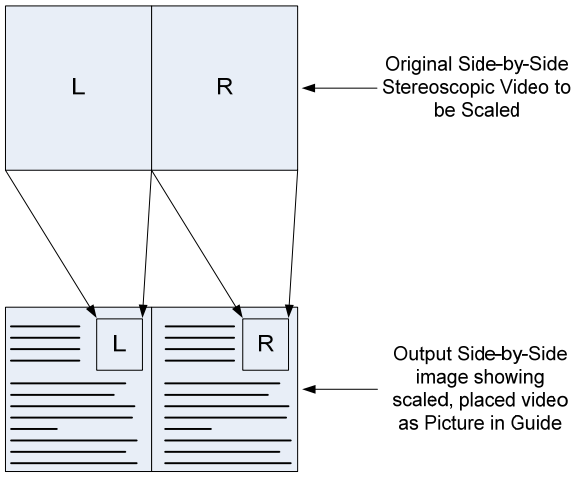


Figure 3(a): Picture in Guide and 3D Scaling

Where the EPG is making use of 3D effects itself, the range of depths used by the EPG may conflict with those of the video, resulting in a strange, and often unpleasant, effect. It may be preferable to simply operate the EPG in a 2D mode, as shown in figure 3

(b) or use only 2D video within the guide as shown in figure 3 (c).

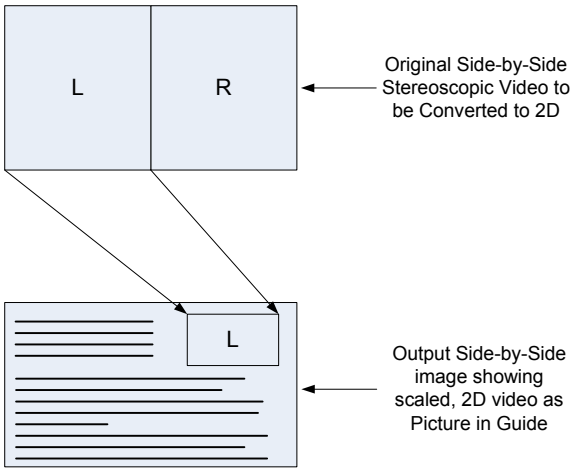


Figure 3(b): Picture in 2D Guide with 3D to 2D conversion

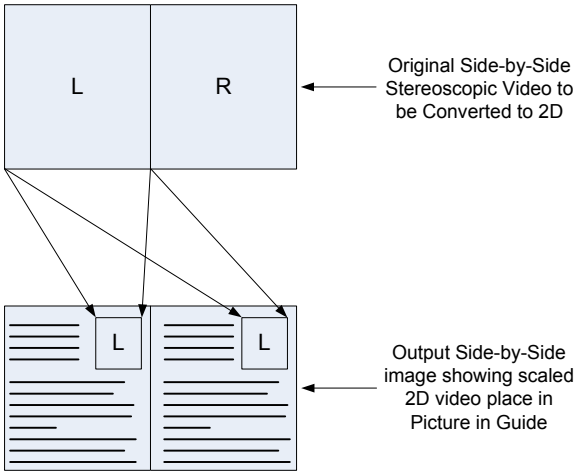


Figure 3(c): 2D from 3D Picture in 3D Guide

Similar visual issues occur with picture-in-picture and where both pictures are in 3D. In the authors' opinion, the conflicts are significantly increased above those with picture-in-guide. Whilst the effect can be partially reduced by providing a border around the inserted picture, to match with the video this border would need to occur in 3D space, and match with the volume shown in the content.

Certain video manipulations, such as those shown in figures 3(a) and 3(c) are significantly different from those normally used in 2D, as they involve two rescaling operations on each video frame. Therefore, these operations can present a significant challenge to existing hardware. In comparison, operations such as those shown in figure 3(b) are common for existing hardware and so easy to implement.

GRAPHICS

One of functions provided by the STB is the provision of graphics or on-screen displays (OSD). This ranges from support for closed captions through channel information banners and interactive applications to electronic program guides (EPGs). There are a range of challenges when handling OSDs and graphics for 3DTV, varying from the relatively simple such as ensuring that they are correctly visible, and that they do not conflict with the video, to including changes in design.

Readability and Format Awareness

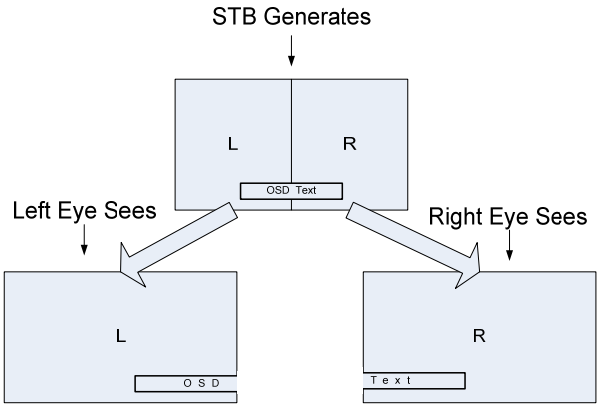


Figure 4: Impact of lack of S3DTV awareness on graphics

Drawing an OSD in a 2D without awareness of the underlying stereoscopic format results in images that are both unreadable and exceptionally disturbing to the viewer. An example of this effect for the case of side-by-side video is shown in figure 4. The top-and-bottom format produces different but still disturbing results with the OSDs only being visible in one eye. By comparison,

frame interleaved formats do not have this readability issue, if handled correctly in the hardware[1].

The readability issue is solved by adapting the graphics stack so that any OSDs are displayed as two images—one image for each eye. This clearly requires accurate signaling of the format so that the STB can alter the details of the graphics to match the underlying video format. When correctly implemented, applications providing graphics do not need to be aware of the 3D nature of the underlying video, as the graphics stack can handle the relevant translations transparently. However, as we shall discuss, there are several reasons why applications may, and in some cases should, choose to be 3D aware.

Depth Conflicts and Awareness

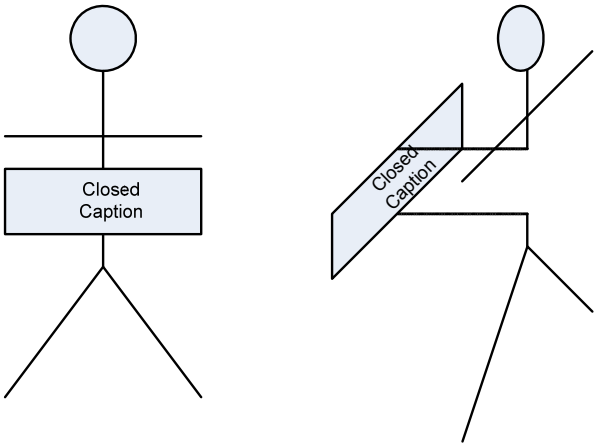


Figure 5: Depth conflicts with Closed Caption Overlays

Once graphics are drawn correctly for the underlying S3DTV video format, the next challenge is to consider where they should be placed. Previously, OSDs only had an x and y location, but with S3DTV they potentially also have a depth, or a z location. Likewise, the video, and the objects in it, occupy a set of depth locations. Careless placement of video can result in a conflict (a visual dissonance) between the objects in the video and the OSD. This happens when graphics are drawn which obscure objects in the video that the viewer knows should be in front of the OSD. An example of this is shown in figure 5.

The perceived depth location of graphics is controlled by the relative positioning of the left and right eye images, and so the STB is able to control the depth at which an object is seen. Figure 6(a) shows the relative offsets of the left and right eye images for a side-by-side format, and figure 6(b) shows the resultant depth that a viewer will see.

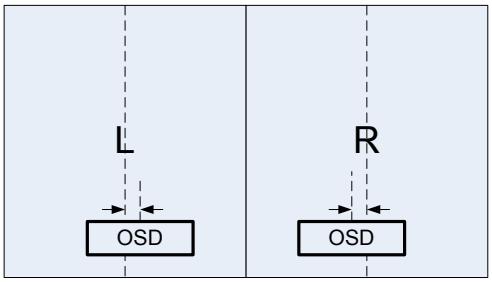


Figure 6(a): Offsetting of OSD in side-by-side S3DTV

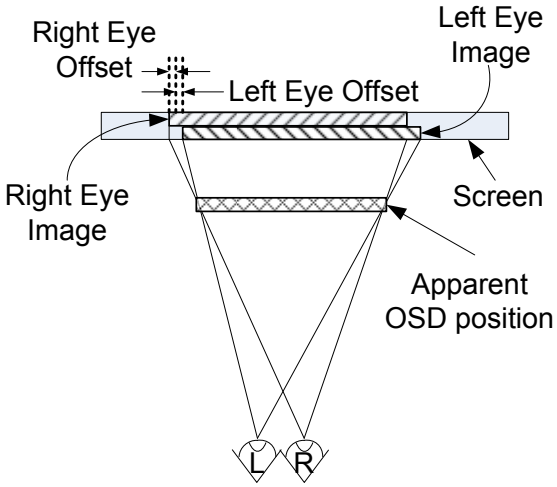


Figure 6(b): Apparent OSD depth position

It is possible to choose a fixed depth position, and always place the OSDs at that location. For example, to minimize the depth conflict, placing the OSD in front of the screen such as shown in figure 6 seems

[1] Some proposed schemes for internal handling treat the frame interleave pair as a single large frame, and so can have identical problems to those of frame compatible formats.

sensible. For short term viewing, this is an adequate general solution, but it does not properly address the entire problem. Firstly, it is possible that occasionally the underlying video will come further forward than the chosen fixed location. Secondly, and more importantly, this can place the OSD significantly forward of objects in the video – unnecessarily increasing the depth budget of the content.

Long durations of high depth budget viewing can increase eyestrain, and consequently approaches to OSD placement which increase the range of depths in use may reduce the appeal of 3DTV. Much content is deliberately created with a careful and conscious choice over the amount of depth that is used. Thus ideally any OSD, be it a closed caption, information banner or an interactive application, should be aware of the depth of the content and seek to fit within, or very close to, the depth range chosen by the content creator.

Design of S3DTV Graphics

The nature of stereoscopic TV is that it provides an illusion of 3D though the brain fusing two images together. In the real world, these two images would change continuously in relation to even the slightest movement of the viewer's head. Without the parallax cue from this continual change, the remaining cues that help the brain fuse the two images together become even more important. In our exploration of the design of 3D graphics[2] we have identified some areas that appear to assist the brain in fusing these images together, or that provide stronger cues, and so result in reduced eyestrain and brain fatigue.

[2] These explorations were performed on polarized, line-interleaved displays, driven by a PC application that rendered a range of graphics at the native format of the display (i.e. no use was made of side-by-side transformations) at 60Hz.

Graphics representing objects with volume present stronger visual cues compared to 2D graphics that are placed at a given depth. Graphics that are 3D objects, or 2D images placed on 3D objects, result in images that are easier to see and appear far more natural.

It is tempting to simply use one image and shift, or offset it, differently for each eye. Whilst this works for a flat planar image, this does not work well for a 3D object, since each eye would normally see a slightly different view. Figure 7 shows, in a somewhat exaggerated fashion, how the left and right eye images differ with a simple box. Thus the design process should ensure that a true 3D model is used and different images are generated with the correct perspective for each eye.

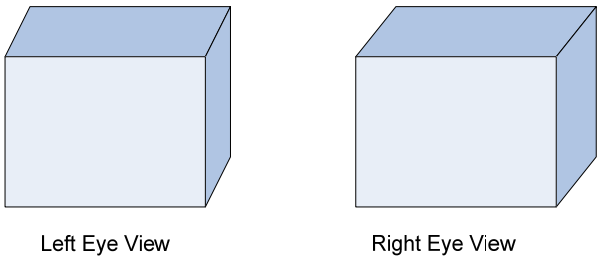


Figure 7: Differences between left and right eye views

When the object is perfectly square onto the screen there are very limited, or no differences between the left and right eye views. Tilting the object in one or more directions introduces differences between each eye's view of the object and provides additional depth cues to help the brain work out the depth placement of the object. Graphics actually look better off-square.

The use of motion on an object provides a much stronger depth and size presence. In part, this is because appropriate motion adds implicit and changing parallax cues through shape changes during movement. So an object that moves into view, changing its depth and rotation as it moves, appears easier on the eye than an equivalent object that simply appears

at a fixed location. Although this effect has a small persistence once motion stops it is not indefinite and it seems beneficial to provide continual animation if only for a small part of the graphic. This provides cues similar to the motion of the viewer's head provides in the real-world, or mimics the continual motion of many objects.

However, there are two aspects of motion that require care. Firstly we found that very fast motion makes it difficult to perceive objects and nearly impossible to determine their depth. Indeed, fast motion provided a worse experience than no motion. Secondly, we found that much of the benefit was lost if the animation process was at too low a frame rate.

The use of a static lighting source helps the reality and natural appearance of objects, especially where lighting related cues such as shadows and color variance are correctly generated and dependent upon true 3D placement. However, it is important that the effects of lighting are updated as objects move and that all objects are subject to the same lighting effects.

The texture of objects also provides depth cues, providing additional reference points on the surface of objects that help in calculating binocular disparity. However, care should be taken to avoid excessive texture, or random noise style textures, as these do not appear to strengthen the 3D presence.

The final aspect is the equivalent of the well-known 2D concept of a safe area. In 3D, we refer to this as a safe volume, and in particular the edges of the display need to be avoided when placing objects in front of the screen. Even objects behind the screen benefit from avoiding the edges. Placing images too close to the viewer is problematic, especially for longer durations (very short periods, especially for disappearing objects is not such a significant problem). Placing graphics far back into the screen is not a significant issue, except for the increased risk of conflict with any underlying video.

Bitmaps and GPU Advances

In the above discussion we have touched upon the need for different representations for each eye to provide some of the correct cues for the stereoscopic illusion, as well as the desire to animate these representations. Earlier, we touched on the need to be able to adapt the placement of graphic objects to differing depths. In many traditional systems, graphics rely heavily (though not entirely) on bitmaps – pre-computed representations of the image to display. Achieving the above goals, especially where multiple formats are to be supported, can easily result in a need for an unacceptably large number of bitmaps.

Newer STB chipsets are becoming available with a powerful graphics processing unit (GPU). These GPUs are able to take complex, abstract representations of a scene, often based on a mesh of triangles with lighting parameters, texture information and camera viewpoint specifications, and then convert this information efficiently into images for display. This approach allows for assets that represent the OSDs to be handled as abstract models and then converted as needed into the appropriate S3DTV format, and placed at the relevant depth.

Moving graphics towards abstract models and exploiting GPUs therefore provides an efficient method to achieve many of the goals above. When implemented correctly, this approach can allow a generic engine to support any abstract model so removing or reducing the need for a new graphics application or set of functions for each graphical asset.

METADATA AND SIGNALING

The sections above have identified a number of STB areas that require updates or changes. Many of these are significantly simpler or sometimes possible only where new or extended signaling or metadata is provided in the broadcast.

Format Signaling

Both AVC/H.264[2] and HDMI[1] provide signaling that is associated with each video frame, and that indicates the format of the stereoscopic data present in the frame. For HDMI this allows a display to perform the correct translations. In the same fashion, the presence, ideally mandatory, of signaling in the video stream allows the STB to adapt its graphics operations to the native, underlying video format.

By providing the signaling in the video, perfect synchronization of the format is enabled. However, this does mean that advance information is not available to the STB, which may be of importance in deciding whether or not to present the item to the viewer. Thus introducing additional signaling is needed in the broadcast. Such information is also useful to assist the box in identifying if it can handle the transmission, and could allow the STB to pre-allocate any increased resources it requires to support S3DTV.

Depth Information

It is possible to provide a single fixed depth value, but as has been discussed above it is desirable for this value to vary to reflect the content. An example of this is shown in figure 8 which represents a bird's eye view of two people walking down a corridor towards the camera. In figure 8(a) the speaker is some distance from the caption, but the caption is placed at the depth position the speaker will reach when the caption disappears. In figure 8(b) the second speaker starts, with a new caption depth placement, which is re-used when the speakers stop walking, as shown in figure 8(c). It is assumed that it is preferable to keep a single caption in at a single depth over the duration of its display, as shown, different captions in the sequence can occur at different locations. However, gentle motion of captions, with a suitable scaling as they approach or recede, is also potentially possible, and ideally the depth information should allow for that eventuality.

The information required for depth could be embedded in the closed caption data stream; however that implies that the values are only available when closed captions are present. An alternative, synchronized stream that might not even be part of the video, allows this information to be used by any OSD, regardless of the presence of caption data. This may be in addition to the depth information contained within a closed caption stream. Various mechanisms for carrying synchronized data exist, and they could easily be extended to carry depth information.

Such a stream of depth information may be generated at the head-end, or during the captioning process. Such depth extraction technology has been demonstrated by Technicolor[3] for off-line caption support and could be implemented within a stereoscopic aware encoder.

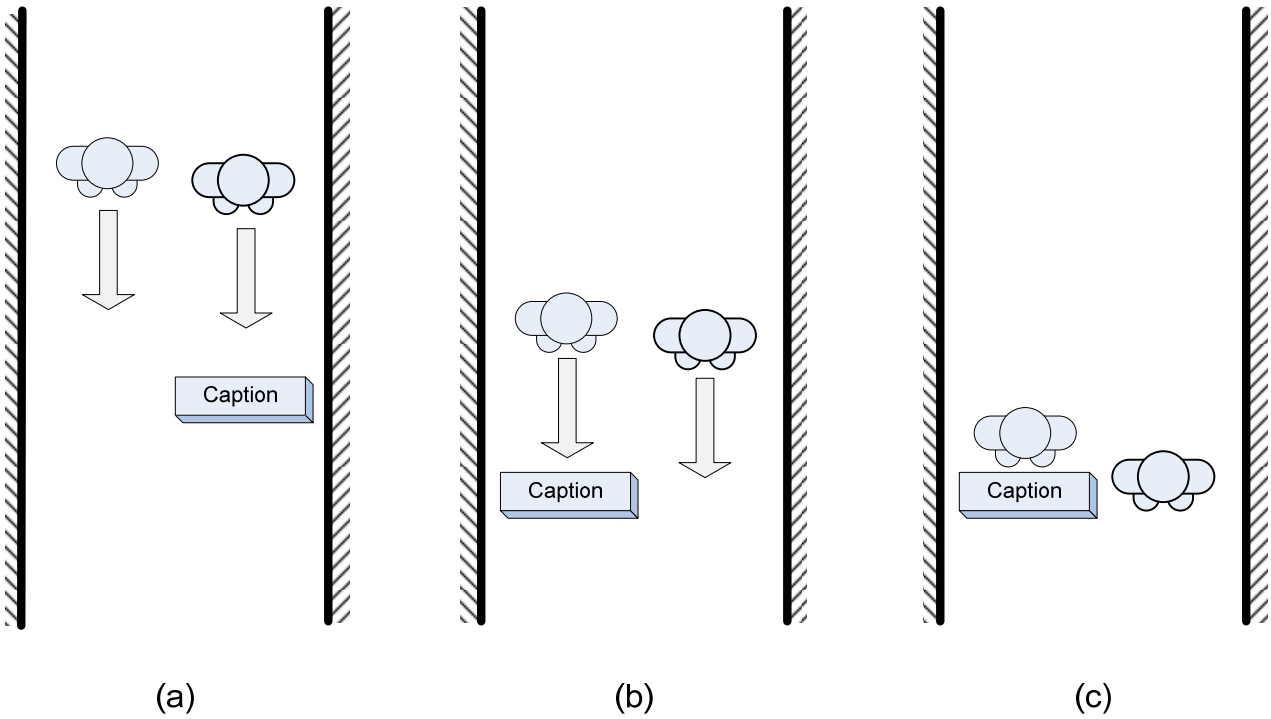


Figure 8: Bird's Eye View Example of Dynamic Depth Caption Placement

Additional Data To Enhance Graphics

When drawing graphics using a GPU, APIs such as openGLES, allow the setting of a wide range of parameters that control how the graphics appear. Some of these parameters correspond to information that is, at least theoretically, available in content creation and production. Examples of such parameters include information typically associated with the camera, such as the focal length of the lens, or that may be known from the setting, such as (e.g. in a studio) the primary lighting directions and sources.

Providing this information to the graphics system allows for a better matching of graphics to the underlying video. This is of most importance where the graphics are closely connected with the video, such as for interactive applications. Developing a system that can carry a wide range of synchronized metadata provides a means for minimizing conflicting cues between the video and the graphics overlays. This, in turn, potentially reduces the unpleasant side-effects from which some people may suffer with longer duration use of mismatched 3D graphics embedded in 3D video.

CLOSING REMARKS

In this paper we have looked at the reasons why the STB needs to be aware of the S3DTV content it is handling, and discussed the basic updates that are required. These start with simple changes to ensure that graphics and OSDs are drawn in a readable fashion, placing them correctly in depth and finally looking at how they can be designed and generated to give a true S3DTV experience. In a similar fashion, we have also looked at issues with processing the video, and explained the limits on performing operations that are simple with 2D but difficult or less desirable to do in S3DTV.

We have also looked at the areas where additional signaling is required, or beneficial. This starts with the S3DTV format in use, and moves through the signaling of depth

information, and discusses some additional metadata that may be useful for graphics.

In closing, it is worth emphasizing how popular S3DTV is likely to be for many viewers. These viewers however will not and should not be aware of the issues outlined in this paper. They will expect everything that they are already familiar with in their television experience to work perfectly in 3D. The areas we have discussed in this paper allow these expectations to be met and help the STB provide its part in the compelling experience of S3DTV.

REFERENCES

1. High-Definition Multimedia Interface Specification Version 1.4a (extract of 3D Signaling Portion)

2. ISO/IEC 14496-10:2009 Amendment 1: Constrained baseline profile, stereo high profile and frame packing arrangement SEI message

3. Technicolor Brings 3D to the Home and Beyond, January 2010

THE USE OF SCALABLE VECTOR GRAPHICS IN FLEXIBLE, THIN-CLIENT ARCHITECTURES FOR TV NAVIGATION

Michael Adams
Solution Area TV, Ericsson

Abstract

Today's subscribers are demanding more and more from their service providers:

- *Personalization: New behaviors from a new generation of "digital natives", (who expect the service to adapt to them!), powerful search capabilities, and recommendations engines.*
- *Communication: Multitasking, social networking, and sharing the viewing experience through chat and instant messaging*
- *Interactivity: Polls, games, and enhanced programming*

And subscribers want all the above services and features to be delivered as a single, integrated service experience across any device, anywhere, and at anytime!

The Internet has shown how to deliver all kinds of services by means of thin-client approaches using the Representational State Transfer (REST) model. Meanwhile, most deployed cable architectures still rely on a "state-full", thick-client approach.

Can thin-client architectures really satisfy large cable system requirements for performance, scalability, high-availability, emergency-alert system requirements, and compatibility with existing interactive application environments such as EBIF?

This paper will show how a thin-client, browser-based approach can:

- *Support rapid development of new applications without the need for new software download to the set-top*
- *Enable personalization of a service to each subscriber's preferences*
- *Allow full customization of the user-interface, including branding*
- *Allow the use of third-party developers, using Web 2.0 service creation methods*
- *Provide set-top independence and multiplatform portability*
- *Decouple CA/DRM certification from new features and applications development*

INTRODUCTION

Today's subscribers are demanding more and more from their service providers. What they want can be grouped into three main categories:

- Personalization
- Communication
- Interactivity

Personalization

- I want my service to adapt to my needs.
- I need recommendations to sort *my* "wheat" from the "chaff".
- I want to watch anything, on any device, at any time, anywhere.

Personalization provides the subscriber with a better experience, tailored to their needs, and creates "stickiness" for the operator.

Communication

- I want to share my experience with my friends in real-time
- I want to interact with social networking sites
- I want to be notified when I have a phone call and to be given an option to pause my movie if I choose to take the call.

Communication is a basic human need. Many subscribers are already using a laptop while watching TV to turn viewing into a social experience. The TV experience can be extended to allow all subscribers to communicate while they are watching TV to a greater or lesser degree, depending on their preferences.

Interactivity

- I want to request more information about products when they are advertised.
- I want to be able to go directly to the movie after I see the promo.
- I want to be able to play along with my favorite game shows
- I want to be able to vote online about important issues.

Numerous studies have shown that interactivity can help to keep subscribers engaged. Interactive programming includes such things as play-along game shows, audience polls, and the like. Interactive advertising is also an important opportunity.

Interactivity is more natural on devices like a mobile phone or tablet PC than the TV, because these devices can allow more natural user-interfaces through the use of touch screens.

User Interface Requirements

In summary, subscribers are ready and willing to extend their TV viewing experience. There is a new generation of savvier, more educated subscribers (often called "Digital Natives" [1]) that desire new features. Nevertheless, TV is still a living room experience and pure web-style navigation doesn't work.

Personalization, communication, and interactivity require a dynamic, flexible, extensible, high-performance user-interface. Moreover, as the subscriber starts to like, and take advantage of new features, the infrastructure that supports them must be scalable. The system cannot slow to a crawl if everyone presses the "red button" at the same time (for example, during a Super Bowl commercial).

