SCALING MOUNT EVEREST: DELIVERING MULTI-SCREEN VIDEO IN AN 'INFINITE CONTENT' WORLD

John Pickens, Chief Technical Strategist VCNBU
Cisco Systems, Inc.
Sree Kotay, Chief Software Architect
Comcast

Abstract

The consumption paradigm for TV is rapidly changing from pure broadcast to time-shifted unicast. This behavioral model is the driver for the new formula, "Cached Unicast equals Multicast". Supporting this trend is the rapid evolution of the network paradigm from a classic siloed broadcast dominated spectrum to a shared spectrum with converged usage of IP transport for all applications including video. The long range vision is tens of thousands of channels, hundreds of millions of assets, and orders of magnitudes more content producers – all delivered to the device of the consumer's choosing. This paper identifies key characteristics of the next generation solution architecture, such as real time enabled cache distribution hierarchies, in order to deliver an infinite world of content and unlimited scale of subscribers and consumption modalities, while delivering many of the economic benefits of today's architectures.

OVERVIEW

The increasingly rapid user adoption of time shift TV, new HD content (requiring multi-carry) and interactive video services [like video on demand (VOD)], coupled with the exploding popularity of blogging and audio/video podcasting, along with higher delivery data rates (e.g. DOCSIS 3.0) and two-way connectivity, requires a revolution in service delivery for media content. The initiative of Switched Digital Video (SDV) for linear video channel delivery is an early recognition of the emerging paradigm of long tail consumption and niche programming in the core TV market. Time shift TV (even popular linear video becomes unicast), the growing libraries of high quality commercial video (movies, original cable shows, made-for-TV, and straight-to-video) and user generated content is accelerating this paradigm shift, thereby stretching the limits of existing multicast and pitcher/catcher video delivery systems to be competitive.

The formula of "cached unicast = multicast", as embodied by Content Distribution Networks (CDNs) like Akamai, becomes more and more desirable as usage patterns change and different device types proliferate. This proposed shift enables the video delivery system to deliver an extreme scale of available assets, including multiple formats, rates, and resolutions for the same asset, with little economic or operational impact. However, traditional Internet CDNs lack the proper control semantics (e.g. you never need to "rewind" a web page), and scale of solution (latency, throughput and cost scalability).

This paper identifies the next generation technologies and paradigms in real-time media delivery that enable cable operators to migrate to this new world. Points highlighted include a massively scaleable authoritative storage network, transition to more distributed architecture designed for media, dynamic caching in the interior and at the edge of segmented content, and n-screen enabling application paradigm, where resource management and authorization enforcement is built into the next-generation network (NGN) media delivery infrastructure.

VIDEO CONSUMPTION

Consumer video consumption behaviors are undergoing a paradigm change, encapsulated by the concepts of time shifting, place shifting, and device shifting. This phenomenon is portrayed within Figure 1.