CURRENT USER INTERFACE

Today, most set-top box (STB) "guides" are limited in their ability to support the requirements of personalization, communication, and interactivity because of the way they are developed:

- The user-interface logic is embedded into a monolithic "resident application" that is downloaded to the STB. Because any changes require extensive testing before they can be unleashed on unsuspecting subscribers, each subscriber ends up getting the same guide as every other subscriber.
- The user-interface is developed in a low-level language such as c, c++, or Java. Highly-paid software developers are needed to modify the user-interface or to create new applications. Any changes must be designed, approved, developed, tested, certified, and signed-off; a process that can take 6 months or more. The result is expensive applications that arrive to market late or not at all.

Embedded user-interfaces were the only option when STBs had a small memory footprint and limited CPU power. For

example, in 1998, a typical STB had only 1-2 MB of memory and a 27 MHz CPU [2].

Today, a system on a chip (SOC) can provide extensive capabilities, and these limitations no longer apply. For example, the Broadcom BCM7400 includes a "dual threaded 350-MHz MIPS32 with FPU class CPU" [3]. Typical memory footprints are at least 256 MB. Despite these changes, the thick-client model has persisted.

USER INTERFACE ALTERNATIVES

Over the past 15 years the user-interface model has been revolutionized by the Internet and the World Wide Web, and the thin-client model has become more and more popular. Netbooks are now the fastest growing category of portable computing device, requiring only a browser on the client while the "heavy-lifting" is done by servers in the network.

The advantages of the thin-client model are:

- Extensible; new user-interface logic can be introduced at any time merely by linking in additional web pages.
- Dynamic; changes in the application can be made without a STB firmware reload or reboot.
- Rapid authoring of new applications by graphic designers (not software developers) with a much faster, shorter testing cycle. The development cycle of new features can be reduced from months to days.
- High availability and horizontal scalability is achieved by means IP load balancing.

The limitations of the Internet model are:

- Performance – even a small round-trip latency makes the user-interface unacceptably slow if a synchronous execution model is used.
- Reliance on a guaranteed high-bandwidth communications path between the server and client. This model cannot continue to provide basic navigation when the return path is lost.
- Need for high-performance server "farms" to keep up with client transactions.
-

TOWARDS A THIN-CLIENT SOLUTION

The advantages of a thin-client solution are attractive; however the limitations must be overcome to make this a viable model for the TV.

The goal of this paper is to demonstrate that careful system design can yield a thin-client solution with all its advantages and none of its limitations.

The proposed solution is based on the following:

1. Asynchronous execution model
2. Browser-based STB software environment
3. Scalable vector graphics
4. Sophisticated authoring environments that support rapid development of new applications
5. Personalization through different user-interface styles
6. Multi-level caching for better user-interface performance and lower-latency

Asynchronous Execution Model

One of the main limitations of early client-server implementations was that they were synchronous, waiting for one event to complete before starting another one. The effect is that all of the network latencies are serialized and can add up to a slow user-experience.

This problem has been addressed by the AJAX (Asynchronous Java script And XML) client-server model, which supports asynchronous execution. (See Figure 1)

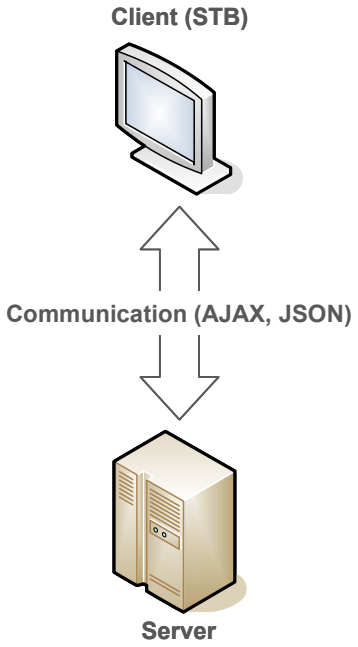


Figure 1: Client-Server Model

The client-side implementation is typically (but not constrained to):

- HTML (HyperText Markup Language) pages, which are navigated by the user as she interacts with the user-interface.
- JavaScript, which provides embedded functions that interact with the Document Object Model (DOM) of the HTML page. Because JavaScript code runs locally in the client environment, user-interface functions can be very responsive.
- CSS (Cascading Style Sheets) allow the format (look and feel) of the user-interface to be separated from the business logic. This allows rapid customization of the graphical aspects (color, fonts, and layout).

Communication protocols are:

- Client-server messages are carried by HTTP using the AJAX format.
- JSON (Java Script Object Notation) is used as data-interchange format.

This standards-based architecture allows each JavaScript object to be transferred asynchronously to the client, eliminating the serialization of transactions and increasing performance of the user-interface.

The server-side implementation is not constrained at all. However, Java Enterprise Edition Server (still commonly known by its old name of J2EE) provides a useful container-based environment.

Browser-based Set-top Box Software Environment

The client that runs on the STB is an advanced web page. Client-server communications are handled by JavaScript and AJAX technology (as previously described). The software framework, both on client- and server-side, is responsible for loading any required JavaScript objects.

The framework also wraps the STB-specific JavaScript Application Programming Interfaces (APIs), enabling new applications to run on any compliant STB.

Advantages of this approach are:

- Allows the use of third-party developers, using Web 2.0 service creation methods.
- Provides set-top independence and multiplatform portability.
- Decouples CA/DRM certification from new features and applications development.

Figure 2 shows the STB software stack. The key points (highlighted on the diagram) are:

- Network Interface – the network interface is based on standard IETF protocols such as SIP, RTSP, and IGMP.

- IMS Applications – IMS (IP Multimedia Subsystem) is used to enable application functions such as presence and messaging. This provides a standards-based way of supporting convergence applications such as caller-id. (IMS is also the foundation of the PacketCable 2.0 specification.)
- Blended Services – the browser runs JavaScript applications and leverages the underlying services layers.
- Characteristics – each STB must meet a baseline set of parameters (such as CPU, memory, and graphics capabilities) to ensure correct performance.
- Browser interface – a set of plug-ins to the browser to allow rapid porting to new STBs.
- Browser capabilities – this will be described in the next section (scalable vector graphics).

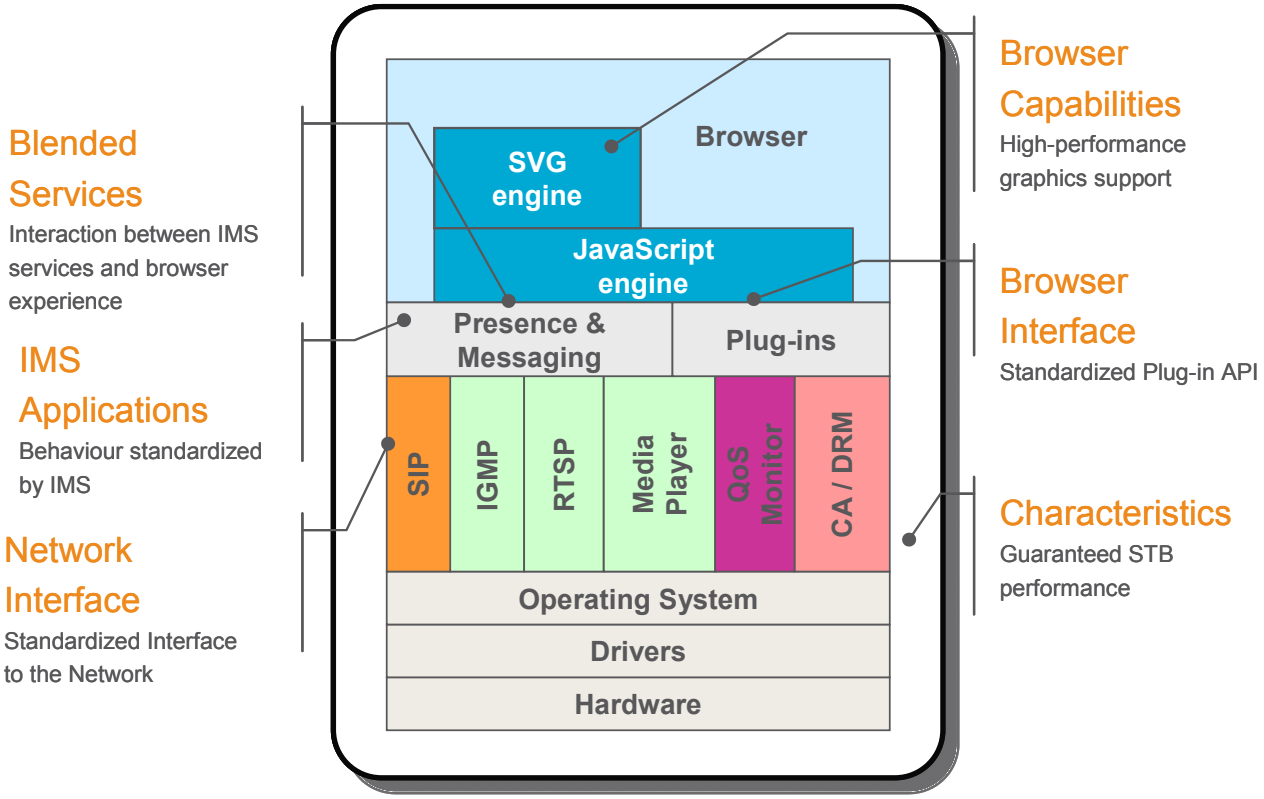


Figure 2: Set-top Box Software Environment

Scalable Vector Graphics (SVG)

SVG is a graphics file format and web development language based on XML. It can be used to create graphically-rich user-interfaces which include animations.

SVG provides similar user-interface functions to Adobe Flash but differs in that it is an open standard that has been under development by the World Wide Web Consortium (W3C) since 1999 [4].

SVG describes two-dimensional graphics and graphical applications in XML. SVG graphics do NOT lose any quality if they are zoomed or resized. Most browsers now support SVG.

Authoring tools

The operator must be able to rapidly introduce new applications, for example (see Figure 3):

1. Casual Games

2. Video Art
3. Real-time Information
4. Web 2.0
5. App Store
6. TV Communities

These example applications can be created by web-applications developers because the environment is identical to that for web development.

Figure 3: Application Examples

Personalization

Different user-interface styles (see examples in Figure 4) can be supported for different user groups or geographical areas.

As previously explained, cascading style sheets (CSS) make it easy to change the appearance and behavior of the HTML pages that define the user-interface.

At the server-side there are other significant advantages to this approach:

1. Different user-groups can have their own unique user-interface style, based on the subscriber profile. For example, a hotel system user-interface may be designed to look completely different from a residential user-interface.
2. Minor changes in appearance and behavior can be made quite simply and rapidly. They are first tested on a lab system, and then with "friendly" subscribers, before being rolled out to the entire subscriber base.
3. Because there is no resident application in the STB, a new user-interface version is published in the same way as updating a web-page. STB firmware downloads and reboots are almost completely eliminated. (The only exception to this is when a browser update is made.)

Figure 4: Personalization of the User Interface

Multi-level Caching

Referring to Figure 5, a multi-level caching scheme is used to improve performance and reduce load on the servers as follows:

1. The browser caches recent HTML pages so that they are retrieved locally if the subscriber goes back to them. The program guide is a good example of this.
2. An HTTP cache stores commonly accessed pages for rapid retrieval without generating any transactions back to the server.
3. A "portal" backend in the server is pre-formatted to increase performance.

It should be noted that all of the cache entries have a time-to-live (TTL) to ensure that information does not become stale. When the TTL timer expires the cache entry is invalidated and the next user request causes a cache-refresh.

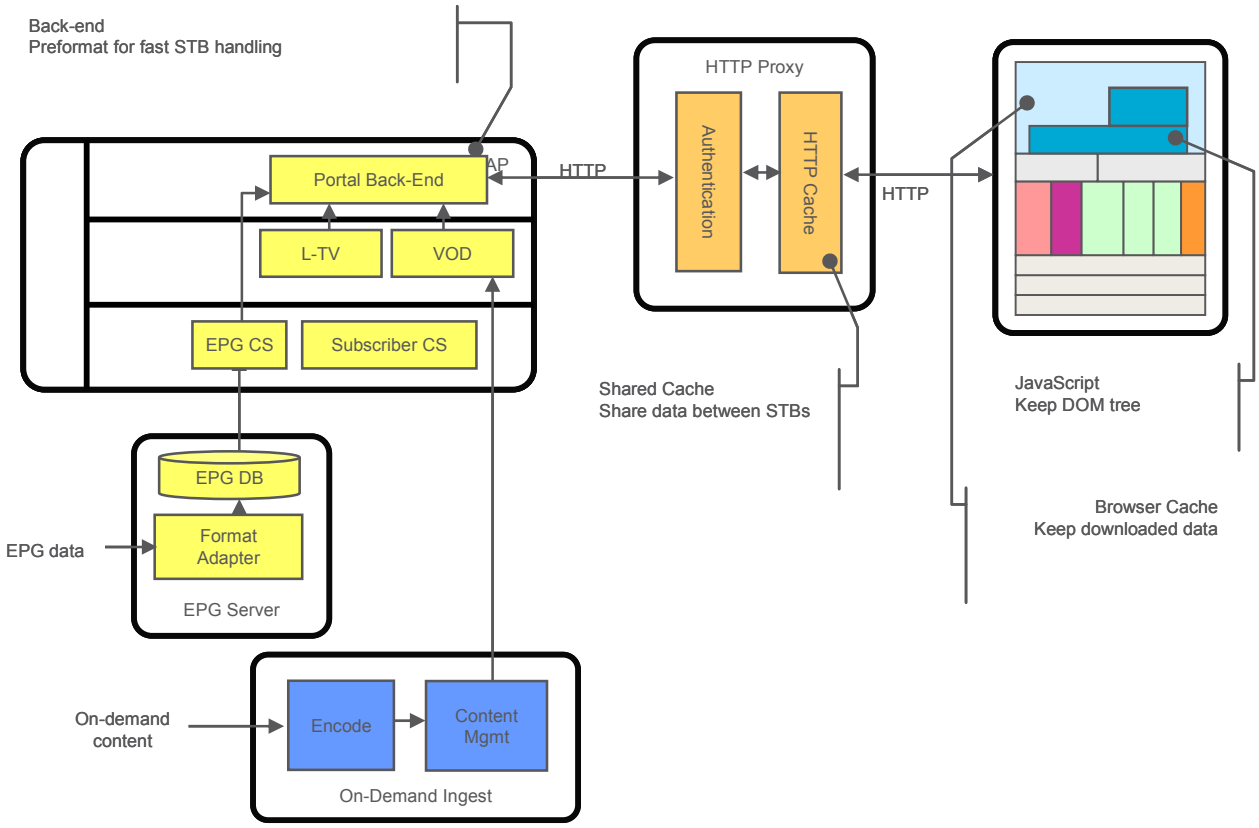


Figure 5: Multi-level Caching Hierarchy

CONCLUSIONS

In this paper we have described how a browser-based, thin-client approach can be used to support a rich, graphical user-interface for the STB.

Making this transition brings with it a number of very significant advantages to the operator:

- Support for rapid development of new applications without the need for new software download to the STB.
- The use of third-party developers, using Web 2.0 service creation methods.
- Personalization of the user-interface.
- Allows full customization of the user-interface, including branding.
- Set-top independence and rapid STB porting through a browser-based approach and plug-ins.
- Decoupling of CA/DRM certification from new features and applications development
- Scalability and performance through the use of multi-level caching strategies.

References

[1] *Digital Natives, Digital Immigrants*, Marc Prensky, 2001

[2] *OpenCable Architecture*, Cisco Press, Michael Adams, page 131

[3] *Broadcom Corporation*, http://www.broadcom.com/products/Cable/Cable-Set-Top-Box-Solutions/BCM7400

[4] *World Wide Web Consortium*, http://www.w3.org/Graphics/SVG/

TRANSPARENT AND REASONABLE NETWORK MANAGEMENT IN DOCSIS 3.0

Don Bowman
Sandvine

Abstract

The high-speed data industry is in a state of flux due to maturing markets, increased regulation, public policy pressure, transparency requirements and DOCSIS 3.0 speeds. With the wide-spread adoption of broadband, and its evolution into wideband, we are seeing the relentless destruction of information-based business models as information shifts to Internet-based delivery. Newspaper media giants are going bankrupt, broadcasters are cutting out their affiliates to deliver directly to the consumer, and the music and movie industry are challenged. Delivery models are shifting to the Internet because of the low cost to the content provider. At the same time, top-line subscriber growth on high-speed data is declining and peak bandwidth is not able to command a proportional premium in price.

This paper will discuss the technical means of achieving the goals of providing continuing, profitable high-speed data service with transparency and reasonable network management, while transitioning consumer experience and expectations to the broadband of tomorrow made possible with the speeds created with DOCSIS 3.0.

NETWORK MANAGEMENT FORCES

No aspect of the evolution of the Internet has ever been slow and quiet. This is as true today as it was in 1969 with the first ARPANET nodes in UCLA and Stanford. However, there is a sea-change happening in that the evolution is now being driven not by common-shared interests amongst a small community, but by divergent interests driven by business, government, and content owners versus content consumers. In "Tussle in cyberspace: defining tomorrow's internet" ([1]), Clark et al discussed the thesis that the Internet will be increasingly defined by tussles that arise among the varying parties with divergent interests, rather than the common shared interest that drove its initial design and evolution, and this in turn drives one of the major set of forces.

Technology forces also play strongly in network management. DOCSIS 3.0 provides wideband access speeds of up to 160Mbps to hybrid-fibre-coax cable networks. As DOCSIS 3.0 is adopted, this in turn changes the oversubscription ratios and 'fairness' concepts between users of DOCSIS 1.0, 1.1, 2.0, and 3.0 (as all may share some portion of the RF spectrum). It also, in the short term, drives an increasing disparity between upstream speed and downstream speed.

Media and content companies have a dichotomy of approach towards network capacity and network management. From a copyright-infringement and digital rights standpoint, they would prefer network management to pro-actively block certain content. From a distribution standpoint, they would prefer for infinite capacity for zero cost (to them) to exist. As a consequence, they both lobby for, and against, network management. For an MSO deploying DOCSIS 3.0, this direct distribution of content (e.g. Hulu) in turn reduces revenue, shifting it from low-cost switched digital video towards high-cost packet-switched infrastructures, while at the same time reducing both subscription and advertising revenue.

The Internet has always resisted regulation and censorship. "Neo-luddism and the demonisation of technology: cultural collision on the information superhighway" ([2]) suggests that "Just as society did not cry for

an end or regulation of the printing press, so too should we not regulate the Internet", and so too here should we not call for regulation to arbitrate these forces.

TRANSPARENCY HELPS DEFINES REASONABLE

A 'reasonableness' test helps define the acceptability of network management. This test stems from the common-law concept of 'what would a typical person agree is reasonable'. It is by consequence a subjective fine-line test.

In the opinion of this author, the best means of defining reasonable network management is by a combination of transparency and contract. If a network management policy is disclosed in such a way that a typical consumer can understand it, and if that same typical consumer then purchases access to that same network, they have de facto defined the practice as reasonable.

Transparency is a challenging concept. The subtle technical nuances of DOCSIS networks (latency, loss, jitter, shared-access, etc.) are difficult to describe in simple enough terms that the layman can understand. Analogies, although helpful to form a basis, rapidly become inappropriate as they diverge from the original problem. Network management practices evolve over time and should not be hampered by overly detailed disclosure and discussion. Thus is becomes important to disclose what is material to understanding a network management policy. Since we are relying on transparency as a means of supporting reasonableness, material becomes any aspect that would affect the actions of the typical consumer.

It is also important to not overly pander to the least-technical nor the most-technical of user of the network. The typical consumer is a concept which has evolved throughout the history of residential Internet access from the earliest days of the enthusiast to today's mass-market penetration. The typical consumer is the one that an MSO targets with television and print advertising.

If typical users, understanding the disclosed network management policies in use, contract for the service, the policy must be reasonable by definition.

Reasonable is defined entirely in the frame of reference of the end-user, the customer of the MSO.

SUCCESS CRITERIA

In order to be successful, a reasonable network management plan must maximise the user experience of the maximum number of users for the maximum amount of time for a given capital investment. It must do so without sacrificing subscriber growth due to competitive forces. It must do so without falling afoul of public policy and regulation. This is a tough set of bounds to stay within, but it is possible.

An acceptable and successful network management plan will take into account the following focus areas.

Narrowly-Tailored

All networks have variation in usage patterns, whether by time of day, by geography, by user demographics, or by other factors. As a consequence, oversubscription and quality of experience are non-uniform across the network. A properly constructed network management plan takes this into account, and focuses as narrowly as possible on the problem to be solved. It does not try to force a one-size-fits-all solution into all areas at all times.

In a DOCSIS 3.0 environment, there are several areas of 'narrowly-tailored' that might be technically considered for addressing subscribers who are causing disproportionate

congestion in times of congestion. These include:

1. Channel bonding and how channels are shared with un-bonded users
2. Mixed DOCSIS 1.0, 1.1, 2.0, and 3.0 usage within the same shared spectrum
3. Subscriber density per node
4. Subscriber demographics per node
5. CMTS backhaul network capacity
6. Unforeseeable events

A reasonable network management practice takes these factors, and more, into account. It applies itself differently, or not at all, depending on the conditions which are currently present. For example, a network management practice might be self-tuning, and disable itself when no congestion is present. It might operate differently when congestion is present on a single user, versus a single RF channel, versus a bonded set of RF channels, versus the CMTS backhaul uplink.

A successful network management practice will narrowly-tailor itself to the situation at hand at the time it is needed. It will not apply in a broad fashion across the broad average of a network.

Proportional and reasonable effect

The impact of a reasonable network management policy must be proportionate to the problem being solved. It would be considered unreasonable by most to take a subscriber causing 15% of the congestion on a network, and manage their bandwidth to 1% of peak rate. It might, however, be considered reasonable to reduce the priority of traffic of the top twenty-percentile of bandwidth users, which as a group might be only 5%, but consume more than half the bandwidth at a given point in time. In reducing their scheduling priority they do not affect the latency and throughput of the other 95%. The network management policy needs to take into account the concept of proportional effect and response.

Legitimate and demonstrable technical need

The operator must have a legitimate and demonstrable technical need for the network management practice. The architectural strengths and weakness of DOCSIS provide the majority of the technical needs for network management, and these are discussed in detail in "An Overview of the DOCSIS (Cable Internet) Platform" ([3]). Additional technical needs arise due to network architecture outside the scope of DOCSIS, for example, implementation-specific details of various CMTS vendors, of backhaul and core network architecture.

To be successful, a network management practice must be described in such a way that it is clear the technical need that is being addressed, and that it is clear the practice is designed to address this need and nothing more.

Transparent disclosure

The operator must make the material information to understand the network management policy publicly available to those impacted by it. The disclosure should be sufficient for a consumer to form an informed opinion on whether the practice will affect them, which applications might be affected, when they might be affected, and what the impact might be, including impact to speed and latency and general experience.

Disclosure might take many concurrent forms. The most popular include network management FAQ web pages, notices included in billing material, acceptable use policies, terms of service, etc. In "Virgin Media Broadband: Traffic Mangement" ([4]),

Virgin Media describes their network management practices in very explicit form. In [5] Cox Communications describes theirs in more general terms. Both provide an end-user with information to understand how the practice will affect them, and both provided pro-active notification to their users in addition to the FAQ web page.

Auditable and demonstrable

Owing to the public scrutiny of capital investment in networks, and network management policies, it becomes important for an MSO to be able to demonstrate that the above criteria were indeed met.

On audit, and MSO should be able to provide:

1. What was the technical need that caused the creation of the network management policy

2. What affect the policy had on the user experience

3. How they have disclosed their policy to the end-user

4. How the policy took into account network and time variances

In addition, the audit should be able to demonstrate the above were met using technical results. These results might include information on the user experience for the typical user for typical locations in the network.

DOCSIS 3.0 CHALLENGES

DOCSIS 3.0 has enjoyed a rapid rollout, giving a large number of consumers access to a wideband experience. It has also created specific challenges for network management policies. Specific network-policy control challenges include:

1. Mixed use of RF spectrum between DOCSIS 2.0 and 3.0 users

2. Availability of downstream channel bonding in advance of upstream channel bonding

3. Lack of DOCSIS 3.0 IPDR availability from CMTS vendors

4. Dynamic channel bonding groups and external signaling of change events

5. Higher oversubscription rates stemming from increased offerings

6. Lack of ability to prioritise traffic in downstream in all CMTS with PacketCable Multimedia

7. Increased burden on backhaul network

As a consequence, this has increased the complexity of network management technology.

In a mixed DOCSIS 2.0 and DOCSIS 3.0 environment, a successful network management implementation needs to be able to:

1. Prioritise per user per RF channel and per bonded channel group in the downstream above the CMTS

2. Prioritise per user per RF channel and per bonded channel group in the upstream in the cable modem and DOCSIS scheduler

3. Detect near congestion state in near real time

4. Detect the users and applications causing the disproportionate congestion in near real time

5. Interact correctly with temporary speed changes such as PowerBoost™, edge-QAM resource management

6. Provide strong reporting and business intelligence to be able to provide accurate capacity planning and auditable results

7. Provide strong subscriber experience measurements

Many MSO are now also using non DOCSIS access technologies such as WiFi, 3GPP HSPA, fibre PON, WiMax, so network management may also need to operate in an access-agnostic fashion.

CONCLUSIONS

Despite its higher speeds (in fact, because of them), active network management is required in DOCSIS 3.0 to maximise the experience of the maximum number of users for the maximum amount of time.

Network management policies must be narrowly tailored, must be proportional and reasonable, must be designed to address a legitimate technical need, must be transparently disclosed, and must be auditable.

Reasonable network management requires disclosure of the policy in such a way the typical user can understand the impact to them, and reasonableness is framed entirely from the end-user perspective.

Access-agnostic network policy control is required to create a network management practice that spans multiple devices, and multiple access technologies.

Strong reporting and business intelligence is required to be coupled to the network management practice to understand demand, capacity, and user experience.

As an MSO deploying DOCSIS 3.0, this may seem like a minefield of requirements, but a few simple up front planning activities can make for a highly successful network management practice.

[1] Clark, D. D., Wroclawski, J., Sollins, K. R., and Braden, R. 2005. Tussle in cyberspace: defining tomorrow's internet. IEEE/ACM Trans. Netw. 13, 3 (Jun. 2005), 462-475. DOI= http://dx.doi.org/10.1109/TNET.2005.850224

[2] Bowman, D. 2009. Neo-luddism and the demonisation of technology: cultural collision on the information superhighway. SIGCOMM Computer Communications Review. 39, 3 (Jun. 2009), 19-21. DOI= http://doi.acm.org/10.1145/1568613.1568618

[3] Tooley, M., Bowman, D. 2008. An Overview of the DOCSIS (Cable Internet) Platform. http://www.tiaonline.org/gov_affairs/fcc_filings/documents/Cable_Architecture_Declaration_01.14.10.pdf

[4] "Virgin Media Broadband: Traffic Mangement", http://allyours.virginmedia.com/html/internet/traffic.html

[5] "Cox Communications – Congestion Management FAQs", http://www.cox.com/policy/congestionmanagement/

[6]

TWO YEARS OF DEPLOYING ITV/EBIF APPLICATIONS – COMCAST'S LESSONS LEARNED

Robert Dandrea, PhD
Distinguished Engineer
National Engineering & Technical Operations
Comcast Cable

Abstract

Comcast has deployed several EBIF-based Interactive TV applications very broadly to its footprint in the past two years. The goal of this paper is to share the technical challenges and lessons learned from this broad deployment experience. Besides reviewing some video platform design considerations that have evolved from this experience we also describe some work done collaboratively with Canoe and Cable Labs that also has aided in creating a scalable EBIF delivery ecosystem.

One challenge has been in extending Comcast's operational support system to cover these new applications. The paper will therefore discuss both platform and process advances that will enable us to support iTV broadly across our footprint. The paper also discusses how Comcast manages scalability as the number of applications and the number of enhanced networks increases.

Comcast's iTV Experience

Over the past two years Comcast has deployed several EBIF-based interactive TV (iTV) applications very broadly throughout its video network. Our currently deployed EBIF user agent – the set top box (STB) client that executes the applications – is compatible with the I05 EBIF specs ([1], [2]) and is deployed on 21 million STBs. Our initial wave of EBIF applications are available in 12 million subscriber homes. The most broadly deployed applications include

1) Caller ID – Comcast digital voice and digital video customers can receive a brief banner popping over any currently viewed video on their TV identifying incoming phone calls.
2) Home Shopping Network's "Shop By Remote" – a bound iTV application carried 24x7 in the HSN video feed which allows viewers to purchase directly through their TV.
3) EBIF enhancements bound to 30 second advertisements locally inserted (via SCTE 30/35) by Comcast's Ad sales team, Spotlight. These include the "RFI – Request for Information" application that offers viewers a direct mail or phone call response from the advertiser and the "Ready – Remind/Record" app that allows a user to automatically set DVR recordings or guide reminders during commercial breaks that advertise upcoming programming.

Currently we are readying more applications for customer deployments, including several bound applications carried in national programming (including some done in partnership with Canoe Ventures, and several unbound applications available over any programming, including enhancements to our Guide and Instant-Info, which is a "News-Weather-Sports" widget-like app.

In the remainder of this paper we shall review some of the major design aspects of

our video delivery platform that help to enable a reliable full-footprint deployment of these applications. One of our major lessons

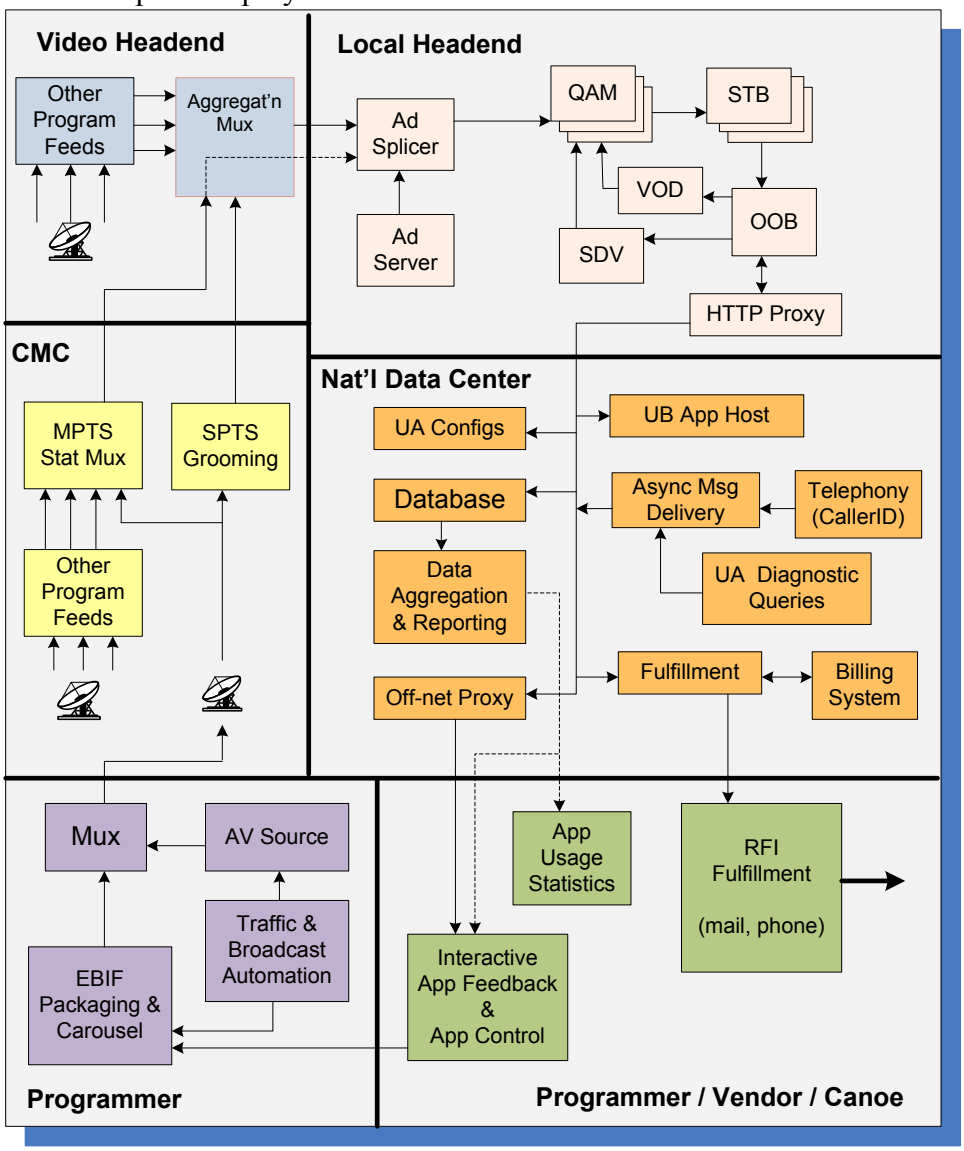


Figure 1: Comcast EBIF Platform - Simplified Functional Architecture

learned has been the need to create an up-front design that meets the operational requirements of reliability, verifiability and problem isolation, we stress platform aspects that contribute to those goals. We also summarize some work done collaboratively with other MSOs, Canoe Ventures and Cable Labs that also has aided in enabling a scalable deployment of many apps across many enhanced networks.

A PLATFORM FOR RELIABLE AND SCALABLE EBIF DELIVERY

Figure 1 shows a simplified functional overview of the Comcast video platform for

delivering both bound and unbound EBIF-based iTV applications. In the following discussion we shall look at this architecture assurance and will enable us to scale to deliver dozens of different applications over dozens of different networks in the near future.

Programmer

The "Programmer" block in Figure 1 shows a typical architecture at a programmer's broadcast center for delivering a bound EBIF application over a nationally distributed video network. That platform consists of

1) an EBIF packaging system (to translate graphic app directives described in proprietary schemas used by application developers into normative EBIF resources).
2) a data carousel to encapsulate those resources into MPEG packets and spin those packets for inclusion in the full transport stream. Generally we have found 100-200 kb/s is sufficient for simple EBIF apps.
3) integration with a broadcast automation system so that iTV applications can be constrained into or out of commercial breaks.
4) the standard AV source and a mux to combine it with the EBIF data carousel.

It has taken a significant effort over the past few years to develop an EBIF broadcasting system simple and reliable enough to be integrated with a programmer's existing AV broadcast gear. Syncing nationally broadcast apps with commercial breaks (to remain in the breaks for enhanced advertisements and to remain out of the breaks in the case of enhanced programming) has been a particular challenge. We have used automated signaling such as GPI or SNMP emanating from the programmer's traffic automation systems to keep EBIF apps in or out of commercial breaks. Initially we have block-by-block, and analyze some of the platform elements that aid in EBIF service

used Comcast-developed subsystems for (1) and (2) above, but in the past year we have worked more collaboratively with vendors and with Canoe to standardize this part of the platform.

Figure 2 shows a sample system design we have used for managing the synchronization of applications across commercial breaks. In this example we've used our own internally developed servers for application packaging and carousel streaming. The design shows the thoroughness at which redundancy is implemented and the manner in which a GPI (general program interconnect) signal from a video broadcast automation system is used to disable and enable an EBIF application. The design shows an app server communicating with a primary and backup app packaging system, and both packagers sending XML descriptions of EBIF page and data resources to both a primary and backup carousel server, where the XML directives are transformed into an MPEG carousel of EBIF and EISS data pids. Both SD-East and SD-West (and potentially HD) versions of the service are simultaneously enhanced. The GPI signals are then used by a redundant pair of ETV filter servers to filter off the EISS signaling passing through them during an ad break. Both primary and backup filtered EBIF carousels are relayed to a mux, where only the primary is used until failover. The GPI triggers need to occur 5 seconds prior to the start of an advertising break due to the EBIF spec requirement that user agents suspend applications (remove them from the TV's on-screen display) within 4 seconds of loss of EISS signaling.

As mentioned above, we have more recently used vendor's EBIF packaging and carousel systems at programmer broadcast systems, and have in this case employed a design for synchronizing the applications to

commercial breaks which relies on the ad break signaling altering the app packaging in a manner to temporarily suspend the app.

Another engineering problem within the programmer's broadcast system that needed to be solved was the determination of a

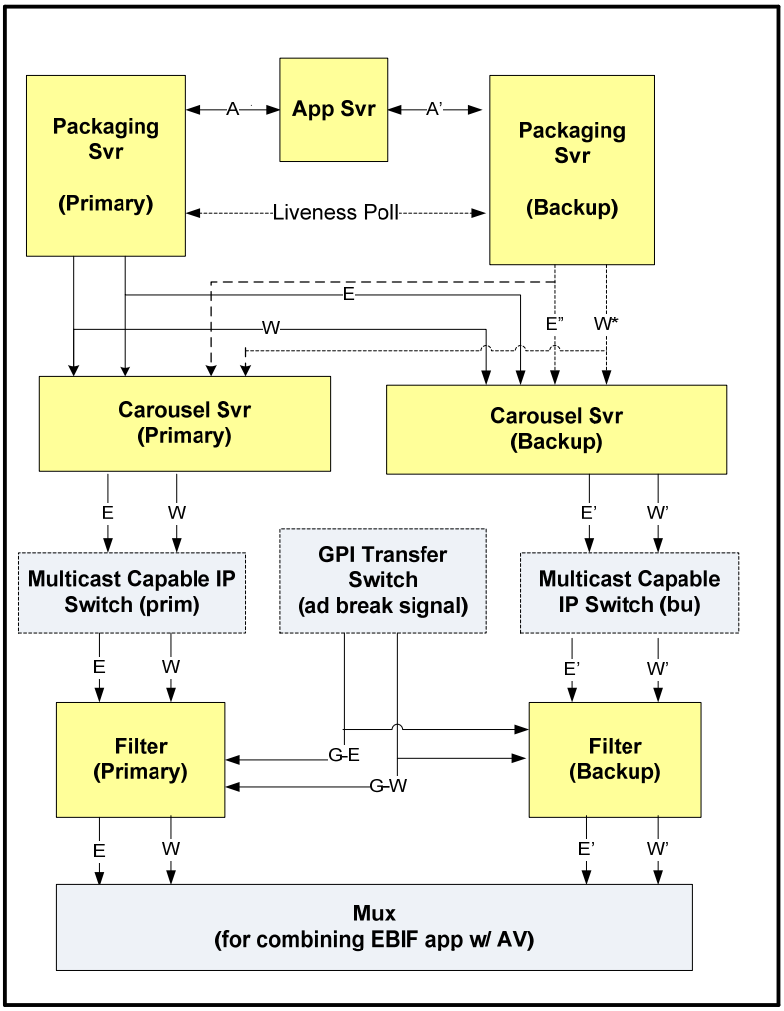


Figure 2: Sample design for EBIF broadcast system synced with ad breaks

mechanism for a programmer to enhance their broadcast feed for some markets or some MSOs without impacting other markets and other MSOs that do not receive the app. We have seen two different general solutions to this: (i) If the programmer has spare satellite bandwidth, they can replicate the program feed on a whole new channel. (ii) More generally the programmer has simply added the new iTV data pids to their single production feed, and added a secondary PMT that includes those new pids (but shares same audio and video pids as primary PMT). Markets / MSOs unauthorized for the app simply continue reading the old PMT without the iTV data pids listed. Other satellite receivers can be separately authorized for the new PMT.

Comcast Media Center (CMC) and Video Headends

The CMC is Comcast's distribution hub for much of its video content. The CMC delivers to local markets nearly all our VOD, HD and SDV content, and a good fraction of our digital SD content too. One major platform change that we are making, partly for the operational advantages it offers for bound application delivery is requiring all local markets to use CMC re-broadcast feeds for EBIF-enhanced programming rather than direct-from-programmer feeds. This simplifies and empowers our operational capabilities, in that a single monitoring point can verify the iTV application at that seminal point in its delivery path, and a single control point offers national on/off control of applications (via pid dropping) at a moment's notice. Furthermore, the use of a standard delivery path through the CMC shortens our application deployment timelines, as local markets needn't be concerned with some EBIF-centric broadcast issues described below.

Use of CMC re-broadcasts also helps manage the in-band bandwidth issues that may arise when multiple programs in a multiplex are enhanced with EBIF applications. The CMC generally offers two types of programmer re-broadcasts: SPTS and MPTS (single and multiple program MPEG transport streams). The SPTS are constant bit rate encoded streams, with the bitrate fixed and common across all streams for bandwidth management reasons (e.g. for Switched Digital Video, SDV). Thus any EBIF app data will simply consume part of the total available bandwidth allocated to that service. A more efficient scheme, however, is to use stat-muxed MPTS transport streams that have been encoded using state-of-the-art multipass MPEG-2 encoding gear that can more optimally find bandwidth space for the EBIF apps in a shared manner across the full multiplex. For our more popular (non-SDV) content, that is our path forward.

Figure 1 shows four different manners that these CMC re-broadcast feeds are treated at the aggregation mux in the market video headend:

1) The SPTS feed can be muxed with other programs into an MPTS feed without any further grooming, thus maintaining its CBR nature. This is the delivery path for services on that are switched onto the RF plant by SDV.
2) The SPTS feed can be re-groomed with other programs into an MPTS stat mux feed.
3) The local aggregation mux can cherry-pick one or several programs out of the CMC MPTS feed, and re-groom them into another MPTS stat mux.
4) The full MPTS from the CMC can be used by the local market without any further grooming.

As described above method (1) above is used for SDV feeds while method (4) above is our path forward for non-SDV services, as it most efficiently uses the available 38 Mb/s of a standard QAM-256 channel. Delivery methods (2) and (3) are only interim steps while moving fully to CMC rebroadcasted programming (and, in the process, dealing with technical challenges like local blackout requirements in some programming).

One technical challenge has been the creation of an encoder/mux system for grooming enhanced programmer feeds that does not substantially de-sync the EBIF app from the underlying audio-video when the video is groomed. Since the EBIF spec uses no timing synchronization to an inherent transport stream timeline like a PCR, proper synchronization of the EBIF app with the underlying audio/video must be maintained carefully when the video is regroomed after the app is embedded. By requiring a maximum of 1-2 seconds of downstream de-

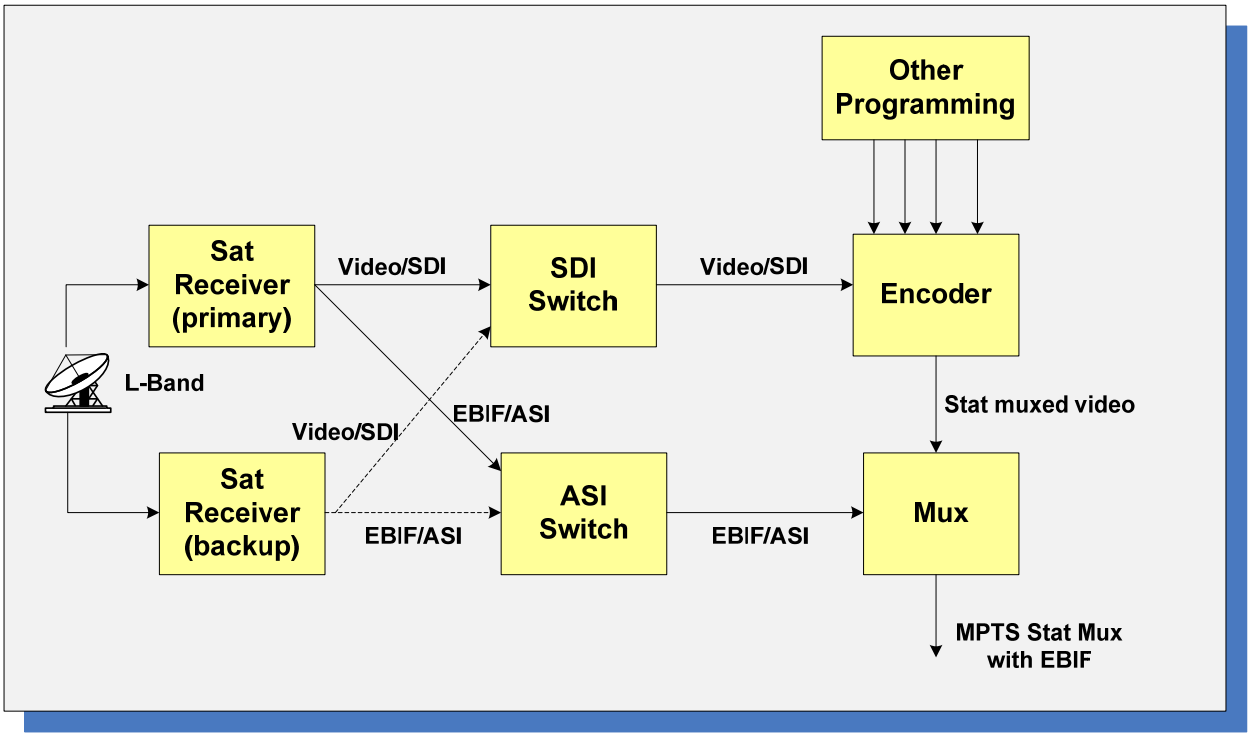


Figure 3: Sample system design for re-grooming EBIF enhanced programming

sync due to further grooming, and requiring that the app is designed so that it can suffer a 1-2 second de-sync without any adverse consequences (e.g. the design for synchronizing the app with commercial breaks has built into it these 1-2 seconds of potential de-sync), we have been able to manage this problem.

A sample design for video grooming of an enhanced service is shown in Figure 3. That design shows an L-band satellite signal as input that feeds two redundant satellite receivers that each have an SDI output for video and an ASI output for the EBIF and EISS app data. SDI and ASI switches then support failover between the two sources. The video is groomed into a stat mux with several other programs, and the EBIF/EISS data is subsequently muxed in. The particular encoder used showed a worst case delay of about 1 second, which would de-sync the EBIF app and audio-video in a manner that would push the app forward up to 1 second in the programming as compared to its original position.

The complexity of the system in Figure 3 demonstrates another advantage to using CMC feeds rather than direct-from-programmer feeds in our local markets: If the market can use the full CMC feed without further grooming, the difficult operational work of building, verifying and maintaining a regrooming system that supports EBIF in each local market is not needed. Centralized distribution of already groomed signals offers a huge operational leverage.

In order to support the broad operational goals of verifying bound application delivery fully from app origin to STB and of enabling traceability of any problems, we are currently developing an MPEG monitoring system to

snoop EISS and EBIF data pids at critical junctures in our video delivery platform. The goal of the system is to improve application uptime by supporting both automated alarming when an application fails to be delivered past some network node, and by enabling post-flight analysis to help troubleshoot any problems that arise. Besides occurring at the centralized CMC re-broadcast point and at video headends, this monitoring can also occur downstream of local headend edge muxes / ad splicers.

A final aspect of aspect our headend-centric platform enhancements has been in standardizing mux configurations (for various mux vendors) to meet new requirements on these muxes stemming from iTV, such as the need to prevent un-authorized EBIF content from passing and the need to insert 24x7 EBIF-enabled placeholder PMTs in programming that will support local insertion of enhanced advertisements if the national feed is not already enhanced.

Local Headends

Comcast's current platform design for inserting local advertisements enhanced with EBIF interactive applications is one that minimizes the operational impact of those enhancements: we simply pre-embed the app into the MPEG asset for the ad before installing the ad on the local SCTE 30/35-based ad server. The impacts are then mux configurations as summarized above and verifying the ad server is updated with now-standard software to support the additional two iTV data pids. Before considering last-minute dynamic insertion of EBIF into local ads at the flight time of the ad, we have felt it wise to first perfect the EBIF package-carousel-mux systems operationally at national broadcasts centers before deploying hundreds of such systems at our edge ad zones.

One aspect of iTV application scalability has been in managing any impact the applications might have on our legacy Aloha (non-DOCSIS) out-of-band (OOB) bandwidth. One platform element that has contributed to that ability has been the buildup of monitoring gear on our OOB paths to measure the impact of these applications. Since we can only estimate concurrent application usage statistics, these measurements alarm us if iTV applications use an unexpectedly high fraction of our OOB bandwidth. We have not seen this occur thus far in our deployment history, but this mechanism should inform us of any need for OOB node size adjustments.

Although our tru2Way user agent can send HTTP direct from the STB over a DOCSIS channel, our legacy user agent does not and so we have developed a proxy that sits in each local headend where the interactive messaging from an application is changed to/from HTTP on the upstream/downstream. The communication between the STB and the HTTP proxy is via a proprietary protocol which adds reliable message acks/retries to the core UDP layer, and minimizes OOB bandwidth by allowing encoding of lengthy URLs and STB MAC addresses. As this HTTP proxy is the conduit through which all interactive messaging flows from legacy STBs at each headend, we have also built into our operational support system the ability to snoop and decode application messaging at this network element, filtering by application or by STB ID.

National Data Centers

Figure 1 displays several server systems deployed nationally to support iTV.

1) All unbound (UB) apps are served from the same web server ("UB App Host" in Figure 1), and this server supports

logging to track UB app usage and server scaling behind a load balancer.

2) Our user agent extracts a STB-specific configuration file from the "UA Configs" server at STB boot-time and daily afterwards. This supports many operational purposes, including enabling market-by-market (and even STB-by-STB within a market) level control over such aspects as which apps are authorized to run and application authorizations for STB resources (such as persistent memory).
   Within our current footprint of STBs with an I05 compatible user agent, we have developed a scheme to use the test-flag described in the I05 EISS (EBIF-AM) spec to manage our needs to limit app display within a market for trial purposes (e.g. limitation to only headend or employee STBs). That scheme has the 8 bits in the test-flag partitioned so that one bit is used for headend STBs, one bit for employee STBs, 3 bits are used for customer-phase rollouts of Comcast-internal apps and the remaining 3 bits used for customer-phase rollout of cross-MSO apps (e.g with Canoe Ventures). The fact that a single national broadcast of a bound app contains a single test-flag value that goes to all markets and to all MSOs limits the ability of this test-flag mechanism to support multiple trials of multiple apps on multiple networks. For that reason we describe future work to improve this platform functionality in the sections below.

3) We have built an asynchronous message delivery system that is shared across several interactive applications (such as Caller ID) and by an OSS server that performs user agent queries to help diagnose and isolate problems. We have also built a system to continuously inject synthetic transactions into this messaging platform throughout our full footprint, to proactively monitor and alarm for any connectivity problems.

4) Fulfillment for both Comcast-internal and Canoe-originated RFI applications occurs through a national server system that optionally may extract billing system customer identification info if the customer so grants.