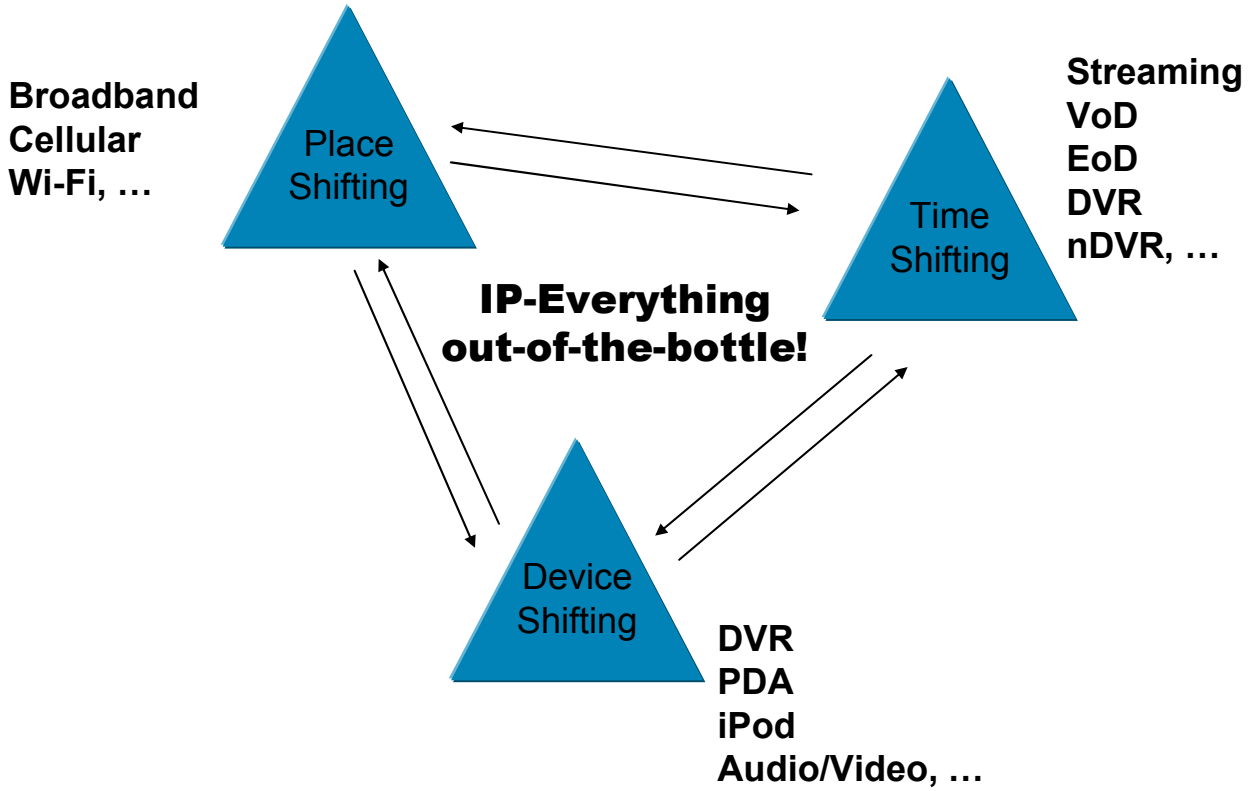

Figure 1 – Consumption Paradigms

For TV, the time shifting phenomenon gained mainstream acceptance with the widespread adoption of digital video recording (DVR). Though a large percentage of homes do not yet have DVR, and those that have a DVR do not have it for all TVs (or a media center), a high percentage of consumers are now very familiar and accustomed to DVR time shift consumption, whether in their own home or in homes of family and friends.

Place shifting is a new emerging trend increasingly being promoted by service providers, typified by "multi-room DVR" and "whole home VoD", where an individual may choose a content to view on one device, then pause the content (bookmark), and then migrate to another device and resume consumption. TV (HD downstairs, SD upstairs), PC (in office or in hotel), and mobile (while riding the shuttle to the airport) are three well known device types for converged consumption of media.

Device shifting is an increasingly discussed paradigm where, in addition to streaming content from the network, it is possible to download the content to different devices, and view the content on those devices. MP3/MP4 players, mobile phone with storage, and laptop PCs are three examples of such devices.

Use cases demonstrating the combination of all three follows. Within the home the content is consumed on an HDTV, paused, and then

resumed upstairs on an SD TV. Alternatively, the subscriber may travel to a hotel and prefer to view the content on his/her PC. Another example includes a subscriber in a limousine/car who prefers to view it on his/her mobile device. Or the content might be downloaded to his/her laptop and he/she views it on the airplane. In all these cases the content is resumed from wherever it was paused or bookmarked, independent of the type of device or location of consumption.

All three consumption paradigms can already be seen in the internet web browsing model for video consumption. The community of Internet users familiar with these models has grown to a staggering numbers, with over 10 billion unique video views consumed monthly, with YouTube accounting for approximately 30 percent of that number.