5) All messaging from an app (including both user-initiated interactions and app-level usage logging that occurs regardless if users interact) flows through a proxy-point that serves several purposes operationally. For HTTP posts that flow to an external 3rd party host, the "Off-Net Proxy" performs the following functions:
   a) It relays the posts to 3rd parties, generally by adding encryption and doing obfuscation of private data (such as STB MAC addresses).
   b) It acts as a firewall to ensure communication is with an authorized host and that HTTP responses from such authorized hosts follow the technical requirements agreed upon (e.g. response size is within allowed range)

   For all posts (kept internal or relayed to 3rd parties) the national proxy point allows for entry of the Post data contents into our internal database. Queries to that database then allows
   a) aggregated app usage reporting for business purposes
   b) handoff of aggregated data for easy feedback into the app (e.g. vote/poll info), either to the programmer, to Canoe or to a vendor (whomever is managing the application broadcast)
   c) aid in validating app delivery and service assurance at an ad zone level (by tracking the app display logging

message at a zip code and ad zone level)

d) the ability to troubleshoot problems by querying the database by STB or application ID

e) The ability to track OOB bandwidth usage by correlating STBs with OOB upstream and downstream nodes

COLLABORATIVE WORK WITH CANOE AND CABLE LABS

Canoe Ventures, founded by Brighthouse Networks, Cablevison, Charter, Comcast, Cox and Time-Warner was created to enhance the MSO's ability to collectively deliver iTV and advanced advertising to a national footprint. Both Canoe and Cable Labs have substantially aided our ability to deploy a scalable ecosystem for iTV.

Canoe has allowed Comcast to process more new applications by offloading some of our business development and application design / development / QA work for national bound applications onto their shoulders, where it is leveraged across many MSOs. The applications initially being developed by Canoe are templated, so they can be re-used across different networks with only minor changes to the app that do nto require new test and trial cycles. This has aided in minimizing onboading times as different networks strive to enhance their programming.

Canoe Ventures has also contributed substantially (along with the vendor community) to creating a more open, reliable and scalable EBIF broadcasting system for integration with a programmer's video broadcast systems. The Canoe architecture for national bound apps minimizes the MSO role in most operational aspects interfacing with the programmer and so streamlines the deployment process. With their MSO partners, Canoe has also collected a wealth of information regarding practical guidelines for application development across MSO user agents, and this will greatly speed development of next generation applications.

Not only has Cable Labs sponsored the forum through which the EBIF specs have been created, they have also recently sponsored the forum that has generated a set of "SaFI" specs [3-6] (Stewardship and Fulfillment Interfaces) for standard messaging between parties in an inter-MSO iTV application ecosystem. Those specs are at the heart of the Canoe inter-MSO application architecture. Cable Labs is also drafting an "ETV Operational Guidelines" specification [7] which aims to standardize some of the iTV platform aspects (such as those discussed in this paper) to make them more scalable and affordable industry-wide.

Some spec enhancements proposed in the most recent (I06) version of the EBIF specs stem from lessons learned from the MSO community during our past two years of deployment experience. One example of an I06 spec enhancements with direct relevance to our ability to operationally support a wide spectrum of EBIF apps is a method to authorize particular applications with permissions to use particular STB resources. As we deploy more apps on more networks, it is important to manage their impact on both STB and network resources, ensuring, for example, that iTV apps do not negatively impact each other or other non-EBIF applications like VOD, SDV and Guide. The "application permissions" proposal to the I06 EBIF spec would enable our platform to manage these resources more intelligently by providing a mechanism for controlled access to STB resources like VOD, DVR, return path, persistent storage, in-band and out-of-band bandwidth and app lifecycle execution privileges.

The test-flag implementation for supporting limiting app display onto a subset of STBs on

plant is contrained due to only having 8 bits in the test-flag and the need for those 8 bits to be shared across all applications and all MSOs. Comcast is also implementing a whitelist/blacklist mechanism as a complementary mechanism to the test-flag, rendering test-flag as a boolean indicator for application test status, similar to how it is treated in OCAP. We expect that other MSOs will implement similar mechanisms that integrate with their OSS applications and processes.

Cable Labs has also recently supported vetting user agents submitted by MSOs against a broad array of test applications (including some they have developed solely for this purpose). We have found that resource to be useful as a research tool as new application functionality and new user agent revisions are developed.

SUMMARY

Comcast has deployed several EBIF-based iTV applications broadly throughout its footprint during the past two years. This paper has reviewed some of the fundamental aspects of our video delivery platform that we have enhanced to enable us to support these and many more future applications in a manner that will ensure an optimal customer experience.

One of our biggest "lessons learned" has been the need to robustly insert proactive monitoring tools in every advantageous network location, including direct application and user agent monitoring and non-realtime operational mining of data already being supplied for purely non-operational purposes.

Another design paradigm that we have learned must pervade our architecture and process is the need to leverage commonality across apps in order to scale them most efficiently. Rather than allow each app to splinter into its own design, we have tried to constrain them to common platform flows with common formatted messaging and common operational support tools and procedures.

Finally, one more key to scalability has been working with other MSO's collaboratively through Cable Labs and Canoe Ventures. The templated apps offered through Canoe Ventures is an ideal example of the above paradigm of leveraging commonality to optimize scalability.

REFERENCES

[1] OpenCable Enhanced TV Binary Interchange Format 1.0. (OC-SP-ETV-BIF1.0-I05-091125)

[2] Enhanced TV Application Messaging Protocol 1.0. (OC-SP-ETV-AM1.0-I04-070921.pdf)

[3] CableLabs Interactive Application Messaging Specification (IAM). AA-IAM-D03.doc.

[4] CableLabs Interactive Application Fulfillment Specification (IAF). AA-IAF-DO3.doc

[5] CableLabs Service Measurement Summary Information (SMSI). AA-SMSI-D03.doc.

[6] CableLabs Campaign Information Package (CIP). AA-CIP-IF-D03.doc.

[7] OpenCable Enhanced TV Operational Guidelines. OC-GL-ETV-OG-V01-060714.pdf.

UBIQUITOUS CONTENT DISCOVERY IN A FEDERATED ONLINE ENVIRONMENT

Michael Kazmier | Chief Technology Officer, Avail-TVN
Christopher Stasi | Senior Vice President Research & Development, Avail-TVN

Abstract

The new paradigm of broadband content delivery brings with it a set of complex challenges. Prominent among them is the need to provide ubiquitous content discovery across a federated online environment.

As both content producers and video service providers rapidly adopt new online services, the challenges of content discovery for the consumer increase exponentially. One constant in this newly connected world is that consumers are neither willing nor able to manage the plethora of locations in which content is made available to them. In order to facilitate the proper discovery of assets throughout an ever changing and evolving environment, an open-standards based platform must be adopted to allow consumers easy access to the content they seek and to which they are entitled. Central to this problem is the hybrid approach a "TV Everywhere" experience brings with it: some content providers wish to both host and deliver content while others wish to distribute and have service providers manage the experience. To facilitate the discovery of content, the consumer must be able to locate assets not only provided through the portal in which they are engaged, but also at the numerous sources located throughout the internet.

*This paper will discuss a standards based approach to allow for ubiquitous content discovery (including search) through*out this fractured, federated environment.

THE NEED FOR A CONTENT CONNECT HUB

Content discovery has been a challenge to Television Provider's since the beginning of the industry. Creative solutions such has the printed TV Guide, electronic programming guide and graphic navigation platforms have been created to assist the consumer in finding the content they desire. Now, consumer desire to watch the television and movie programming they want, when they want it, wherever they are, has never been stronger and continues to grow every day. These first moves to satisfy these consumer desires have emerged as "TV Everywhere" online platforms. As major content producers as well as multi-channel video programming distributors ("MVPD") embrace broadband delivery as a means to satisfy the consumer demand for "TV Everywhere", we find ourselves facing a new set of industry challenges, including content discovery.

There is not a single business model that properly meets the needs of all MVPDs and content producers, thus a hybrid and fractured ecosystem whereby content is delivered from a multitude of locations has been developed. This federated environment works well to meet the needs the numerous business models, but challenges a key value proposition to the consumer: Ease of Use. The consumer's ability to discover content and its location online is the prominent impediment to solve as consumers flock toward online content delivery.

The industry is responding to the consumer demand for access to more content through the development of standards around authentication and authorization. These standards support a federated environment that allows each entity to provide solutions which align with their business desires. With important foundational technologies, these standards leave room for innovative solutions toward other areas such as content discovery. One of the key ways that MVPDs can continue to add value to their subscribers is to facilitate the "TV Everywhere" experience, ensuring that their subscribers can locate and access the content they desire. It is clear that consumers are unwilling to access (browse or search) multiple individual locations in order to locate the content that they desire, thus numerous entities have built proprietary methods for aggregating as well as publishing content availability. While innovative

proprietary solutions embrace the entrepreneurial spirit, utilizing an open-standard infrastructure allows all parties to leverage the scale necessary for success.

The Content Connect Hub is an open-standard based platform providing ubiquitous content discovery for the next-generation of "TV Everywhere." The Content Connect Hub will allow any aggregator, distributor, or provider of "TV Everywhere" services to display and search content from all participating providers throughout the federation. Handling the content discovery in an open, standards based format allows the aggregator to focus on other key aspects of their solution.

Each party in the "TV Everywhere" value chain has slightly differing goals. For the content producer, their key goals are to preserve and drive their content brand as well as control the monetization of their assets. For the MVPD, their key goal is to provide a complete, high quality experience for the consumer along with additional value that they cannot find with other providers. The Content Connect Hub provides critical linkage to the "TV Everywhere" environment that satisfies both of the needs. The Content Connect Hub allows content producers to host the asset themselves, ensuring their ability to secure and monetization it, while still allowing MVPDs to display, search and reference the remotely hosted content utilizing the branding of the content producer. Importantly, the Content Connect Hub does this through an open and cost-free interface.

INSIDE THE CCH GEARBOX

The Content Connect Hub (CCH) utilizes and extends a number of existing standards. At its core, the CCH leverages the SCTE-130 Part 4, Content Information Service (CIS) specification. Just like SCTE-130, CCH leverages standard SOAP XML messaging to allow standardized communication between a client (the consumer portal) and the server (the national CCH service). The CCH provides a basic CIS interface for content queries utilizing the same syntax as a typical CIS, however not all functionality as specified in SCTE-130 is provided through the CCH. For example, the CCH does not allow for the advanced query functionality that CIS does nor does it allow for content notifications (or registration for notification). The CCH is a stateless information service.

The CCH works by maintaining a subset of asset metadata which any hosted provider may publish to. This metadata subset is then used to perform content discovery throughout the federation. When the client (consumer portal) requires detailed asset information, it queries this from the content producer's own CIS environment. This is handled through an extension CIS standard where the CCH will return a Content Metadata Location element as opposed to the standard Content Location element that a CIS would typically return.

To utilize the CCH, a content producer or an MVPD must first register. During the registration process, the content producer is able to setup some basic business rules in addition to generic branding and security information. The content producer is given a Metadata API from which to manage its metadata. The CCH utilizes metadata in the CableLabs ADI v1.1 format. Content Producer's simply publish asset information, including the availability of assets, through the Metadata API. The Metadata API also supports metadata updates through versioning of the ADI files, so content producers are always able to keep their asset information current.

The CCH allows entities to also enforce business rules within the federation. Each entity is allowed to create access control lists (ACL) as to which may permit or restrict their queries. For publishers of metadata, the ACLs may be explicitly exclude certain clients (and therefore everyone else is implicitly allowed access to view/search their content) or explicitly include certain clients (and therefore everyone else is implicitly denied access to view/search their content). Likewise the client of the CCH may limit their searches to a subset of the total providers, either explicitly restricting or selecting the entities to work with.

Once content metadata is published to the CCH, clients (consumer portals) who are also registered with the CCH may then query the CCH for information about remote assets. It's important to note that clients of the CCH must agree to the End User License Agreement which requires the asset branding to be preserved.

Clients can then query the CCH for nearly any metadata criteria. For basic browsing, the client may ask for specific genre or categories. For searches, the title, producer or actor fields are likely targets. Other filters such as the advisories or the rating information may also be specified.

A typical query to the Content Connect Hub looks like this:

```
<ContentQueryRequest
messageId=" consumer_portal.com"
system=" portal_system_1" version
=" 1.1" identity=" 60EA930E-01AF-5050-
A7EB-5D5B4A225311" >
  <ContentQuery expandOutput=" true"
contentQueryId=" 1" >
    <core:ContentDataModel
type=" CLADI_1.1" >URI</core:ContentDat
aModel>
    <QueryFilter>
      <FilterElement name=" Title"
value=" Abyss" />
    </QueryFilter>
  </ContentQuery>
</ContentQueryRequest>
```

And the response from the Content Connect Hub should look similar to the response from an SCTE-130 Content Information Service:

```
<ContentQueryResult resultSetSize=" 2"
contentQueryRef=" 1" >
  <BasicQueryResultList>
    <core:Content>
      <core:AssetRef
providerID=" avail-tvn.com"
assetID=" AABFCA001" />
      <core:ContentMetadataLocation>
        http://cis.avail-tvn.com/cis
      </core:ContentMetadataLocation>
      <core:Ext><ADI> <Metadata> <AMS
Asset_Name=" ..." />
...</Metadata></ADI>
      </core:Ext>
    </core:Content>
    <core:Content>
      <core:AssetRef
providerID=" avail-tvn.com"
assetID=" AA000001" />
      <core:ContentMetadataLocation>
        http://cis.avail-tvn.com/cis
      </core:ContentMetadataLocation>
      <core:Ext><ADI> <Metadata> <AMS
Asset_Name=" ..." />
...</Metadata></ADI>
      </core:Ext>
    </core:Content>
     <core:Ext>

<ProviderBranding>...</ProviderBranding
>
     </core:Ext>
  </BasicQueryResultList>
</ContentQueryResult>
```

While the details are omitted, the response would include the asset metadata in the core:Ext element along with the basic branding information including background color, font, foreground colors for the provider's assets. This information allows the consumer portal display the asset information in the style that the provider intended.

CONCLUSION

The Content Connect Hub is necessary for providing ubiquitous content discovery in a federated environment. By leveraging open-standards, and providing its key services for free, the CCH bridges the gap left open by existing "TV Everywhere" deployments. It is our hope at Avail-TVN that the Content Connect Hub becomes an industry standard for online, multiplatform content discovery and we are eager to work with all parties to ensure its success.

VIDEO QUALITY IMPAIRMENTS 101 FOR MSO'S

Daniel Howard
VQLink, Inc.

Abstract

Key video quality impairments that impact video networks are described to help MSOs in monitoring and analysis of video quality over their networks. An in-depth discussion and taxonomy are given of video quality compression and network artifacts that are detectable by new, no-reference video quality technology that employs a hybrid of both bitstream and pixel processing and thus provides full video analysis of live and file-based MPEG2 and MPEG4 video content. The importance of measurement accuracy and minimal Type I and Type II errors for detection of these artifacts are developed and specific issues in transitioning from MPEG2 to MPEG4 are addressed with respect to these artifacts. Also discussed is how compression and network artifacts are perceived and detected differently in MPEG2 and MPEG4, and the specific video quality challenges for MSOs using transcoded video. Use cases for accurate measurement and classification of video impairments are given for network capacity planning, verification and maintenance of no material degradation (NMD) constraints, and stream bandwidth reductions for delivery of Internet video. Ultimately, MSOs can use newer video measurement and monitoring technology that provides accurate detection and classification of video quality impairments throughout their network to ensure that they affordably deliver the video quality required to remain competitive.

INTRODUCTION

The video quality and quantity wars are underway between cable operators and their telco and satellite competitors, much as the high speed data bandwidth wars began almost a decade ago, and continue to this day. Already the number of HD channels is a key market feature, and now video quality, especially of high definition content, has emerged as a market differentiator and competitive advantage. But just as MSOs discovered the key to effectively competing for high speed data subscribers was to offer the maximum amount of bandwidth they could affordably deliver, the key to winning the video quality wars will be for MSOs to be able to offer the maximum quality that they can affordably deliver. And just as in the bandwidth wars, where MSOs needed accurate tools to measure and adapt bandwidth delivered to subscribers, they will need an accurate tool to measure and monitor video quality in their networks, and this video quality measurement and monitoring tool can then also be used to adapt video quality to ensure network health as well as offer new features to subscribers over time.

Video quality measurement and monitoring technologies fall into several categories.

1) Network proxies for video quality monitoring and measurement: these technologies use network packet monitoring as a proxy for video quality measurement. The difficulties with this approach are as follows:

a) video quality impairments such as compression artifacts are not measured;

b) network artifacts in the video that do not have corresponding packet errors are missed, which happens when packet errors are re-encoded by an MSO or content provider, and packets are renumbered; and

c) packet errors that have little to no impact on actual video quality are reported, thereby 'crying wolf'.

2) Full reference video quality measurement: these technologies rely on access to a copy of the original video in order to compare pre and post processing versions, and use methods such as peak signal to noise ratio (PSNR) to evaluate the video quality. The difficulties with this method are as follows:

a) MSOs cannot measure the video quality of ingested video using full reference technologies since they do not have a copy of the original video from the content providers;

b) PSNR does not work when format changes in video occur because pixels are not in the same location; and

c) PSNR does not accurately reflect the human visual system (HVS) and thus may not correlate well with results from subjective testing.

3) Partial reference video quality measurement technologies: these are similar to full reference in that information derived from the original video is required, and thus they suffer from many of the same disadvantages.

4) No reference, full analysis video quality measurement technologies: these are ideally suited to deployment anywhere in the MSO's network since they do not require a copy of the original video. However, in the past they have suffered from inaccuracy, including the inability to detect even gross video quality impairments, and false alarms from structures in the video that are similar to codec induced blockiness or blurriness.

Fortunately, new no-reference technology based on based on the human visual system (HVS) and employing a hybrid bitstream-pixel processing approach is now available that can accurately detect a variety of video quality impairments anywhere in the network. MSOs can now fully and reliably characterize ingested video using machine algorithms instead of, or in addition to their so called 'golden eye' human video monitors, without requiring a copy of the original clean video. More importantly, they can place the equivalent of their golden eyes anywhere in their networks, and tune the results so that the maximum amount of video quality can be delivered under varying network conditions and competitive threats.

But using such a tool requires complete understandings of the video artifacts that can be detected by a no-reference, hybrid video quality technology, especially when grounded on human visual system modeling and testing, and so details and examples of these artifacts are given in this paper. Emphasis will be on the video image artifacts only; audio artifacts and synchronization issues are not covered.

VIDEO QUALITY IMPAIRMENTS

The two key types of visual digital video impairments that occur in otherwise properly functioning video networks are compression artifacts (CA) and network artifacts (NA). While less common, interlacing artifacts such as 'mice teeth' are also noticeable in modern digital video systems. Also far less common are impairments due to equipment malfunctions such as encoder errors and/or improper configurations, although they do occasionally occur. But to affordably offer video services to subscribers, some amount of CA and NA must be tolerated, and therefore a good understanding of the different types of CA and NA are needed.

It is important to note that even in top quality video, many video quality (VQ) impairments are visible on individual frames. But the human visual system (HVS) misses them if the video is properly encoded and the viewing distance is that of typical viewers,

i.e., 2-3 times the diagonal length of the display. Humans can also miss artifacts if the impairment is away from the center of the screen, or if the temporal duration of the impairment is very short. On the other hand, if a video expert is viewing the digital video from a very close distance, many more artifacts can be seen, including weaker ones that would be missed at a greater distance or missed by untrained viewers, regardless of viewing distance. A properly designed video quality measuring and monitoring system permits a cable operator to vary the results so that the operator can automatically emulate either a video expert, or a typical subscriber, or something in between.

Compression Artifacts

Compression artifacts (CA) occur when the bit rate of the encoded video is insufficient to provide smooth video motion without unnatural jerks, blocks, blur, noise or jagged edges being visible in the video stream. In the worst cases (lowest bit rates), blocky CA can be seen even in static video, but this is very atypical of modern high quality video. The CA types most often visible to viewers (as evidenced by blogs on the subject) can be delineated into video artifacts that are blocky, blurry, and choppy. Unlike network artifacts that can appear somewhat randomly across the screen, most CA are typically seen in conjunction with motion in the video and are usually associated with the object in motion. Circular motion, explosions, scene changes, roaring fires, and fireworks often reveal compression artifacts when they are otherwise not detectable. Trails, which are vertical lines or colored blocks that trail from a moving object in the video, are actually detectable as either blocky or blurry artifacts and are thus included in these two categories. Jagged edges are also a form of blocky compression artifacts. Posterization, which is a loss of color depth, and color errors are also quite noticeable to viewers, but are less common in current video networks when properly configured. Mice teeth, which is an interlacing artifact, is also quite noticeable to viewers when present.

On the other hand, video experts looking up close can see blocky and blurry compression artifacts even when they are not visible to untrained viewers at normal viewing distance, which is to say that they can detect CA of far lower strengths. They can also detect more subtle CA that include the two manifestations of Gibbs effect, namely mosquito noise, which is random speckling around the edges of objects, and ringing, which is spatial ripple away from the edges of text or other sharp edges in the video. Another subtle artifact has been called occasional blur, or "background breathing", which is frequent sudden blur in, for example, background foliage in the video. Although more subtle and often missed by casual viewers, these artifacts are also described below since once artifacts can be detected by machine algorithms, it is a simple matter to convert these detections into appropriate metrics that reflect either expert viewers or more typical home viewers.

1) Blockiness: The most obvious compression artifact for both experts and typical viewers is blockiness, where the bit rate is too low for the level of action or spatial variation in the video. Strong blockiness is obvious to even untrained viewers, while slight blockiness is detectable only up close or by video experts. In Figure 1, there is actually slight blockiness in the first image on the top, but this would normally only be detectable by a video expert looking closely at the video. The middle image shows moderate blockiness that many viewers would miss if it were brief, while the bottom image shows blockiness that all viewers would notice.

In properly configured video networks, a slight amount of blockiness is often acceptable because it either happens infrequently, or is only detectable when video experts look up close at a screen.

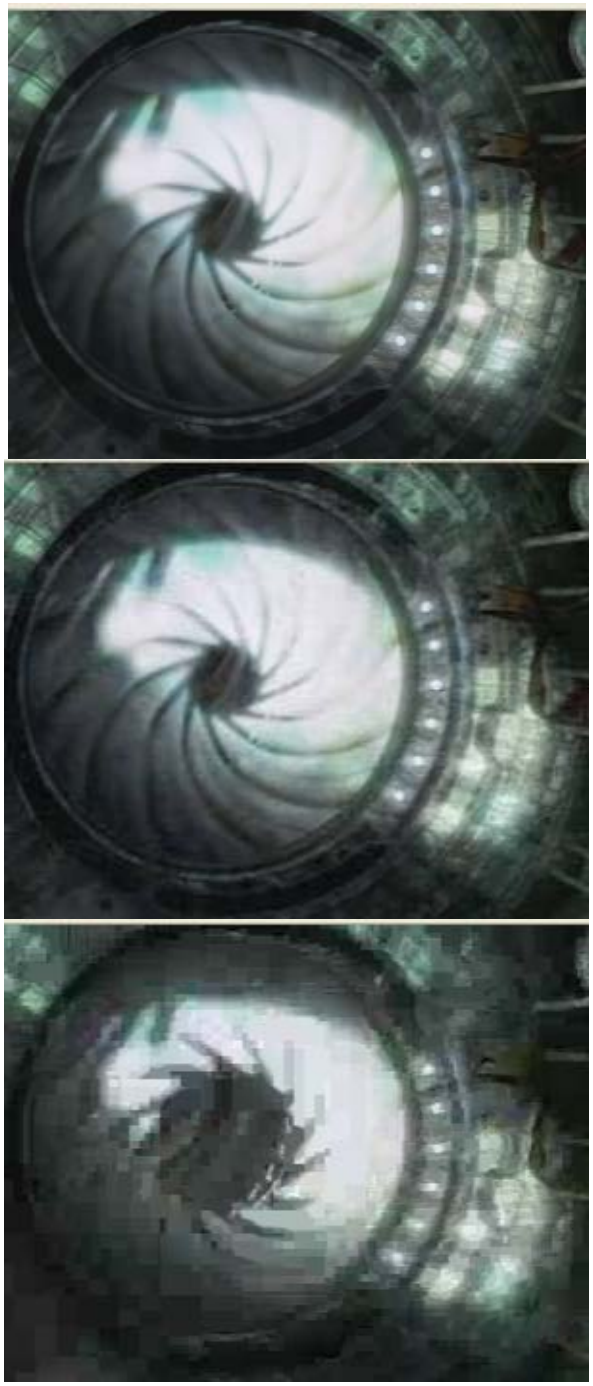


Figure 1. Slight Blockiness (top), moderate blockiness (middle) and strong blockiness (bottom)

2) Blurriness: Codec induced blurriness similarly occurs when the bit rate is too low, however the effect is often more subtle to viewers and thus a greater amount can be tolerated by untrained or less picky viewers. Especially in fast moving scenes, even moderate blurriness can be missed entirely by many viewers. Video experts however, can not only detect even slight blurriness, but can readily tell the difference between naturally blurry features in the video (objects that are intentionally out of focus, e.g.) and codec induced blurriness. Therefore, video quality systems should also be capable of detecting even slight codec-induced blurriness if they are to mimic experts.

Figure 2 shows natural blurriness in the image as seen near the portal opening at the top left in the original image on the top, while the image on the bottom shows codec-induced blurriness as seen around the four lights on the rightmost image in Figure 2.

Figure 2. Natural vs. MPEG4 codec-induced blurriness in original (top) and low bit rate (bottom) video.