Also driving the change in user experience is the explosive growth of HD content and exponential expansion in the number of content producers – a consumption feedback cycle highlighted by the early trends of blogging and podcasting, now extending to video.

All these paradigm shifts require a revolution in mechanisms for enabling service delivery of media content, because traditional multicast pub/sub models lack the cost and operational scalability to compete effectively.

Subscriber Quality of Experience

The TV consumption experience of users served by content within the "broadcast" network is significantly different and higher quality that that of today's users who are served by internet or mobile services.

In the internet model (e.g., YouTube) the consumer has been conditioned to accept lower quality consumption experiences. Experiential examples include long latencies while waiting for the picture to display after the start of streaming, the inability to seamlessly transition into trick mode behavior, forward or rewind, and experiencing random display pauses while repairing under-runs of the elasticity buffer in the PC.

For TV quality consumption, by contrast, the user experience delivered by the network is expected to be extremely high quality. An example is the requirement for low latency (subsecond) delay from stream event to stream action. Examples of stream events are stream start, trick modes (fast forward, rewind), and interactivity (e.g. pause → pause-ad).

Subscriber Infinite Content World

The content universe is growing. Whereas a typical library size for Video on Demand (VoD) used to be a few thousands of hours, it is now targeted to be much larger, on the order of hundreds of thousands of hours and eventually millions of hours. [1]. In early 2007 [2,3] Netflix announced an online library of 70,000 titles. Now the estimated library size is well over 90,000 titles, and a high percentage of newly added titles are HD Blu-ray format reflecting the popularity of high definition programming. Comcast in January 2008 announced plans for Project Infinity to grow the On Demand library to 6000 titles (3000 in HD) in 2009, with that number expected to scale dramatically thereafter [4].

One of the key design differences between today's video delivery systems and those of tomorrow is the split between content discovery (asset metadata, availability, and associated information) and content distribution (the physical movement of the asset from source to consumer.

As the number of available assets grows, it becomes untenable (and undesirable) economically and operationally to scale edge

capacity against the number of assets. Instead, edge capacity must scale against the number of ***unique*** assets consumed. This design criterion demands a separation of data flow from media flow.

An interesting number foreshadowing future content volume growth pertains to the amount of user-generated video content on the Internet. While the quality is not as good as professionally produced content, its growth is explosively accelerating. Based upon unpublished monitoring done by search companies, in early 2007 the estimated number of titles was in the order of 40,000,000. By the beginning of 2008 the number of titles had grown to around 120,000,000.

In addition, operators are beginning to offer managed services that enable users to generate their own content and make it broadly available either downloaded online or as part of user-generated channels.

Increasingly, professional content producers are opening up their content archives to consumers, both directly (called over-the-top) and via managed relationships with service providers (assured quality of experience).

IP NGN VIDEO ARCHITECTURE

Three key initiatives for achieving a video enabled IP NGN architecture are defined. First is a series of infrastructure convergence initiatives required in order to increase the diversity of content delivery and user consumption experiences. Second is a real time enabled caching architecture for content distribution. Third is a transformation to make the content format, place, and device independent in order to deliver n-screen delivery.

Convergence Initiative

In order to deliver the universe of infinite content, and assure the DVR-like experience from within the network, a number of convergence initiatives are underway.

Perhaps most enabling convergence activity is the rapid evolution to a wideband all-IP infrastructure. This transition was foreshadowed in the Video-QAM universe by the migration toward IP enabled QAMs (IP to QAMs, traditional MPEG to home). Switched Digital Video (SDV) [5] for linear video channel delivery was the next step recognizing of the need to rapidly evolve infrastructure in order to free up bandwidth for next generation services. The key insight of SDV, versus traditional broadcast to the home, is that ***it is desirable to scale content against consumption, instead of against the total corpus of availability***. It is now accelerating with the evolution toward DOCSIS 3.0 (wideband all the way to the home) and universal QAMs, which allow service channel sharing across VOD, high speed data, and video services.