While previously it was difficult for machine algorithms to detect such subtle differences between codec-induced blurriness and natural

blurriness, new video quality measuring technology is quite capable of discerning codec-induced blurriness from natural blurriness and alerting operators to the presence of blurry compression artifacts without undue false alarms from natural blurriness in the video.

Note that in the bottom image in Figure 2 there is both blockiness and blurriness produced by the lower bit rate. It is often the case that several compression artifacts are simultaneously present in a video when the bit rate is too low, and a properly designed video quality measuring system should indicate specifically which artifacts are present. This is so that an operator can, for example, control the bit rate carefully to minimize blockiness that can be seen by most viewers but permit a higher level of blurriness since far fewer viewers will notice the latter.

3) Choppiness: A lack of smoothness in motion in a video (sometimes called jerky video) produces choppiness in the video that can also be detected by newer video quality measuring technologies. While difficult to depict in still images here, it can be imagined as slow motion video effects where they should not be. New video quality measuring technology can also detect this type of choppy compression artifact.

4) More subtle compression artifacts: Unless the bit rate is grossly under the required level for quality viewing, more subtle compression artifacts are often missed by most viewers, but are nonetheless detectable by video experts. Examples include mosquito noise and ringing (which are both manifestations of Gibbs effect), and sudden blurriness or 'background breathing' which is often noticeable in MPEG4 encoded video at lower bit rates. Mosquito noise and ringing are shown in Figure 3. On the top, mosquito noise is seen as random speckle around the text, while on the bottom there is a periodicity to the smearing of the edge of the image, especially at the top of the letter "O" for example.

Figure 3. Mosquito noise (top) and ringing artifact (bottom) around text.

Another compression artifact which can be subtle is the so-called 'background breathing' a somewhat frequent blurring of the background, which can be seen in MPEG4 video in particular when there is foliage in the video background. Figure 4 shows a closeup of the fixed background in two consecutive frames of MPEG4 video. The entire region appears to 'jump' slightly in the video due to a sudden increase in blurriness, which can be seen in the highlighted region in particular, but in actuality the entire background is perceptibly different.

Figure 4. Background breathing example in MPEG4 video.

5) Interlacing artifacts: Another video artifact in interlaced formats (480i, 1080i) that is not so subtle and also detectable by modern video quality monitoring technology is the so-called 'mice teeth', shown in Figure 5 below, and when present, it is also highly visible to even untrained viewers from a distance.

Figure 5. Interlacing artifacts (mice teeth).

Network Artifacts

Network artifacts in video may occur when packet errors occur such as loss, excessive delay or jitter, or even that packets are sent out of order. The video may or may not be affected depending on the type of MPEG frame involved, error masking techniques in use, and so on. Thus, not all network errors lead to actual video artifacts. On the other hand, if network artifacts in the video are re-encoded such that packets are renumbered, there may be network errors in the video that do not have corresponding packet error indications, which means the network artifacts can only be detected via pixel analysis of the video.

The worst network errors result in stuck frames, lost frames, and blank frames, and these are all detectable by most modern video quality measurement systems. Amazingly, entire frames can be lost and still the viewer not notice it if the video is relatively static. This is also true of the most typical network errors, where portions of the frame are affected, especially those in motion. Error concealment can essentially repeat the affected portion of the frame from a previous frame and thus the network error looks like a slight jerk in one portion of the video. With error concealment turned off however, the network errors appear as either streaks, blocks around the edge of a moving object, or a checkerboard type pattern. Figure 6 shows several network artifacts (NA) visible in MPEG2 and MPEG4 video.

Figure 6. Checkerboard and streaky (top) MPEG2 network artifacts, and H.264 network artifacts (bottom).

Note that these figures depict the errors as they actually appear, i.e., when error concealment in the decoder is disabled. In the checkerboard artifact, the entire screen is affected, however even this can be concealed or missed if it occurs in a single frame only. Network streaks, as seen in the middle image of Figure 6 are far easier to mask when they are localized, and if they occur at the edge of the screen they may not even be perceived by typical viewers.

In a sense, MPEG2 network artifacts are more straightforward to detect, even in the absence of packet errors in the video stream, because of the regularity of macroblock features seen in the video. Network artifacts in H.264 (MPEG4 part 10 or AVC) video, on the other hand, are much more challenging to detect via pixel analysis because the artifact pattern is much more varied and non-regular, as seen in the bottom image in Figure 6.

Nevertheless, network artifacts in H.264 can be detected by advanced video analysis algorithms, and alternately when packet errors occur, the video can be analyzed to determine what impact, if any, the error had on the H.264 video.

Other Video Artifacts

The other class of video artifacts that occur occasionally are due to errors or failures in the encoders themselves, which can take the form of super macroblocks in the video, often seen at the edge. However, many of these patterns are specific to the encoder and thus are not shown here.

MEASURING IMPAIRMENTS WITH NO-REFERENCE TECHNOLOGIES

As mentioned in the introduction, the key benefit of using no-reference, hybrid bitstream-pixel video quality measurement technology grounded in subjective testing is that the video may be analyzed anywhere in the network, from initial video ingestion point to the home, and it provides an accurate match to what humans would have perceived. And unlike approaches that rely solely on packet errors, and therefore require such devices throughout the network for locating the source of video problems, no-reference hybrid techniques can detect and classify compression and network artifacts even after subsequent packet processing, which means monitoring devices can potentially be more sparsely deployed in a network and still accurately detect the vast majority of artifacts.

However, since no-reference techniques do not have the original, assumedly clean video for comparison, there will a non zero error rate in both the probability of false alarm (Type I errors) and the probability of missed detections (Type II errors). It will be up to the individual MSO to decide where best to draw the line between minimizing Type I errors vs. minimizing Type II errors, depending on whether maximum visibility of artifacts is desired, or a mimic of untrained viewers who are likely to miss many subtle errors. The required missed detection and false alarm probabilities are also likely to be a function of the value of the video asset, level of artifacts already present in the video, the time of day, and tradeoffs between certain video services and other services such as high speed data for business customers. MSOs should therefore tune the quality monitoring system to their specific headend needs or the needs of their subscribers, which may vary considerably.

The results of video quality measurements can be mapped to a variety of metrics for MSOs, such as number of artifacted seconds, or Mean Opinion Score (MOS). Many more granular and higher level metrics are also possible and have been implemented in new technology from VQLink, and these metrics can be used to characterize video streams absolutely, relative to each other, and also over time for trend analysis. Ideally, the measurement system should not only detect artifacts, but also classify them. But for any of these metrics to have meaning, the video quality measuring system must have adequate Type I and II error performance for detection and classification of video artifacts, and this should be grounded on, and verified with human subjective testing to ensure accuracy.

CHALLENGES WITH MEASURING VIDEO QUALITY OF TRANSCODED CONTENT

One of the challenges for no-reference video quality measurement is when the video has been transcoded. In the simplest example, low bit rate video could be reencoded at a higher bit rate prior to ingestion by the MSO, thereby preventing the MSO from knowing the original bit rate. An accurate video quality measurement system would not be fooled by the higher bit rate, but rather would detect the video artifacts correctly regardless of the bit rate of the video. It is much more challenging when not only the bit rate, but also the type of encoder used is changed by the transcoding. What can then happen is that the compression artifacts typical of one type of encoder at a low bit rate are present in a video stream which uses a different type of encoder and bit rate. While this would previously mean higher Type I/II error rates, new video quality measurement algorithms are much more intelligent and can perform adequately even for this type of content, although the error rates will likely vary depending on the specifics of the original encoder, the final encoder, and any bit rate changes that occur.

USING ACCURATE VIDEO QUALITY MEASUREMENTS TO IMPROVE NETWORK EFFICIENCY

Once an accurate video quality (VQ) measurement and monitoring system is in place, there are several use cases for this technology of interest to MSOs, content providers and content aggregators. First, the MSO can use VQ measurements to characterize the level of artifacts in ingested video. This gives the MSO the ability to groom the channel as needed and compare the results to the ingested stream to ensure that no material degradation has occurred, both in the headend and at the edge of the network. Since the quality of ingested video varies considerably, characterizing the video at this point with maximum accuracy is particularly important.

Second, with accurate VQ measurements, content providers and MSOs that offer video streams over the Internet can trim the bit rates so that quality needs specific to Internet delivery are met, which may be different from those on their cable networks, for example. For content providers, this represents potential savings of bandwidth costs for delivery over the Internet. As before, if the video delivered is already artifacted to some extent, there is generally room for additional bandwidth savings as long as the video quality can be accurately measured and maintained.

Third, MSOs seeking to add more HD channels, or increase the number of QAMs allocated to high speed data service (in the daytime for business customers, for example) may need to alter their QAM lineups using accurate video quality measurement of channels and their trends over time. While the 3DTV frame compatible system currently proposed for cable is designed to avoid

additional bandwidth requirements for 3D content, there are proposals for enhancements to frame compatible 3DTV that would add additional bandwidth requirements. And if the market demands full 3DTV over time, the bandwidth needs of 3D channels will certainly increase.

CONCLUSIONS

New, no-reference video quality measurement technology that employs a hybrid of bitstream and pixel processing to perform full analysis of video is available that offers MSOs the ability to accurately detect, classify, and monitor video artifacts anywhere in their networks, from initial ingestion to edge delivery. However, to use such technology maximally, the MSO must understand the types of video artifacts common in cable networks, their impact on video quality as a function of artifact type and strength, and use this information to configure video quality systems so that the data reported is both accurate and actionable. Too many false alarms mean that such systems can become ignored by technicians, while inability to detect artifacts under all conditions, including transcoded video, leads to a false sense of security and lack of trust in the monitoring system.

With the emerging video quality wars between competing video service providers, the provider who can *affordably* deliver superior video quality to subscribers, and monitor and enhance that quality over time will have the edge. Further, as more home viewers get ever larger TVs, and blogs, forums and other social networks as well as marketing tactics focus the viewer's attention on video quality, the differences between the professional video expert and the home viewer is likely to decrease over time. Hence, understanding video artifacts fully and having an ability to accurately measure and monitor them will be critical to both the current and the future success of MSOs in their delivery of video services.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. 0848558. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

WHAT 3D IS AND WHY IT MATTERS

Mark Schubin
SchubinCafe.com

Abstract

The term "3D" is venerable but ambiguous. It can be confusing to viewers, referring to a broad range of technologies, many of them unrelated to depth perception.

Unlike motion, sound, color, and HD, current 3D TV systems do not generally provide more information to viewers, just a sensation of depth, varying with screen size, viewing distance, and pupillary distance of the viewer. Under some conditions and for some viewers 3D can cause visual discomfort.

An understanding of the different meanings and technologies of 3D can help reduce viewer problems and confusion. Future 3D television technologies may be very different.

SEMANTIC AMBIGUITY OF "3D"

3D without Depth Perception

Enter the term "3D circuit" into a search engine, and the first results will likely be related to a Third Circuit court of law. The first non-judicial result might well refer to techniques for stacking transistors or other electronic components.

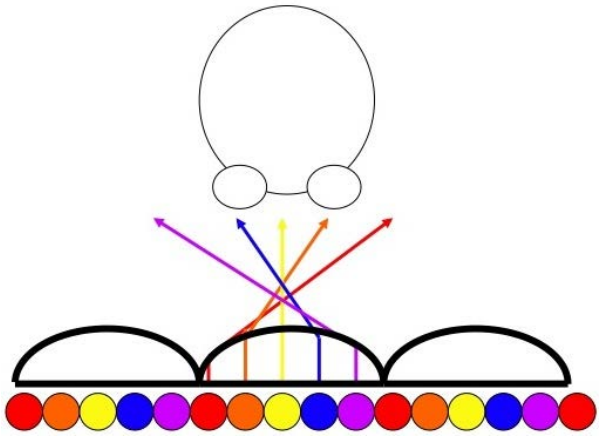
Restricting results to the television field doesn't necessarily help. A major television-set manufacturer still touts its use of 3D comb filtering in its latest digital, flat-panel HDTVs.[1] In that case, the three dimensions are vertical, horizontal, and time.

Three Academy Awards (Scientific & Technical) were presented this year for 3D achievements, but they, too, were unrelated to depth perception. All were awarded for image processing using 3D look-up tables, wherein the three dimensions were axes of color.[2]

Depth of "3D" in Television

Even when the term 3D applies to both depth perception and television, there is a large range of possible meanings. Earlier this year, Study Group 6 of ITU-R (the International Telecommunications Union's Radiocommunication Sector) issued a report "outlining a roadmap for future 3D TV implementation."

Its third-generation future signal format is called "object wave profile," perhaps better known as electronic holography. The second generation is multiview autostereoscopy. Both the second and third generations allow viewers to get different views by shifting their heads (as in "real," non-television vision).



example of multiview autostereoscopy, in this case five-view lenticular (lens-based)[3]

The first-generation signal format in the report is called "plano-stereoscopic." Even that format is assigned four levels based on degree of compatibility with existing displays, existing video frames, and existing signal-distribution standards.[4]

Not even the ITU-R report covers the full range of 3D viewing options. At the high end, for example, there is already full-color, full-motion, high-definition holography, though in its current commercial implementation neither

capable of being transmitted live nor of television-program duration.[5]

At the low end, there are 3D-sensation-providing techniques that might be considered less than even plano-stereoscopic, such as the Pulfrich illusion, chromostereopsis, temporal view shifting, and microstereopsis, all of which have been applied to television programming. They will be described later.

Below even those techniques are ordinary television pictures, which, nevertheless, provide multiple depth cues. The terms "3D graphics" and "3D animation," for example, have been applied to digitally generated objects and scenes using such techniques as shading and perspective to create the appearance of depth even in 2D images.

Human Depth Perception Cues

The strongest indicator of depth in human vision at all viewing distances is occlusion or interposition. It is simple to understand. If one object is blocking another, the one being blocked is behind the one doing the blocking. There are many other depth cues, as illustrated in the painting below.

Joachim Patinir, *Charon Crossing the Styx*

Even though it is an early-16th-century painting, reproduced here at a tiny fraction of its size, viewers should have little difficulty distinguishing foreground from background using such cues as occlusion, object size, textural perspective (detail in the vegetation, rocks, and water becoming less distinct towards the background), and aerial perspective (objects in the distance becoming both hazier and bluer). Moving-image media, such as television, offer even more depth cues, including motion parallax (the ability of a camera to "see" around objects as it moves) and temporal size change (e.g., a car getting bigger as it approaches the camera).

The "real" world adds three more cues. One, accommodation, is the focusing of the eyes' lenses on something. The focus muscles send distance feedback to the brain. Another, vergence (or convergence), is the aiming of the eyes. Again, muscular feedback provides a depth cue.

The third cue is called stereopsis or binocular disparity. In viewers with binocular (two-eyed) vision, each eye gets a slightly different view, just as one camera gets two different views at two different moments when it is moving. Stereopsis, however, does not require camera or object motion.

random-dot stereogram

In the absence of any other cue, stereopsis can provide depth information. If the left image above is viewed by the left eye and the other by the right, a square in the center should appear to float above the field of dots.

VIEWING 3D TELEVISION

Revelation vs. Sensation

3D is sometimes characterized as the next step in image media, following such developments as motion, sound, color, and

high-definition. From a business standpoint, it very well might be.

At this year's Consumer Electronics Show, Panasonic CTO Eisuke Tsuyuzaki said, "It's a challenging market [for TV-set sales]. We need something to kick us out of this. To me the thing that's going to get us there is 3D."[6]

In terms of viewing experience, however, 3D is different from sound, color, and high definition. Consider this black-&-white still image from 1954.

Without motion, viewers don't know what just happened or is likely to happen. Without sound, they don't know what she's saying (or singing). Without color, they don't really know whether the hair is blonde or blue-tinted white. Without more definition, they can't say whether there is an earring or a trickle of blood. It's clear, however, based on relative sizes, that the face at the lower right is farther away than the main subject.

Unlike motion, sound, color, and high definition, 3D (except for holography, multiview, and similar technologies) does not reveal new information to viewers. It *does* provide a sensation, however, as do such other important moving-image developments as music and directing.

Range of 3D TV Sensations

The report of ITU-R Study Group 6 indicates that "object wave profile" 3D TV, the only form that might approach the "real" world visual sensation of depth, is "technically some 15-20 years away."[4]

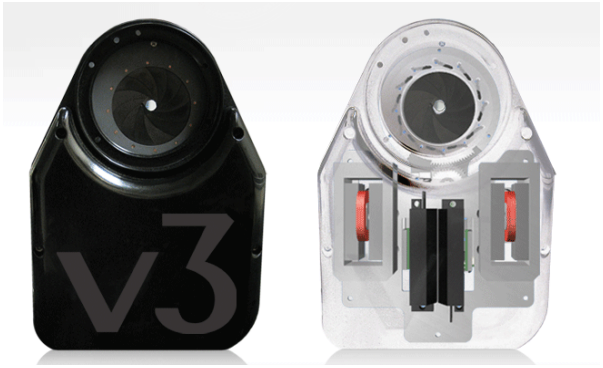
At the 2009 convention of the National Association of Broadcasters (NAB), Japan's National Institute of Information and Communications Technology demonstrated live 3D holography. The image was tiny and crude, had a limited viewing angle, and required a room full of laboratory equipment to produce. Here is a still photo of the image.

Across the aisle, NHK (the Japan Broadcasting Corporation) demonstrated a form of multiview 3D they called "integral" television. It used a multi-lens array in both shooting and display to provide an image that *did* reveal more information as the viewer's head moved.

Unfortunately, the multiple views divide the system resolution. Despite the use of an ultra-high-definition camera (one with 16 times the number of picture elements of so-called "full 1080-line" HDTV), the final image appeared to have less resolution than even a low-grade home VHS videocassette recording. Here is a full-screen image.[7]

At the other end of viewer 3D sensation, below even the ITU-R "first-generation" plano-stereoscopic systems, are view shifting, the Pulfrich illusion, and chromostereopsis. All three can be used with unmodified television distribution systems and displays.

View shifting is a restricted version of motion parallax. If the shift is sufficiently large, and the image is otherwise stationary, it provides a strong sensation of depth. In entertainment-grade moving images, the effect is subdued.[8]



v3 view-shifting lens adaptor

In Pulfrich-illusion 3D, the viewer darkens one eye. Foreground portions of the scene must move in one horizontal direction and background in the opposite. A carousel is ideal Pulfrich material.

According to one theory of how it works, the darkened eye is forced from photopic (retinal cone-based) vision towards scotopic (rod-based). The photochemical reaction in cones is faster than that in rods. The clear eye, therefore, sees what *is*, and the darkened eye sees what *was*, effectively providing a form of motion parallax.

Pulfrich 3D has been used often in television. These glasses are from Discovery *Shark Week* in 3D.[9]

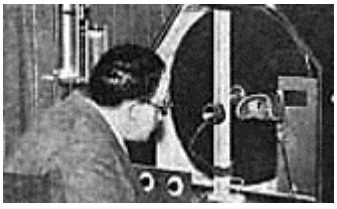
Chromostereopsis is based on the inability of any simple lens to focus different colors at the same point at the same time. Red focuses closer than blue, so a scene in which the foreground is reddish and the background bluish will provide a mild depth sensation.

Just as Pulfrich 3D requires attention to foreground-background choreography, chromostereoscopic 3D requires attention to color placement within a scene. The effect can be enhanced with glasses that shift colors in different directions. Here are red-shifting ChromaDepth glasses used for VH1's *I Love the 80s* in 3D.[10]

View shifting, the Pulfrich illusion, and chromostereopsis not only utilize ordinary TV sets and distribution systems, but also, even when they use glasses, provide images compatible with viewers not wearing glasses. Other 3D systems (not counting multiview, holography, and so-called volumetric displays) use stereoscopy.

Stereoscopy attempts to duplicate binocular human vision. Separate left-eye and right-eye views are captured (microstereopsis, which will be described later, uses variations of that process).

For presentation to the viewer, some mechanism must be used to direct the appropriate view to the appropriate eye. These range from goggles with built-in screens for each eye, to a broad range of view-controlling glasses, to glasses-free autostereoscopic displays based on visual barriers or lenses. The first 3D TV broadcast, in 1928, used a stereoscope for viewing, as shown below. That forced the viewer into a fixed position, head against the hood.

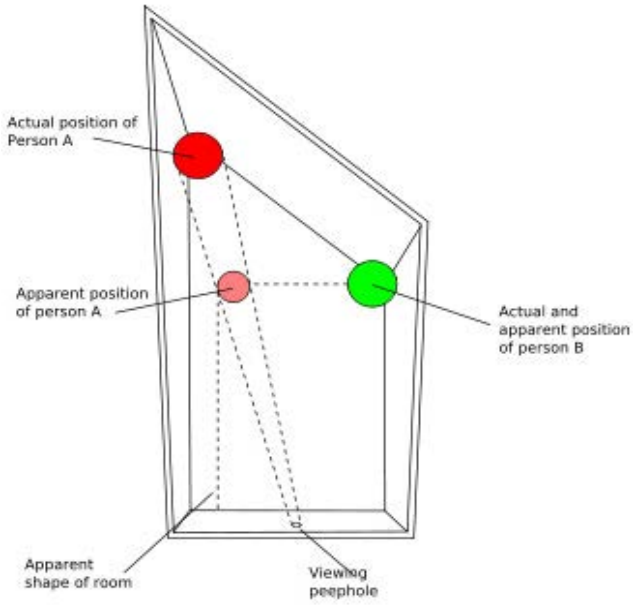


Cue Conflict

In the "real" world, all visual depth cues should work together to provide the same information. In the world of produced imagery, however, they need not.

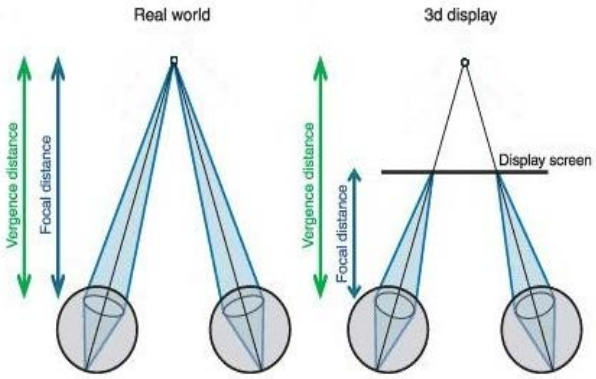
Sometimes cue conflict is used intentionally to provide a desired effect. The Ames-room illusion, for example, uses a distorted version of a normal room so that the parallel-line perspective cue fails.

In an Ames room, surfaces that appear to be parallel, like the walls, floor, and ceiling, are not. The illusion was used intentionally in the *Lord of the Rings* movies to make certain characters appear smaller than others. An Ames-room plan is shown below.[11]



Sometimes, however, there can be undesirable cue conflicts. In "real" world vision, for example, accommodation and vergence should always provide the same muscular depth feedback. In stereoscopic 3D, eyes always focus on the screen, but vergence may be at, behind, or in front of the screen,

The illustration at the top of the next column indicates the conflict. At the left is the "real" world; at the right is a 3D screen. In this case, the stereoscopic convergence point is located behind the screen.[12]



The vergence-accommodation cue conflict has been proven to cause discomfort under certain viewing conditions. At theatrical viewing distances, it is generally not a problem unless the stereoscopic convergence point is very far in front of the screen. At close TV-viewing distances, however, a convergence point far behind the screen might cause discomfort.[12]

A different cue conflict is shown below, this time between two of the most powerful cues, occlusion and close-range stereopsis. In the upper pair of images, the text is difficult to read when viewed stereoscopically, because its positioned 3D depth is behind the snowball it is occluding. In the lower pair, the text depth has been moved to the front, and it is easier to read.[13]

THE OTHER THREE DIMENSIONS OF 3D

Pupillary Distance

If the goal of shooting stereoscopic 3D is to capture the disparate views of binocular human vision, it is important for the stereographer to have a sense of how disparate those views normally are. It is not necessary to duplicate them precisely. As in other image media, it might be desirable to exaggerate or diminish an effect.

Although there are many human-vision depth cues, they don't all have the same influence at different distances. Stereopsis, or binocular parallax, is an extremely strong cue, second only to occlusion at very close distances, where the space between the pupils of the eyes is significant. As distance increases, stereopsis diminishes, disappearing completely at a few hundred meters.[14, 15]

When the left- and right-eye images are captured with lenses spaced beyond the appropriate distance, the result is "hyperstereo;" stereopsis is extended to greater distances, but the viewer experiences a sensation of diminished depth, as though looking at a doll house or through the eyes of a giant. "Hypostereo" is the opposite.

A stereographer can choose an appropriate view-separation distance based on lens magnification and human pupillary distance, PD, the distance between the centers of the pupils of our eyes. Unfortunately, there is no single PD. According to one researcher, the "range of 45-80 mm is *likely* to include (almost) all adults, and the minimum… for children (down to five years old) is around 40 mm" [emphasis added].[16] Variation in human pupillary distance is the reason binoculars have adjustable hinges.

Thus, stereography based on a 40 mm PD might seem like hypostereo to an adult viewer with a larger PD. Conversely, something shot based on an adult PD might seem like hyperstereo to a young child viewing the same screen.

Screen Size

When looking at an infinite distance, both eyes of a viewer with normal vision will point straight ahead. The vergence angle will be zero. In human vision, furthermore, "infinity" can be considered to be as close as ">20 ft."[17]

A stereographer can measure a viewer's PD, shoot accordingly, and arrange to have the content presented to a viewer with objects at infinity separated by the PD on any screen that is wide enough. Unfortunately, there is a great range of screen sizes.

The largest cinema-auditorium screens are more than 100 feet wide. The smallest screens (exclusive of built-in-screen goggles) on which some people watch entertainment programming are probably on mobile phones, just about an inch wide. In between are television screens ranging from a few to 152 inches in diagonal.

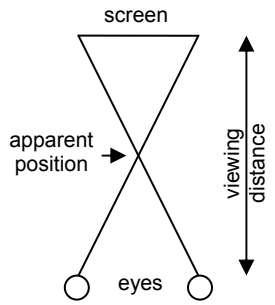
Display adjustment of stereo disparity based on screen size is not yet the case in television sets. A conflict between PD-based stereopsis and screen-size-based convergence, therefore, is likely.

Infinity eye-view disparity set to 40 mm on a 100-foot screen will become 0.93 mm on a 32-inch-diagonal 16:9 TV screen. For a viewer with the author's 68-mm PD, viewing the screen from the nominal-viewing Lechner Distance of nine feet, the vergence-based feedback depth for infinity will be just over nine feet, barely behind the screen.

Conversely, an infinity separation set for a small screen can scale up on larger screens beyond the viewer's PD. That calls for the eyes to *di*verge rather than *con*verge, an unnatural condition.

Viewing Distance

In the example above, it was necessary to specify PD, screen size, and viewing distance, because vergence angles are based on all three. It is easiest to illustrate the effect with negative parallax, the condition in which a point in the right-eye's image is to the left of the left eye's image.



As shown above, if the negative parallax equals the viewer's PD, then the vergence point will be exactly halfway between the screen and the viewer. Of course, that distance will vary with the viewer's distance to the screen.

A viewer watching at a distance of four feet will get vergence feedback indicating that the point is two feet in front of the screen. A viewer with the same PD watching the same screen at a distance of 12 feet will get vergence feedback indicating that the same point is six feet in front of the screen.

VIEW-CONTROL MECHANISMS

Fixed Position

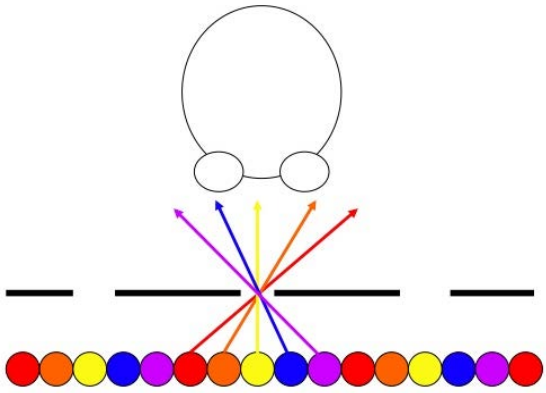
Again, the first 3D TV broadcast, in 1928, used a stereoscope to control which view got to which eye. The viewer's head was placed against the stereoscope hood. Prismatic lenses directed each eye to its appropriate view and increased the accommodation distance to reduce the possibility of vergence-accommodation conflict, and adjustments could be made for different viewer PDs.

The modern equivalent is video goggles, wherein each eye gets its own screen, with an adjustable lens system to provide a significant accommodation distance. An example, the Vuzix Wrap 920, is shown below.

It is intended to provide the sensation of a 67-inch screen viewed from a ten-foot distance and has a suggested price of about $350.[18] An upcoming model is to include stereoscopic cameras for "augmented reality."

Autostereoscopic Systems

Autostereoscopic displays use lens-based (lenticular, illustrated previously) or barrier technology (shown below) to direct the eye views. If more than two views are used, there is leeway in viewer position, and shifting the head can reveal new information, as in the "real" world.[3]



parallax-barrier multiview autostereoscopic display

Unfortunately, the overall display resolution must be divided by the number of views presented. That's why the image on the NHK "integral" television display appears so crude despite utilizing ultra-high-definition equipment.

Having just two views reduces the resolution problem. Unfortunately, it also makes the viewing "sweet spot" narrower.

As can be seen from the diagram on the previous page, successful view control in autostereoscopic displays is also affected by viewing distance and deviation from the orthogonal axis. Relatively large autostereoscopic displays viewed at relatively short distances can make tracking of objects across the screen difficult.

Glasses

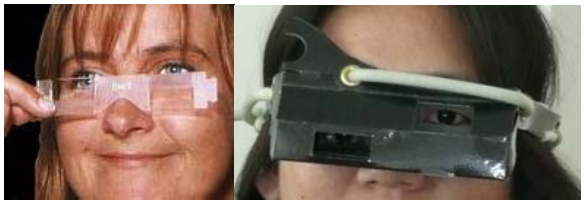
Not counting the special glasses for Pulfrich (one lens darkened) and chromostereoscopy (color spread), glasses for 3D viewing may be divided into four categories, each with sub-categories. They are: view-directive, polarized, color-filtered, and shuttering.

When *Business Week* magazine ran the headline "3-D Invades TV" in 1953, the glasses being used were view directive, with prisms to direct the eyes to side-by-side views on a TV screen. They are shown below on the left. Next to them are modern prismatic glasses for side-by-side viewing.[19]

In side-by-side displays, viewers without glasses can choose to watch just one of the images, perhaps covering the other to avoid distraction. When the right-eye view is to the left of the left-eye view, viewers without glasses may also occasionally cross their eyes to fuse the two into a 3D image.

One problem with side-by-side displays is that they have a vertically oriented aspect ratio. The trend in television has been towards ever wider aspect ratios, with two manufacturers offering 21:9 screens.[13]

An alternative, therefore, has been over-under displays of the two images. On a 4:3-aspect-ratio TV screen, each image would have the shape of the widest-screen movies. Again, prismatic glasses have been used, as shown at left below.[20] Cross-eyed viewing without glasses, however, is impossible.



In both side-by-side and over-under displays viewed through prismatic glasses, there is a restricted "sweet-spot" viewing distance. Inexpensive, adjustable-distance mirror-based prototype glasses have been demonstrated (above right), but they have not yet been manufactured.

Polarized glasses are used in systems with matching display polarizers. As an example, if the left eye's view is polarized horizontally and the right eye's view is polarized vertically, then, if a viewer wears matching glasses, only horizontally polarized light will reach the left eye, and only vertically polarized light will reach the right.

When polarized images are projected, it is necessary for the screen not to depolarize the light. Non-depolarizing screens are sometimes called "silver" screens. In practice, linearly polarized glasses are usually polarized along 45- and 135-degree axes.

With the left-eye and right-eye images superimposed on the screen, any light of the wrong polarization that reaches a viewer's eye (optical crosstalk) might be perceived as a ghost. Many factors affect ghosting, including the polarization materials used, their alignment, and scene content (a white object in one eye's view spatially coincident with a black field in the other eye's view is an extreme example). Ghost-reduction systems can be used.[21]

In addition to linear polarizers, there are also circular polarizers. The alignment of circular polarizers is not critical in the manufacture of glasses. On the display side, it's also possible to use a switchable optical delay to reverse the direction of circular polarization on a temporal basis (e.g., alternate video field or frame).

Ghosting can also be wavelength dependent (many circular-polarizing filters do not cancel shorter wavelengths like blues as well as do linear polarizers) and affected by the angle of the viewer's head and the glasses on it. Appropriate 3D viewing requires upright viewers, not heads on shoulders.

Linear polarizers provide a self-correction function to viewers; as they see double images, they align their heads to reduce them. Circularly polarized glasses do not provide that indication of head misalignment.

Glasses with colored filters, such as those shown above, have been used for 3D for so long that they are called "anaglyph" (literally three-dimensional carving). Even inexpensive color filters have very distinct pass bands (that allow certain colors through) and stop bands (that prevent colors from getting through).[22]

When the filters were used for black-&-white movies, therefore, with matching filters on the projection side, there was little ghosting. Unfortunately, color television displays do not match the colored filters as well, resulting in increased ghosting.

Anaglyph technology may be used with existing TV sets and distribution systems. Besides ghosting, however, there have been other issues associated with anaglyph glasses, including poor color rendition and brightness mismatch between the eyes. There is also the chromostereoscopic issue of red accommodation being different from blue.

For those reasons (and others), there are many different anaglyph color combinations. A green-magenta pair, such as the one shown below, reduces the brightness mismatch.[23]

Magnification is sometimes used to adjust red accommodation. A dark amber-blue pair, such as the one shown below, is said to offer better color rendition.[24]

Colored and polarized filters cause reduced transmission of light to the eyes due to both the glasses and the filters used on the screen or in projection. In cinemas, where light levels are lower than those of TV screens even in unfiltered projection, the transmission reduction is sufficient to cause a noticeable color-desaturation effect. It has been mentioned in reviews that consider both 2D and 3D versions of the same 3D movies.[25]

One form of color-filter glasses is *not* referred to as anaglyph. It uses interference-filter technology to create something that might be considered an optical comb filter. Each eye gets red, green, and blue primary colors, but the color primaries for one eye do not match those for the other. Thus far, the system, glasses shown below, has been used only with projection displays.[26]

All of the previous 3D glasses – view-directive, polarized, and color-filtered – are electronically passive, and the anaglyph and polarized versions can be made very

inexpensively. The last type, shuttering (eclipse) glasses, are electronically active.

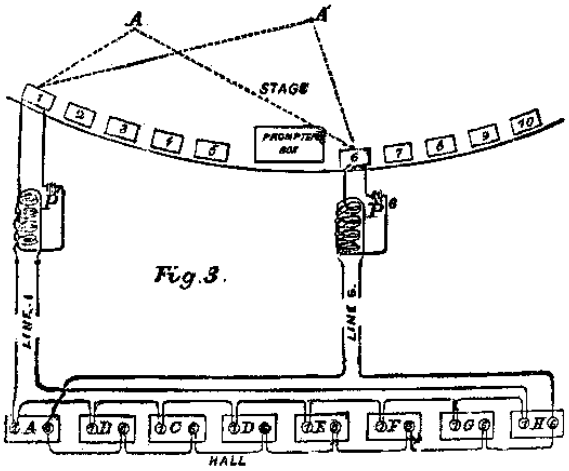
They alternately block the view of one eye and then the other in synchronization with the presentation of the alternating views on the display screen. They may be wired to the display or use a wireless (infra-red or radio-frequency) sync system. Wireless active glasses must use batteries for the shutters.

In consumer TV sets, shuttering depends not only on the flashing speed of the glasses but also on the refresh rate and optical characteristics of the display. A review of this year's first 3D TV sets shows differences in ghosting based on display technology.[27]

MICROSTEREOPSIS

Evolution of Stereo Sound

In 1881, at the International Electricity Congress, a demonstration of what we would today call stereo sound was conducted. Pairs of microphones at the lip of the Paris Opera stage, as shown below, fed pairs of telephone receivers in listening rooms. The process was then called "binauricular auduition," but it was likened to the 3D stereoscope.[28]



Crowds (among them author Victor Hugo) were impressed by the ability of stereo sound to provide localization of sound. There was a "Wow" factor.

Early stereo sound in the electronic era also used widely separated microphones and speakers. The result, sometimes called "ping-pong stereo" (when sounds were heard first coming from one speaker and then from the other), also provided an exciting, if unnatural, source-localization sensation.

Now that stereo sound is common, it is often acquired through "single-source" techniques (more closely spaced microphones). They deliver less of a "Wow" factor but what is often considered a more natural sound. A tiny stereo microphone is shown below, larger than actual size.[29]

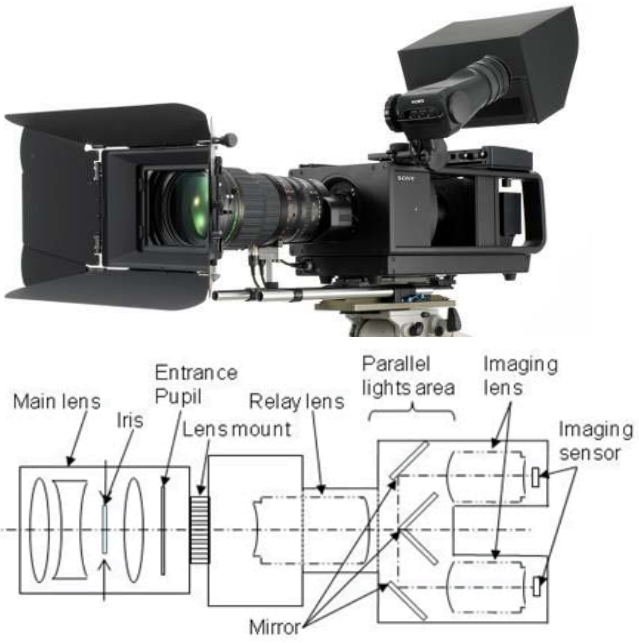
"Kinder Gentler Stereo"[30]

As noted previously, conflict between 3D vergence and accommodation can cause viewer discomfort, and a mismatch between separation during image acquisition and viewer vergence on the same material can change the perceived depth of the 3D sensation. Is there an alternative?

One proposal has been called "microstereopsis."[30] It is based on the principle that binocular human vision is extremely sensitive to stereoscopic disparity – under some conditions down to a fraction of one arcsecond (an arcsecond is 1/3600 of a degree of arc).[31]

A sensation of depth, therefore, can be conveyed with very little separation between left- and right-eye views. When displayed,

the image is, furthermore, generally accepted by viewers without glasses as normal 2D.



One manufacturer showed a single-lens (but dual image sensors) 3D video camera based on microstereopsis (above) at the CEATEC show in Japan last October, and 3D pictures from it were demonstrated at the 2009 Consumer Electronics Show.[32] There have been other versions used (with single sets of image sensors) at least as early as the 1970s for television broadcasts.[33]

THE BUSINESS OF 3D TV

The Current Push for 3D TV

Consumers have used 3D technology at least since the publication of the invention of the stereoscope in 1838.[34] All of the view-control technologies mentioned previously (directive, polarized, colored, and shuttering) were developed and used for the display of 3D images before the end of the 19th century.[35] The origins of stereoscopic cinema also date back to the 19th century,[36] and those of stereoscopic television to 1928, almost the first television broadcast.[37] Why, then, does there seem to be a push only now for 3D TV?

A historical search would show that there has actually been development work on 3D TV almost continuously. A 1930 book discusses "several possible methods of accomplishing Stereoscopic Television."[38] A 1938 RCA patent covered 3D TV.[39] Live 3D TV was demonstrated at a 1950 meeting of the Institute of Radio Engineers. *The New York Times* ran a story on April 22, 1980 headlined "3-D TV Thrives Outside the U.S." about work in Australia, Italy, Japan, Mexico and at the ITU.[40] Below are images of 3D video cameras from 1989 (left) and 2001.[41]

There *are* some new developments in the field, however, suggesting that the current era of 3D TV might be different from previous ones. The current highest grossing movie of all time, for example, is a 3D production.[42]

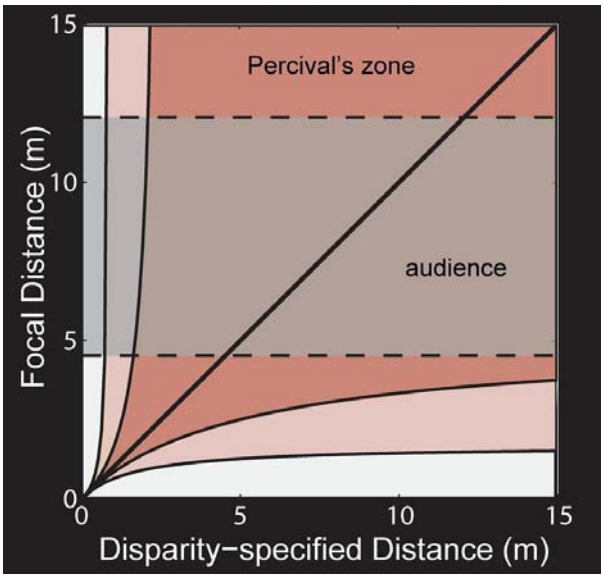
Solid-state imagers and memories allow tiny side-by-side 3D cameras and camcorders, as shown below.[43] Digital processing allows correction of optical distortions and other 3D shooting problems. Advances in display technology make inexpensive 3D TV sets possible, and the redundancy of the two views in stereoscopic imagery suggests low additional bit-rate for 3D signal distribution.

It is possible, therefore, that, from a technology standpoint, and, perhaps, even in

terms of consumer desire, the time has finally come for 3D TV to achieve significant household penetration. As noted, however, the push by TV-set manufacturers is being driven by not those but a desire to overcome the poor sales in "a challenging market." [6]

Audience Issues

Although some of the current push for 3D TV might be based on the success of some 3D movies in cinemas, 3D in movie theaters is very different from 3D in the home. First, there is viewing distance.



As can be seen from the diagram above, a cinema audience falls within Percival's zone of comfort (dark area) for all conditions except vergence ("disparity-specified") distances extremely close to the viewer. A home audience falls outside the zone, however, at many vergence distances, especially at close viewing ranges ("focal distances"). [12]

It's conceivable that, as viewers become accustomed to 3D imagery, their visual systems will tolerate greater vergence-accommodation conflict. Human perception has certainly been trained in other areas. Listeners to Edison's "tone tests" (beginning in 1915), for example, were reportedly unable to distinguish the sounds of phonograph recordings from those of live singers either in concert halls or at close range. [44]

There are many other differences between cinema and TV audiences. Glasses are provided for all 3D cinema viewers. Home viewers must obtain their own.

Inexpensive anaglyph glasses might be provided free to viewers by advertisers, but newer 3D TVs often use active shutter glasses. They allow the set manufacturer to offer 3D capability at low cost (needing just an emitter for the synchronizing signal), but they are an expensive addition for viewers, some currently priced at over $150 per pair. [45]

Even if consumers purchase a number sufficient for all members of a household, that doesn't cover guests, and – at the moment, at least – there is no guarantee that glasses that work on one TV will work on another, so guests can't even bring their own. Standards might eliminate that issue, but there is still one of battery life, currently on the order of *weekly* U.S. household TV-viewing time. [45]

There are also viewers with stereo-visual impairments. Some, for example those blind in one eye, might not perceive 3D effects but can enjoy one eye's view when wearing 3D glasses. Others, however, cannot seem to tolerate 3D images even when wearing 3D glasses. [46] Those viewers can usually self-select non-3D cinema auditoriums in which to view movies also available in 3D versions. At home, alone, such viewers should also be able to view 2D versions of 3D shows (assuming appropriate signal-distribution methods).

If the majority of a group opts to watch 3D, however, those 2D-only viewers cannot watch at all. An option might be 2D glasses, delivering the same view to both eyes. [47] In the case of active shutter glasses, however, such an option might reduce the image frequency to the point where flicker is visible (also a problem for viewers blind in one eye).

Some other differences of home viewing from cinema include multitasking, which might be difficult when wearing 3D glasses, channel-changing between 2D and 3D imagery, and closed captions. As noted, graphics obscuring 3D images must be placed in front of whatever is being obscured.

There might also be viewers who simply do not like 3D. According to a 2009 report, "Approximately 20% of the people who attended a 3D movie did not like it, citing eye fatigue, the eyeglasses and other issues," and "About 5% of people are 'stereoblind' and cannot see in relief."[48]

Distribution Issues

Even within any given image format and frame rate, there are multiple mechanisms for distributing 3D. In the uncompressed domain, these include the side-by-side and over-under systems described previously (in both shrunken and anamorphic versions), field alternation between eye views, frame alternation between eye views, side-by-side with image rotation (in multiple forms), quincunx (alternating-square checkerboard patterns, e.g., left-eye on the red squares and right-eye on the black), dual feed, and left-eye view plus depth information (there is also a variant: left-eye view plus depth information plus graphics information).

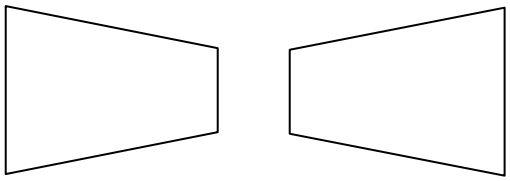
Each has advantages and disadvantages. The disadvantages can include reductions in spatial and temporal resolution. Despite what seems to be a confusing array of formats, however, relatively inexpensive converters are already available. HDMI Licensing has established a small number of mandatory 3D formats in HDMI 1.4a and allows others.[49]

In the compressed (bit-rate-reduced) domain (e.g, MPEG-2, MPEG-4 AVC, etc.), the tremendous redundancy between the left- and right-eye views would be expected to reduce the overhead of the second eye's view well below 100%. Some test results seem to confirm that expectation. Others, however, do not.[50] Some work on finding optimum compression parameters might be required.

Production and Post Production

Post-production processing for 3D includes format conversion, convergence correction, view matching, and 2D-to-3D conversion. Remarkably, even the chromostereoscopic concept of reddish foregrounds and bluish backgrounds, alone, has been used with some success in automatic 2D-to-3D conversion.[51]

There is a debate among stereographers about the value of convergence during shooting. Still-picture 3D cameras typically have not used convergence so as to avoid the image distortion shown below. Digital image processing, however, can correct the problem.



exaggerated view of convergence-based image distortion of rectangular frames

Based on the reduced effects of both stereopsis and vergence angle with increased distance, 3D cameras (or dual-camera rigs) are often placed closer to scenes being shot than are conventional 2D cameras. Many involved in the acquisition of 3D TV imagery also note that 3D seems to require fewer cameras and less cutting between camera positions than does 2D. The need for even a different sound mix for 3D production has been discussed.[52]

Remarkably, the same was said of HDTV in its early days.[53] It might well be the case.

Unfortunately, as the history of HDTV sports and entertainment production shows, 3D-only television programming is unlikely after the experimental period ends. Whereas

sports were once picked up in dual productions, one optimized for HD and the other for standard-definition (SD), today a single production serves both audiences with the faster cutting and greater number of cameras of SD production, and with consideration of its narrower screen as well. 3D productions will likely, similarly, have to serve 2D audiences.

Although sources vary, it appears that U.S. household penetration of HDTV sets only recently exceeded 50%, and, even within those households, not all TV sets are HDTV and not all HDTV sets are used to receive HDTV programming.[54] In 3D, not only is the technology much younger, but vision issues might also prevent some viewers from *ever* purchasing a 3D set. Content programmers will be faced with a largely 2D television audience for many years.

Advancing technology, however, might make possible the use of "virtual-camera" technology for 3D acquisition. Virtual cameras have been used for years. A number of physical cameras capture information about a space and everything within it. Processors can then create a virtual camera that can effectively "shoot" from any position within the space – even "moving" during a shot. The same technique can be applied to stereoscopic virtual cameras.[55] It can also be used to create the multiple views of a multiview autostereoscopic display.

The Little Things

Cameras, lenses, post-production equipment, distribution equipment, and displays might seem to cover all of 3D, but there is more. In addition to major 3D equipment, there are many minor elements required for the complete 3D chain.

Stereoscopic test and monitoring equipment is just becoming available. Stereoscopic viewfinders, due to their very close focal distances, can be difficult to use (and, due to the difficulty of implementing some stereoscopic display technologies, many 3D viewfinders use anaglyph glasses).

A cable manufacturer has just introduced 3D coax (color-coded bound coax pairs). A mobile production facility engineer found a need to create 3D half glasses (shown below) so that crews could look up at 3D displays but also down at dimly lit control surfaces.[56]

Should electronic program guides be in 3D? Should displays scale stereoscopic disparity according to screen size? Despite its 82-year history, 3D TV remains a young field, with discoveries still being made.

THE BOTTOM LINE

TV set manufacturers might make money from selling 3D TV sets. Glasses makers might make money from selling glasses. Movie producers, distributors, and exhibitors might make money from 3D, though even that is not guaranteed.[57]

As for the rest of the industry, although experiments generate both useful information and publicity, it's not yet clear whether 3D will increase revenue, slow a decrease in revenue, or simply cost money. And viewers exposed to multiple 3D technologies might get confused.

Some recent 3D events carried on cable systems have used anaglyph (in different color-pair formats), Pulfrich, and color-shifting glasses, all with images viewable on ordinary TV sets. Newer 3D events are being carried as side-by-side images intended to be converted to something else by special 3D

TVs. Glasses that work for one form of 3DTV won't necessarily work for another, yet even movie-theater glasses have been tried for such home television 3D material as the Michael Jackson tribute in this year's Grammy Awards broadcast.[58]

Gary Shapiro, president and CEO of the Consumer Electronics Association (CEA), addressed the current 3D TV situation in his column in the March/April edition of *Vision*, a CEA publication. "It is early and many challenges must be overcome. We must agree on standards so consumers can invest in glasses. We must understand that those with eye issues, monovision or susceptibility to motion sickness may not appreciate 3D. We need to qualify consumers and set their expectations to avoid 3DTV returns. We need to understand the benefits and any potential harm from 3D viewing."

Shapiro wrote, "3DTV will be a hit." He asked the industry, however, to "back away from irrational exuberance…." "Otherwise," he continued, "we risk launching a new feature that will not meet lofty expectations."[59]

REFERENCES

1. Haier HLC32B, http://bit.ly/aFDsve
2. Academy of Motion Picture Arts & Sciences, 2010 SciTech Awards, http://bit.ly/8EIVtH
3. Diagram from Thomas Edwards, FOX Network Engineering & Operations, "Approaches to Non-glasses-based 3D displays," presented at SMPTE 2009 Tech Conference, Hollywood
4. ITU-R Study Group 6 Report, 14 January, http://bit.ly/8ynsKJ
5. RabbitHoles, http://www.rabbitholes.com/
6. *Multichannel News,* January 7, 2010, http://bit.ly/7HAXHq
7. NHK integral television, http://bit.ly/cnFrno
8. v3 Depth Enhanced Imaging, http://www.inv3.com/index.html
9. American Paper Optics Pulfrich, http://bit.ly/cJ0uMr
10. American Paper Optics ChromaDepth, http://bit.ly/bwIsgP
11. A short video explaining and illustrating the Ames-room illusion may be found here: http://bit.ly/19JEA0; it might be worth noting that stereoscopic image capture is incompatible with this illusion.
12. Martin S. Banks, Vision Science Program, University of California - Berkeley, "User Issues in 3D TV & Cinema," presented at the 2010 HPA Tech Retreat, Rancho Mirage, California; much of the information is also available in David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision,* vol. 8, no. 3, http://bit.ly/dAQWm7
13. Jeroen Stessen, Philips Consumer Lifestyle Advanced Technology Lab, Eindhoven, Netherlands, From "21:9 TV," presented at the 2010 HPA Tech Retreat, Rancho Mirage, California
14. Shojiro Nagata, "How to reinforce perception of depth in single two-dimensional pictures," *Pictorial communication in virtual and real environments,* Taylor & Francis, 1991
15. James E. Cutting & Peter M. Vishton, "Perceiving layout and knowing distance: The integration, relative potency and contextual use of different information about depth," *Perception of Space and Motion,* Academic Press, 1995, http://bit.ly/cCo3w9
16. Neil A. Dodgson, "Variation and extrema of human interpupillary distance," *Proc. SPIE,* 2004, vol. 5291, no. 36, http://bit.ly/brHuaH
17. Eugene M. Helveston, editor-in-chief, "Refractive and Refractive Accommodative Esotropia," *The Strabismus Minute,* vol. 2, no. 24, http://bit.ly/bXpBkZ
18. Vuzix Wrap 920, http://bit.ly/ioDRD
19. 1953 prism glasses image from James Butterfield; Prisma-Chrome glasses from Anachrome, http://bit.ly/8N8idR
20. LeaVision prismatic viewers from KMQ, http://bit.ly/ckeiKz; LeaVision mirrored prototype
21. Inition paper on polarization extinction and stereoscopic ghosting, http://bit.ly/8WWN5g
22. Andrew J. Woods & Tegan Rourke, "Ghosting in Anaglyphic Stereoscopic Images," http://bit.ly/atYYIE; Louis Ducos du Hauron applied the name "anaglyphe" to a color-based view-control stereoscopic system in 1893 (more in reference 35)
23. American Paper Optics TrioScopics, http://bit.ly/9VNHbD
24. Glasses shown are American Paper Optics ColorCode, http://bit.ly/dwfIH6; some versions of 3D Magic's SpaceSpex appear to be similar, http://bit.ly/aE3oBU
25. Review of *Up* by James Berardinelli, for example, http://bit.ly/gCNEG
26. Dolby 3D, http://bit.ly/2d4SXN
27. "Big Differences in 3D TVs," ConsumerReports.org, http://bit.ly/cOmi0J

28. "The Telephone at the Paris Opera," *Scientific American,* December 31, 1881, http://earlyradiohistory.us/1881opr.htm
29. Olympus ME-51S stereo microphone
30. Mel Siegel, Yoshikazu Tobinada, and Takeo Akiya, "Kinder Gentler Stereo," *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems VI,* January 1999, http://bit.ly/cjxCqy, and Mel Siegel and Shojiro Nagata, "Just Enough Reality: Comfortable 3D Viewing via Microstereopsis," *IEEE Transactions on circuits and systems for video technology,* 2000, vol. 10, no 3, http://bit.ly/aPazme
31. Ian P. Howard and Brian J. Rogers, *Binocular Vision and Stereopsis*, Oxford University Press, 1995
32. Sony 240-fps single-lens stereoscopic camera demonstration, http://bit.ly/9kof6D
33. Jimmie D. Songer, Jr., U.S. patent 3,712,199, issued January 23, 1973, http://bit.ly/9pGpXF; a similar system, with horizontal-slit iris, was used by Digital Optical Technology Systems for television broadcasts in Europe and Australia and was exhibited at "The Future of Film Technology" in New York, Jan. 1980.
34. Charles Wheatstone, "Contributions to the Physiology of Vision. Part the First. On some remarkable, and hitherto unobserved, Phenomena of Binocular Vision," *Philosophical Transactions of the Royal Society,* vol. 128, 1838, http://bit.ly/d0hWDY
35. Franz Paul Liesegang, translated by Hermann Hecht, *Dates & Sources: a contribution to the history of the art of projection and to cinematography*, The Magic Lantern Society of Great Britain, 1986
36. H. Mark Gosser, *Selected Attempts at Stereoscopic Moving Pictures and Their Relationship to the Development of Motion Picture Technology, 1852-1903,* Arno Press, 1977, http://bit.ly/co1HRp
37. R.F. Tiltman, "How 'Stereoscopic' Television Is Shown," *Radio News,* Nov. 1928, http://bit.ly/dbSagH
38. Thomas W. Benson, *Fundamentals of Television*, Mancall Publishing, 1930
39. Vladimir K. Zworykin, U.S. patent 2,107,464, February 8, 1938, http://bit.ly/9CPr8B
40. *The New York Times*, http://www.nytimes.com/
41. Ikegami LK-33, http://bit.ly/apydTH, and Canon 3D lens for XL1 camcorder, http://bit.ly/a0IqNj
42. *Avatar,* http://bit.ly/4EeQSs; the 2nd highest grossing movie of all time is the same director's non-3D *Titanic,* however, and no other 3D movie has yet entered the top-10.
43. Lux Media Plan LP-1 stereoscopic camera with 2/3-inch-format imagers, shown at IBC 2009 with the author's forefinger in front
44. "The History of the Edison Disc Phonograph," Library of Congress, http://bit.ly/boiVuU
45. Pete Putman, ROAM Consulting, "3D in the Home: Trends and Products," presented at the 2010 HPA Tech Retreat, Rancho Mirage, California
46. Rafe Needleman, "TV industry turns blind eye to non-3D viewers," CNet News, January 15, http://bit.ly/6Patze
47. Mark Schubin, "2D (not 3D) Glasses," *Schubin Café,* Feb. 3, 2010, http://bit.ly/9Kxd2F
48. "Eyes wide open: 3D tipping points loom," PricewaterhouseCoopers, http://bit.ly/d2vHNI
49. The 3D portion of the HDMI 1.4a specification is available for free download here: http://bit.ly/cectlV.
50. Peter Symes, SMPTE, "3D in the Home: Standards for 3D," presented at the 2010 SportsTechLA, USC; video coverage of the presentation may be found under the title "3D and SMPTE" here: http://bit.ly/cRe4QK.
51. "CRC-CSDM: Colour-based Surrogate Depth Maps for real-time 2D to 3D conversion," IBC 2009, http://bit.ly/caarb0
52. "TV jumps on 3D bandwagon for live sport," *Variety*, March 25, 2010, http://bit.ly/cVS3N6
53. European Eureka Project, for example, [HDTV] "requires fewer cameras," http://bit.ly/bXybaE
54. Leichtman Research, *HDTV 2009: Consumer Awareness, Interest and Ownership,* November 30, 2009, http://bit.ly/6SblPN; at roughly the same time (late 2009), Leichtman put U.S. household HDTV penetration at 46%, Magid at 43%, the Consumer Electronics Association at 63%, and Opinion Research Corporation at 62%, http://bit.ly/dADhJB
55. Oliver Grau and Vinoba Vinayagamoorthy, "Stereoscopic 3D sports content without stereo rigs," BBC Research & Development White Paper 180, originally published in the *Proceedings of IBC 2009,* http://bit.ly/cIrJMk
56. Belden 1694D; 3D "bifocals" by Richie Wirth, Jr., engineer-in-charge of All-Mobile Video's Titan
57. Although the 3D movie *Avatar* is the highest grossing movie of all time, another 2009 3D movie, *Call of the Wild 3D*, which opened on June 12, grossed just $28,682 domestically through March 29 of this year, http://bit.ly/cImyvV
58. Deborah McAdams, "Grammys Get Big Numbers," *Television Broadcast*, Feb. 1, 2010, http://www.televisionbroadcast.com/article/94056
59. Gary Shapiro, "Is 3DTV Over-Hyped?" Shapiro's Spectrum on the Consumer Electronics Horizon, *Vision,* March/April 2010, http://bit.ly/cfhJQQ

AUTHOR

Multiple-Emmy-Award-winning SMPTE Fellow Mark Schubin is an independent technology consultant based in New York. He is the chief information server at http://SchubinCafe.com.

WIMAX LINKS AND OFDM OVERLAY FOR HFC NETWORKS: MOBILITY AND HIGHER US CAPACITY

Ayham Al-Banna
ARRIS Group, Inc.

Abstract

This article proposes a dual-approach solution to augment the US bandwidth in HFC networks. The solution is based on using OFDM channels over the existing cable plant and WiMAX channels over the air.

The proposed approach not only extends the US bandwidth, but also supports backward compatibility, smooth migration, efficient US bandwidth utilization, DOCSIS® 3.0 US channel bonding, load-balancing, US path redundancy, mobility, and. low cost and optimized implementation.

INTRODUCTION

Recent studies show consistent growth in subscribers' bandwidth caused by continuous offerings of killer applications that require faster data rates. Higher speeds are constantly needed to accommodate the transfer of the bandwidth-intensive and latency-sensitive contents associated with these newly developed applications. Not only is the offered bandwidth increasing in the Downstream (DS) direction, but it is also increasing in the Upstream (US) direction, because many applications depend on the Transport Control Protocol (TCP) for data transmission, where the US and DS bandwidths are tightly related. The US bandwidth is also increasing as a result of more services and applications that send more US traffic such as business services, online gaming, and Peer-to-Peer (P2P) transfers.

Hybrid Fiber Coaxial (HFC) networks have very limited return path bandwidth, which presents a serious challenge to Multiple Service Operators (MSOs) and places them at a competitive disadvantage. This paper proposes a solution to extend the US bandwidth in future HFC networks using two parallel (and complementary) approaches, which are described below and shown in Fig. 1:

1. Orthogonal Frequency Division Multiplexing (OFDM) overlay for the cable.
2. Wireless interoperability for Microwave Access (WiMAX) network over the air.

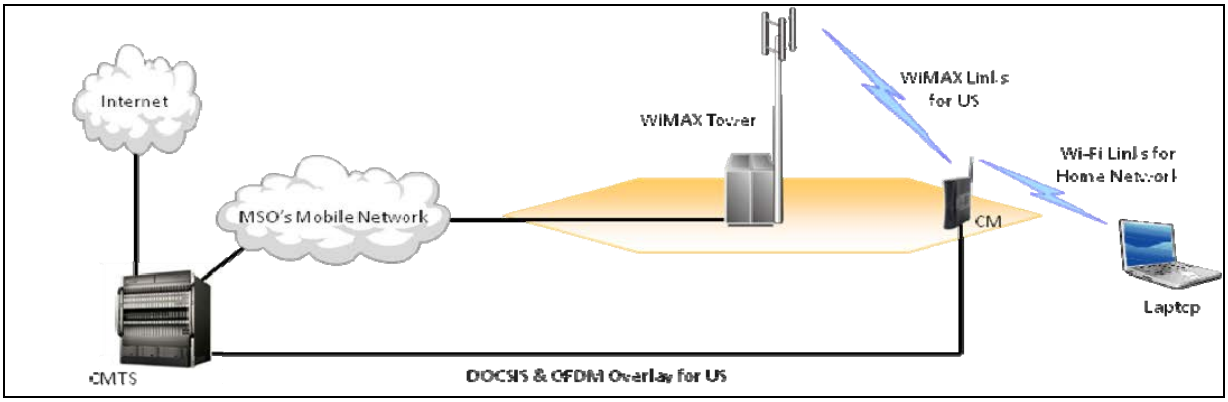


Figure 1. Augmenting the US bandwidth in future HFC networks: OFDM overlay over the cable & WiMAX links over the air

OFDM OVERLAY FOR THE CABLE

The OFDM overlay approach introduces OFDM channels that overlap with the existing conventional DOCSIS® channels but do not interfere with them. This is achieved through the ability of turning on/off the subcarriers that compose the OFDM signal as shown in Fig. 2.

The capability of turning on/off subcarriers is a natural consequence of the architecture of the OFDM signal, which is composed of overlapping and orthogonal subcarriers. In particular, the OFDM signal can be represented as [1] [2]:

$$x(t) = \sum_{i=-\infty}^{\infty} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} a_{k,i} \cdot x_k\left(t - iT_F\right), \qquad (1)$$

where $a_{k,i}$ is the complex symbol modulating the k^{th} subcarrier ($x_k(t)$) in the i^{th} OFDM symbol block. The k^{th} subcarrier $x_k(t)$ is expressed as:

$$x_k(t) = \frac{1}{\sqrt{N}} w(t) \cdot e^{j2\pi f_k t}, \qquad (2)$$

where $w(t)$ is a unity amplitude rectangular window that is nonzero in the range $0 \le t \le T_F$ and $f_k = k/T_F$ is the k^{th} subcarrier frequency.

The $1/\sqrt{N}$ term is a normalization factor included to ensure that the OFDM signal power is independent of the number of subcarriers.

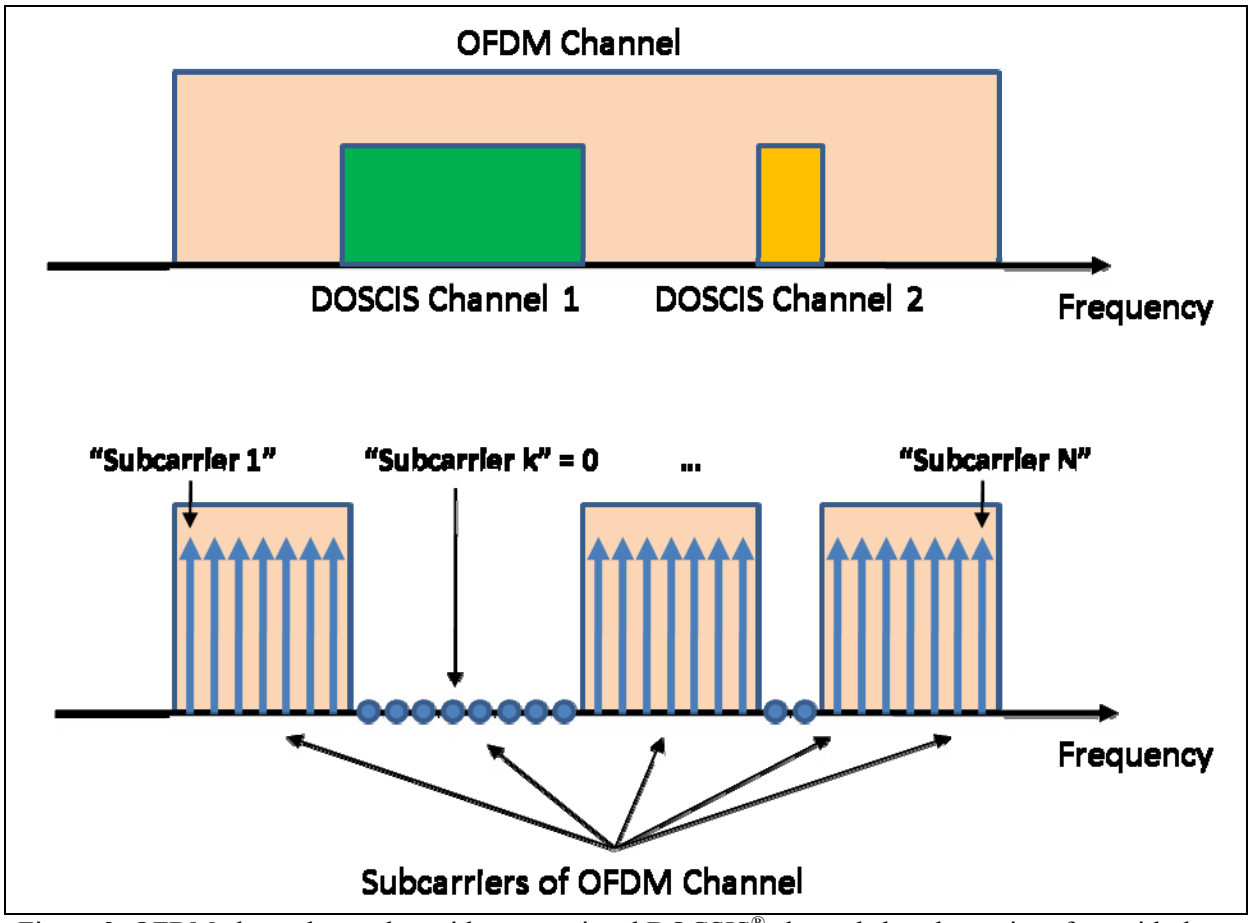


Figure 2. OFDM channels overlap with conventional DOCSIS® channels but do not interfere with them

WIMAX LINKS OVER THE AIR

The WiMAX channels form an alternative return path to the cable. They are based on the OFDM technology and present many benefits as explained in the following sections.

BENEFITS OF THE PROPOSED SOLUTION

This section demonstrates that this dual-approach proposal provides many advantages beyond the basic benefit of expanding the US bandwidth. These benefits are:

1. Backward Compatibility & Smooth Migration

An important advantage of this proposal is backward compatibility, where conventional DOCSIS® cable modems can share the cable with the new modems (Cable & Wireless) through the ability of OFDM to turn on/off subcarriers as explained earlier. The OFDM overlay approach of this solution provides smooth migration toward a new HFC network architecture. It may also provide an easy migration when midsplit bands may be introd-

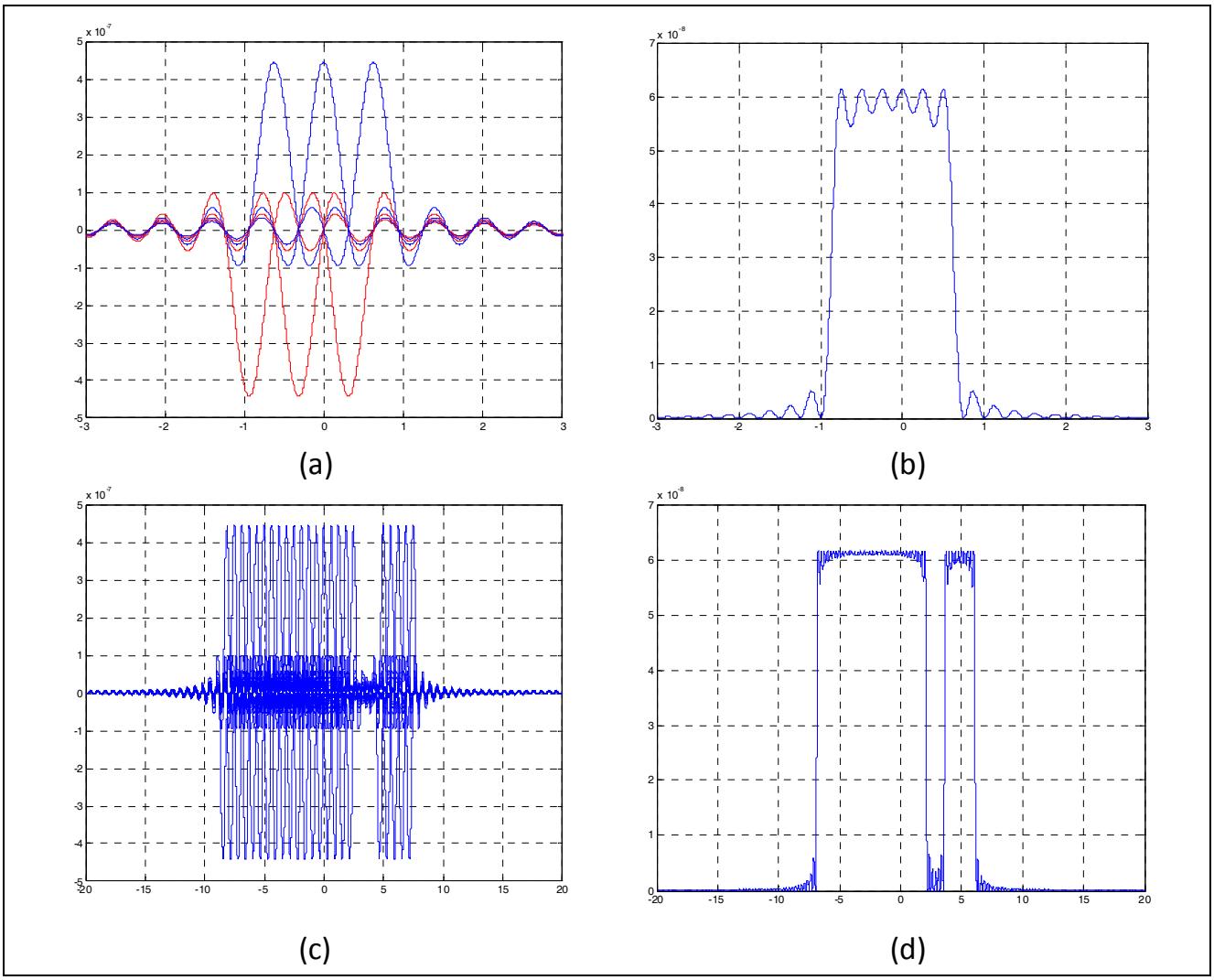


Figure 3. Architecture of OFDM signals: (a) OFDM subcarriers overlap but orthogonal. (b) PSD (w/Hz) of OFDM signal composed of the subcarriers in (a). (c) OFDM signal composed of 52 subcarriers, where 6 of them are off. (d) PSD (w/Hz) of the OFDM signal composed of subcarriers in (c) show a frequency gap within the channel. Note: all graphs are plotted versus frequency (MHz).

uced to the cable, where a single wide OFDM channel may cover two separate bands.

2. Efficient US Bandwidth Utilization

The ability of OFDM to support various channel widths and the ability of OFDM to turn individual subcarriers on and off provides better spectrum utilization by exploiting unused holes in the spectrum which may exist near the band edges or between conventional DOCSIS® channels. Moreover, the feature of dynamically allocating OFDM subcarriers to different modems provides better spectrum efficiency especially when a particular modem does not need all the subcarriers on a single channel. Efficient bandwidth utilization is also achieved using high resolution US data grants (subcarrier mini-slots) offered by the OFDM overlay technology. In particular, a subcarrier mini-slot is defined as a time-slot over a subcarrier bandwidth but not over the full channel bandwidth, as is the case in conventional DOCSIS® channels.

The inherent immunity of OFDM (and Coded OFDM) to different types of noise found in HFC plants enables the operation in challenging noise environments and provides higher goodput (useful throughput) values as little Forward Error Correction (FEC) may be needed [2].

3. DOCSIS® 3.0 US Channel Bonding

The dual-approach proposed in this article provides the flexibility to perform DOCSIS® 3.0 US channel bonding over multiple media, where service flows can be bonded over multiple channels on the cable, WiMAX link(s), or both. This, in turn, introduces evolutionary growth to the priceless US bandwidth in HFC networks.

4. Load-Balancing

Load balancing is another promising feature of this dual approach. Load balancing distributes the user bandwidth across different channels over the cable, WiMAX link(s), or both. Offering more channels through both approaches of OFDM overlay and WiMAX link provides more flexibility to the load-balancing feature, which helps in providing good Quality of Experience (QoE) service to the subscribers.

5. US Path Redundancy

One important advantage offered by this proposal is the redundancy in the return path. This feature contributes significantly in offering high-availability and better QoE services to subscribers because it eliminates or immensely reduces the service downtime, which can be very large especially in cable-related failures. One example of US cable-related failures is a noisy US band resulting from old cabling and corroded RF connections which introduce Common Path distortion (CPD) noise in addition to ingress noise and impulse noise that couple into the deteriorated cable. All these types of noise add to the common Additive White Gaussian Noise (AWGN) causing failure in communication in the US direction.

6. Wireless & Mobility

This solution adds the wireless features and mobility to HFC networks through WiMAX links over the air. Integrating this wireless technology with HFC networks enables the MSOs to utilize all the benefits offered by the WiMAX technology such as mobility, flexible channel widths and modulation profiles, dynamic resource allocation, Quality of Service (QoS), security, etc. This addition also allows the MSOs to use multi-element antennas to offer more bandwidth through spatial multiplexing and intelligent interference mitigation techniques [1]. In particular, spatial multiplexing may increase the channel capacity multiple times through utilizing various wireless paths between the transmitter and receiver using multi-element antennas. Additionally, multi-element antennas are also used with tapped delay lines to provide directional radiation patterns that increase the antennas transmit and receive gains and mitigate interference from other wireless devices [1].

7. Low Cost and Optimized Implementation

Since OFDM is the base technology for both parts of this solution (OFDM overlay and WiMAX), systems designers and RF engineers

may be able to debug, design, and apply the similar solutions and optimization algorithms to both parts of the network even though they exist on different physical media. For example, the principle of allocating OFDM subcarriers dynamically based on subscribers' needs or noise pattern can be applied to both OFDM overlay channels over the cable and WiMAX links over the air.

The architecture of OFDM signals which are composed of narrow subcarriers may eliminate the need for complex pre-equalizers, which results in much simpler and cheaper PHY systems. Additionally, OFDM is based on the Inverse Fast Fourier Transform (IFFT) algorithm, which can be implemented easily and efficiently on processors and therefore produce faster and lower cost systems. Another potential benefit of using the OFDM-overlay technology is implementing a simpler US mapper (scheduler), where certain subcarriers can be assigned to different modems, services, MAC management messages, etc.

Utilizing WiMAX for the return path may provide for easier integration within the subscriber's home network, where a wireless medium is used to connect a variety of subscriber devices. In this scenario, the home network can be based on different wireless technologies such as WiMAX or Wireless Fidelity (Wi-Fi) as shown in Fig. 1. Observe that higher data rates in Wi-Fi devices use the OFDM technology which is also used in WiMAX [1]. Therefore, the wireless home network and the US return path (OFDM overlay and WiMAX) are all based on the OFDM technology, which in turn can help in the evolution of RF gateways.

Finally, employing the standards-based WiMAX technology can speed up the evolution process of HFC plants by avoiding the creation of new standards or specifications and also exploiting well-established equipment offered by different WiMAX vendors.

CONCLUSIONS

A dual-approach solution to augment the limited US bandwidth in HFC networks was proposed. The solution offered two parallel approaches: (1) OFDM overlay channels over the cable plant, and (2) WiMAX links over the air. The solution augments the current US bandwidth in HFC network through adding extra bandwidth resources (WiMAX network) and providing better bandwidth utilization (OFDM overlay). There are many advantages offered by this solution such as: backward compatibility and smooth migration, efficient US bandwidth utilization, DOCSIS® 3.0 US channel bonding, load-balancing, US path redundancy, mobility, and low cost and optimized implementation.

REFERENCES

[1] Ayham Al-Banna, *Interference in IEEE 802.11 WLANs: Characterization and Mitigation*, ISBN: 978-3-639-13280-9, VDM Verlag, 2009.

[2] Ayham Al-Banna and Tom Cloonan, *"Performance Analysis of Multi- Carrier Systems when Applied to HFC Networks"*, SCTE-ET NCTA Conference, April 2009.

Author's Contact Info:

ARRIS Group, Inc.
Address: 2400 Ogden Ave., Suite 180, Lisle, IL 60532, USA
Tel: 630.281.3009
Fax: 630.281.3362
E-mail: Ayham.Al-Banna@arrisi.com

Biography:

Ayham Al-Banna, Ph.D.: Sr. Systems Architect at ARRIS Group, Inc., Chicago. His research interests include RF Communication, Traffic Management, QoE, and QoS. Ayham has published a book and numerous publications in the area of Wireless and Cable Communications.

XML SCHEMA REPRESENTING AN EBIF TEMPLATE DEFINITION, METHOD FOR AUTO-GENERATING SCHEMATIC INSTANCES FROM ORIGINAL EBIF SOURCE CODE AND CONSTRAINING CUSTOMIZED INSTANTIATIONS OF THE RESULTING TEMPLATE

Mike McMahon and Lea Anne Dobbins
Comcast Media Center

Abstract

While there is no standardized source code language for EBIF, the commonly used authoring tools and compilers use XML as their source syntax. Peripheral XML standards such as XML Schema, XPath and XSLT can therefore be leveraged in validation, transformation and marshalling of EBIF source trees. Presented here is a methodology in which an arbitrary EBIF application, developed by any iTV vendor can be automatically "templatized" such that its original source tree is subsequently used to generate data driven, customized instantiations. Our ambition is to alleviate toolset incompatibility resulting from proprietary syntax, compilers and customization toolsets, thereby restoring the spirit of open standards to the end-to-end workflows associated with templating and customizing EBIF applications.

THE NEED FOR TEMPLATING IN EBIF

In order to re-purpose, brand or skin interactive television applications, a non-technical person (e.g. a brand manager) should be able to simply select an underlying template and supply the desired text, graphics and colors needed to generate a customized instantiation. In order to achieve this aim it is necessary to isolate the core logical and functional attributes of an application as a "template," whereby the customized data and stylistic attributes are supplied separately, in a non-technical and user friendly way, in order to define a given "instance." These capabilities need to be fluid enough to demonstrate unique branding and creative elements in the template "instance."

PROBLEM WITH CURRENT SOLUTIONS

Data, logic and presentation

There is a mechanism within EBIF to separate application logic, represented in binary form as a .PR file from application data, represented as a .DR file. This type of separation is certainly useful in order to iterate through data sets in applications such as those that fetch dynamic RSS feeds.

Nonetheless, logic, data and presentation remain largely coupled within EBIF. Consequently, true isolation of a core, logical template in order to expose only the stylistic and data qualities of an application to a non-technical brand manager is not feasible within the current construct. Presentational qualities and logical data binding directives are necessarily part of the core source code, as opposed to an accompanying properties file.

Even if such qualities were to be encapsulated in an external properties file there would still be need to enforce constraints such as string length, image format and dimensions, etc on any given set of instantiation properties. The most effective and safe way to generate a custom application is, therefore, to go back to the original application author with a set of requirements.

While this need not amount to much more than a copy, paste, compile effort on the part of the developer it is neither an efficient use of the developer nor the technology. In

addition, such an approach clearly raises a variety of quality assurance concerns as the underlying code base cannot be assumed to be a static entity. The custom instantiation would, therefore, need to go through a test cycle and although somewhat redundant, in many cases, this test cycle will need to be as comprehensive as that conducted on the original code base.

Proprietary Template Toolsets

There are several "template toolset" products available. Such products do indeed solve many of the problems identified above. They typically provide a simple, non-technical customization interface allowing a user to select from a pre-defined set of underlying templates and provide custom graphics, text and color schemes. These values are then used in a find and replace mechanism against the application's original source code such that a new, unique code base is generated and sent to a compiler, generating a custom instantiation. The more mature systems are additionally validating and constraining user input thereby enforcing output quality.

Such template toolset products unfortunately ship with a fixed set of baseline templates and corresponding customization GUIs. Customized instantiations are confined to the functional capabilities of those pre-canned templates. In order to add an additional functional template to the toolset one must work with that vendor to define and develop it.

It is generally not possible to ingest an application developed by a third party into these types of proprietary toolsets. Moreover, because "templatization directives" are not exposed, any new applications that are added to the system must be authored in a proprietary SDK, typically provided by that same vendor.

Therein lay the primary problem we seek to address: the lack of interoperability between EBIF authoring and templating or customization tools.

THE TEMPLATE DEFINITION SCHEMA

We aim to define an open standard XML schema which is intended to serve as a structured, strongly typed interchange between EBIF authoring tools and the template toolsets used for re-skinning and customizing specific instantiations of those applications.

Specifically presented here is an XML data model we call templateDefinition.xml and a functional reference implementation as it relates to ingesting a new application into a template toolset. The data model and methodology described herein allows an arbitrary EBIF code base, authored outside of and independently from the template toolset, to be ingested into the template toolset in a manner in which the "templatizable aspects and constraints" are identified and understood.

The application can then be added to the set of available templates, allowing the template toolset to reliably render, capture and validate the instantiation parameters in its customization GUI.

Our belief is that such a schematic representation of "where and how an application is customizable" should be adopted as an open standard, thereby allowing applications developed in any EBIF authoring tool to be ingested into external, third party customization tools.

Template Definition XML

Figure 1 below illustrates the crux of the data model. It provides document pointers to each file within the code base and, for each file, XPaths to the precise nodes and/or

attributes that could be reasonably and successfully customized. For each of these "templatizable items" the necessary constraints on the instantiation parameters are additionally defined.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <codeBase>
  + <doc location="TemplateTest/application/app-desc.xml">
  - <doc location="TemplateTest/application/page-resource/page/index.xml">
    - <templateItems>
      - <item id="31">
          <xpath>/mac/var[1]</xpath>
          <nodeType>var</nodeType>
        - <passThroughValues>
          - <attributes>
              <name>templateItem-theText</name>
              <type>string</type>
            </attributes>
          </passThroughValues>
          <transfromation default="Go Blue" type="element" />
        - <constraints>
            <constraint type="required">true</constraint>
            <constraint type="length">7</constraint>
          </constraints>
        - <originalNode>
            <var name="templateItem-theText" type="string">Go Blue</var>
          </originalNode>
          <name>templateItem-theText</name>
          <description>templateItem-theText</description>
        </item>
      </templateItems>
    </doc>
  + <doc location="TemplateTest/application/page-resource/style/palette.xml">
  </codeBase>
```

Figure 1: templateDefinition.xml

Physical Location of Source Code

The template toolkit needs build time access to the original source code in order to compile a given instance. We use URIs as pointers to the location of the original source code. With respect to third party IP, there are a couple options in terms of the physical location of the source code. It could be hosted by the original application developer and dynamically accessible to a template toolkit when compiling a custom instance. Alternatively, application authors could upload source code to the template toolkit. Either scenario necessitates contractual protection of the source code and associated IP.

Generating the Template Definition XML

The original developer of a given application is, clearly, the authority with respect to establishing which elements within the application could or should be safely customized. We therefore seek a mechanism whereby the original developer can compile a default instantiation while establishing specific text strings, variables, integers, colors or graphics as "customizable."

Likewise, the original developer is best able to define necessary constraints during the customization process. For example, the default value of a given text message within the application might be ten characters long. A message of twelve characters would be perfectly acceptable but a message of fifteen or more characters would cause line wrapping, detrimental to the visual appearance of the overall screen. It is therefore necessary to solicit not only the "templatizable aspects" of the application from the original developer, but also the corresponding constraints.

Our template definition XML, of course, describes both. The question, however, becomes how is that definition itself generated? Our view is that if such a data model were widely adopted it would likely be the case that EBIF authoring tools would automatically generate the template definition XML as a supplementary output of the authoring and compilation process. Perhaps application authors would highlight blocks of code and right click to bring up a dialogue box in order to capture the constraint definitions.

In lieu of template definition files generated from an authoring tool we had need to supply our own by way of an external, supplementary file. It is far from desirable to introduce the risk of human typos when creating the template definition XML as a freehand effort. It was also not possible to inject innocuous markup into the source code as the compilers would reject the syntax.

In order to automatically generate a compliant template definition XML file and remain both agnostic and innocuous to existing compilers we introduced a naming convention in the authoring syntax such that the application author prefixes potentially

customizable areas of the source code with *templateItem-*. This allows the original application author to surgically pinpoint specific areas of the source code that can, should or must be customized. Additionally, because this is a naming convention as opposed to an extension of the source syntax itself, it does not affect the existing compilers and can be used within any XML based EBIF source syntax.

Given such a naming convention within the underlying source tree we are able to programmatically traverse the whole of the application's source and extract the precise location of all "templatizable aspects" of the application as defined by the original application author. Figures 2 and 3 below illustrate an XSLT script that will traverse an EBIF source tree written in the TVWorks MAX syntax and generate a normalized template definition XML file. The logic in the XSLT will automatically derive the XPath and constraints. It takes a first pass through the source tree, indentifying each node flagged as "templatizable" by the author and holding them in memory.

```
<xsl:param name="srcDir" as="xs:string" required="yes" />
- <xsl:variable name="theCollection">
    <xsl:value-of select="$srcDir" />
    ?select=*.xml;recurse=yes
  </xsl:variable>
  <xsl:output method="xml" indent="yes" encoding="ISO-8859-1" />
- <xsl:template match="/">
  - <xsl:variable name="start">
    - <codeBase>
      - <xsl:for-each select="collection(iri-to-uri($theCollection))">
        - <xsl:variable name="theDoc">
            <xsl:value-of select="document-uri(.)" />
          </xsl:variable>
        - <xsl:element name="doc">
          - <xsl:attribute name="location">
              <xsl:value-of select="$theDoc" />
            </xsl:attribute>
          - <xsl:element name="templateItems">
            - <xsl:for-each select="document($theDoc)//node()">
              - <xsl:variable name="theAtt">
                  <xsl:value-of select="substring(@name,1,13)" />
                </xsl:variable>
              - <xsl:if test="($theAtt='templateItem-')">
                - <xsl:element name="item">
                  - <xsl:element name="xpath">
                      <xsl:value-of select="saxon:path()" />
                    </xsl:element>
                  - <xsl:element name="originalNode">
                      <xsl:copy-of select="." />
                    </xsl:element>
                  - <xsl:element name="nodeType">
                      <xsl:copy />
                    </xsl:element>
                  </xsl:element>
                </xsl:if>
              </xsl:for-each>
            </xsl:element>
          </xsl:element>
        </xsl:for-each>
      </codeBase>
```

Figure 2: XSLT first pass traverse

We then take a second pass through the memory tree in order to analyze individual node context, group and define each of the items:

```
- <xsl:if test="nodeType/var">
    <xsl:element name="nodeType">var</xsl:element>
  - <xsl:element name="passThroughValues">
    - <xsl:element name="attributes">
      - <xsl:element name="name">
          <xsl:value-of select="originalNode/var/@name" />
        </xsl:element>
      - <xsl:element name="type">
          <xsl:value-of select="originalNode/var/@type" />
        </xsl:element>
      </xsl:element>
    </xsl:element>
  - <xsl:element name="transformation">
      <xsl:attribute name="type">element</xsl:attribute>
    - <xsl:attribute name="default">
        <xsl:value-of select="originalNode/var" />
      </xsl:attribute>
    </xsl:element>
  - <xsl:element name="constraints">
    - <xsl:element name="constraint">
        <xsl:attribute name="type">required</xsl:attribute>
        true
      </xsl:element>
    - <xsl:element name="constraint">
        <xsl:attribute name="type">length</xsl:attribute>
        <xsl:value-of select="string-length(originalNode/var)" />
      </xsl:element>
    </xsl:element>
  </xsl:if>
- <xsl:if test="nodeType/color">
    <xsl:element name="nodeType">color</xsl:element>
  - <xsl:element name="passThroughValues">
    - <xsl:element name="attributes">
      - <xsl:element name="name">
          <xsl:value-of select="originalNode/color/@name" />
        </xsl:element>
      </xsl:element>
    </xsl:element>
  - <xsl:element name="transformation">
      <xsl:attribute name="type">attribute</xsl:attribute>
      <xsl:attribute name="targetAttribute">value</xsl:attribute>
    - <xsl:attribute name="default">
        <xsl:value-of select="originalNode/color/@value" />
      </xsl:attribute>
    </xsl:element>
  - <xsl:element name="constraints">
    - <xsl:element name="constraint">
        <xsl:attribute name="type">required</xsl:attribute>
        true
      </xsl:element>
    - <xsl:element name="constraint">
        <xsl:attribute name="type">fixedlength</xsl:attribute>
        6
      </xsl:element>
```

Figure 3: XSLT analyze and generate

This process results in a single, normalized template definition XML file. It is this file, in conjunction with the original source tree, which a third party template toolset can now ingest, interpret and reliably expose a corresponding customization interface. Insofar as the authoring tool and templating toolkit are independent pieces of proprietary

software from two different vendors, the template definition XML file serves as a data interchange able to abstract away those proprietary underpinnings and achieve interoperability between these two crucial components of the overall workflow.

REFERENCE IMPLEMENTATION

In order to exercise the data model and prove out the interoperability we have implemented a basic template toolkit to ingest an application and its template definition XML file. This is done as a web system with two login roles. The first role is that of an application developer wishing to upload and "templatize" their application. The second is for a "customizer," the individual interested in selecting from the overall set of available templates and generating a customized instantiation.

In this implementation Tomcat is used as web server and servlet container, Saxon as an XSLT processor and Oracle as a database. The database maintains names and descriptions of available templates as well as pointers to the original source tree and corresponding template definition XML file.

Application Upload and Template Definition

Application developers are presented with a simple HTML form page to enter the name and brief description of their application. The source tree is uploaded as a single .zip file which is unzipped into the server's file system and parsed by the XSLT script. The script discovers any "templatizable items" within the source tree and generates a single, templateDefinition.xml file. The developer is asked to provide some additional, human readable information to assist a customizer in understanding the significance of each customized item. We ask the developer to define a name and briefly describe each of the items. Once done, the template definition XML file is updated with the additional information and the developer has completed the upload. Figure 4 below illustrates the two upload screens as presented to the application developer.

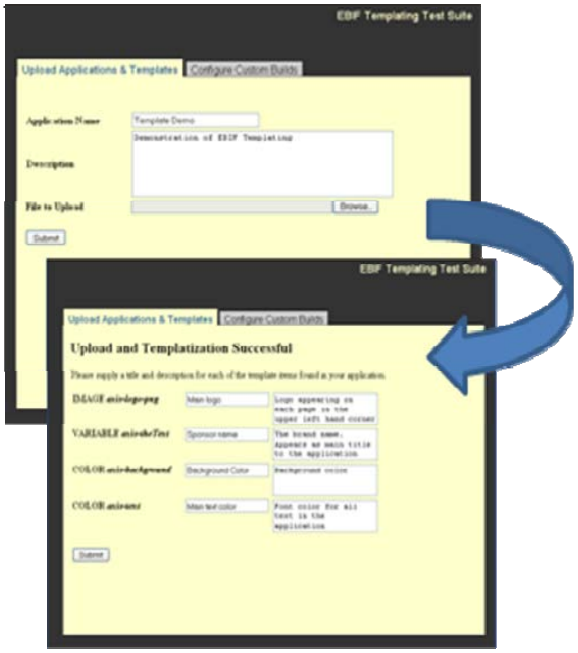


Figure 4: Uploading an application

Figure 5 below represents the logic and data flow associated with ingesting a new application and generating its template definition XML. Ultimately, we persist a name and description of the EBIF application as well as URI pointers to both the .zip file of the original source tree and generated template definition XML file.

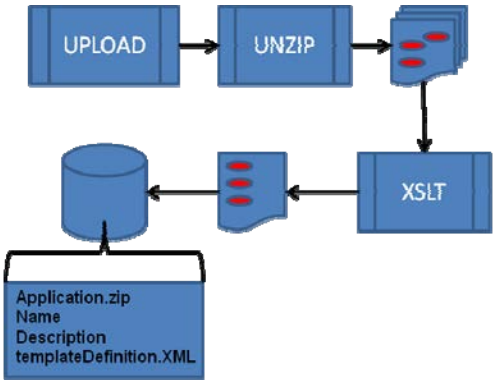


Figure 5: Ingest logic and data flow

The application is then deemed a template and ready for customization. Logging in as a customizer, the user is presented with a list of all templates in the database. Selecting any template will present the user with a screen for supplying the necessary customization values. Figure 6 below illustrates the two screens as presented to the customizer.

Figure 6: Customizing an application

It should be noted on the second screen that the template definition XML file itself is used to generate the customization interface. The customizer interacts with familiar HTML forms, where each input field is tailored to the attribute in question and constrained accordingly. Graphics have file upload fields, strings are constrained text input fields and colors are defined through a standard JavaScript color picker widget. The template definition XML is also used to generate field by field validation JavaScript such that when the customizer submits the form each field is validated by the web browser and it is impossible to post any values breaking the constraints defined by the original application author.

XSLT is then used as a find and replace mechanism against the original source tree, replacing the result of all XPath expressions found in the template definition file with new values in the instance definition file. The resulting source tree can then be compiled into the customized EBIF binary. Figure 7 represents the logic and data flow associated with the customization process.

Figure 7: Customization logic and data flow

Managing MSO and User Agent Variations

This methodology can be summarized as a normalized find and replace system with the actual customization achieved at build time. This technique achieves customization with respect to re-skinning and brand repurposing motivations. With respect to true end-to-end interoperability, however, we would be remiss if we did not address variations among such things as MSO navigational paradigms, user agent execution and integration with third party guide, DVR and VOD systems.

Each of these begs for their own level of unique customization, quite different and more complex than the surface level look and feel modifications desired by the non-technical brand manager. In addressing interoperability across navigational paradigms, consistency in UI dialogue screens, backend interoperability, etc, we nonetheless believe the methodology described here is a promising approach.

Where the non-technical brand manager seeks to replace specific application resources; an MSO, user agent or backend

system would need to replace whole components or methods calls within the application in order to achieve UI consistency across the plant or interoperability with backend systems, the user agent or other software on the set top box. For example, specific method calls to set DVR recordings or perform VOD telescopes may vary depending on the particulars of the guide, DVR or VOD system. Similarly, in the interest of uniformity, MSOs may seek to standardize navigational paradigms or such things as button labels, placement and on-click behaviors.

This is achievable at build time, whereby sets of pre-established blocks of source code, representing the unique, desired method calls and UI components are compiled into the application. Again, this is fundamentally a build time, find and replace mechanism not unlike the system we have described above. We believe, furthermore, that a well defined, standardized naming convention at the source code level would accommodate these needs. Application developers would indicate such replaceable blocks of code with naming conventions such as:

- templateItem-ConfirmationScreen
- templateItem-VODTelescope
- templateItem-DVRSetting

CONCLUSIONS

In order for EBIF to achieve critical mass it is essential to reduce the time to market and minimize the QC associated with individual applications while maximizing the level of creativity and flexibility in appearance of customized instantiations. This is best achieved by wide adoption of pre-approved templates and strong validation in customization tools. This aim must additionally and necessarily encourage innovation amongst a wide range of iTV vendors and independent application developers. This is the ecosystem from which new, compelling features and revenue models will be born and their applications will need to be made readily available as core templates within customization tools.

The naming convention, XML schema and XSLT transformations described here represent the underpinnings of potential standardizations, whereby application developers could easily designate "templatizable" aspects of their applications in a manner in which compilers and customization tools could reliably ingest, parse and process them. By implementing this in an open standard approach as presented here, authoring tools are decoupled from customization tools such that discrete components of the overall value chain become truly interoperable.

FUTURE WORK

Direct support in authoring tools

The XSLT used in this reference implementation to generate the template definition XML assumes that the TVWorks XDK is used as the authoring tool. While the general technique is theoretically agnostic to the particular XML authoring syntax, the node inspection logic within the XSLT is specific to the TVWorks MAX syntax.

This is only required to automatically generate a template definition XML file. The template definition XML itself is its own, standalone data model such that the XPath expressions and constraint definitions can be applied to any underlying XML based source syntax.

Ultimately, our view is certainly that EBIF authoring tools would intrinsically generate such template definition files and there would not be a requirement for the template toolset to generate one.

Validating color palettes and graphics

Color palettes are defined in the EBIF source code and all graphics included in the application are confined to those defined colors. The methodology described here allows a non-technical person to manipulate the underlying color palette and provide custom graphics. There exists potential conflict and limitations where an author might have a "core graphic" which must be preserved in all instantiations and the customizer finds the remaining colors cannot accommodate their desired graphic. The symptom is more pronounced on low-end environments, limited to sixteen colors.

Conventions surrounding graphics and color palettes should be explored. In a sixteen color environment things are clearly highly constrained, but it should be possible for an application author to provide and define (in the template definition XML) an acceptable set of potential palettes and any corresponding "core graphics." In the richer 256 color palettes the symptom is greatly alleviated and the solution is potentially as simple as earmarking a conventional set of a certain number of colors (64 or 128) as customizable.

Validation and Instantiation via Web Services

In our reference implementation we have used XSLT to inspect a set of EBIF source documents for particular naming conventions and node patterns. We did this in order to auto generate normalized template definition files in lieu of them being created by authoring tools. Whether or not authoring tools implement the schema, the inspection technique itself appears useful with regards to potentially automating some basic tests. For example, source code could be dynamically inspected using similar XSLT scripts for valid organization IDs, appropriate calls to terminate() methods, approved VOD handlers, correct HTTP POST parameters, allowable remote control keys, etc. These are examples of any number of scripted tests which could be used to perform automated validation based on an inspection of source syntax.

REFERENCES

[1] T. Bray et al, Extensible Markup Language (XML) 1.0, W3C Recommendation, November 2008 http://www.w3.org/TR/2008/REC-xml-20081126/

[2] J. Clark XSL Transformations (XSLT) 1.0, W3C Recommendation, November 1999 http://www.w3.org/TR/xslt/

[3] J. Clark and S. DeRose, XML Path Language (XPath) 1.0, W3C Recommendation, November 1999 http://www.w3.org/TR/xpath/

[4] H. Thompson et al, XML Schema Part 1: Structures, W3C Recommendation, October 2004 http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/

[5] P. Biron and A. Malhotra, XML Schema Part 2: Datatypes, W3C Recommendation, October 2004 http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/

ISBN 0-940272-01-6; 0-940272-08-3; 0-940272-10-5; 0-940272-11-3; 0-940272-12-1; 0-940272-14-8; 0-940272-15-6; 0-940272-16-4; 0-940272-18-0; 0-940272-19-9; 0-940272-20-2; 0-940272-21-0; 0-940272-22-9; 0-940272-23-7; 0-940272-24-5; 0-940272-25-3; 0-940272-26-1; 0-940272-27-X; 0-940272-28-8; 0-940272-29-6; 0-940272-32-6; 0-940272-33-4; 0-940272-34-2; 0-940272-35-0; 0-940272-36-9; 0-940272-37-7; 0-940272-38-5; 0-940272-39-3; 0-940272-40-7; 0-940272-41-5; 0-940272-42-3; 0-940272-43-1; 0-940272-44-X; 0-940272-45-8; 0-940272-46-6; 0-940272-47-4; 0-940272-48-2; 0-940272-49-0; 0-940272-50-4; 0-940272-51-2; 0-940272-52-0; 0-940272-53-9; 0-940272-54-7