As currently portrayed in Figure 2, a DOCSIS enabled wideband infrastructure will enable 6 Gbps aggregate IP enabled spectrum downstream – competitive with other service providers – on a 950 MHz plant.

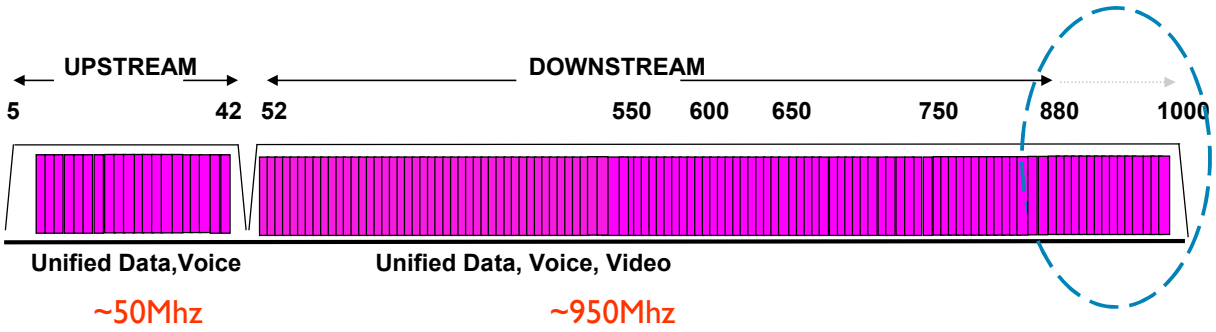

Figure 2 – 6Mbps DOCSIS convergence

A second enabling initiative is the continuing effort to reduce fiber node size. Two driving factors are service enablement and the competitive need to offer higher broadband bitrates.

For service enablement, even at 250 homes passed with 100 percent subscriber penetration, a 6 Gbps infrastructure can serve 750 MPEG-4 HD streams (8Mbps per stream) – 3HD streams per home. The reality is that subscriber penetration is less than 100 percent, and all TVs are not HD, and thus sufficient bandwidth exists even at higher HHP ratios.

For addressing competition a motivation for reduction of fiber node size is the need to further increase peak bandwidth offered per subscriber. For example, optical fiber technologies such as EPON are migrating from today's 2 Gbps:1 Gbps:32 to tomorrow's 10 Gbps:10 Gbps:32 (ratio of down:up:homes). A 6 Gbps DOCSIS® downstream is already higher than the 2 Gbps downstream offered in EPON architectures today and is well in the league of Ethernet Passive Optical Network (EPON) 10 Gbps downstream architectures. The only significant difference between DOCSIS and EPON will be the number of homes sharing the bandwidth, and the amount of spectrum offered for DOCSIS® enabled converged IP delivery.

Other convergence initiatives not addressed in this paper include bandwidth management, metadata, standard advertising interfaces, digital rights management, real time streaming protocol, Digital Living Network Alliance/Universal Plug and Play, conditional access systems, etc. Ultimately these issues need to be addressed as challenges abound. Unlike video services of the past, new services must be delivered to all types of devices in myriad locations -- with high quality and to massive scale. A new architectural approach is needed.

Real Time Caching Initiative

Web caching is a well understood and widely deployed paradigm which features the transient storage of web objects such as HTML documents for subsequent retrieval. Caching enables reduced bandwidth consumption, reduced load on servers with the authoritative storage of content, and reduced interactive latency. Overall it increases the user quality of experience, and reduces network infrastructure cost. [6, 7]

Web caching can be deployed in a variety of modes, from client, to proxy, to arrays of front ending servers. In this paper we focus on caches placed within the network.

Delivery of real time video via caching has similar benefits to delivery of web objects via caching. Consumption characteristics exhibit a Zipf curve phenomenon [8] where more popular content (e.g., a show now playing, though time shifted) is viewed by more people. The first person to consume a video causes it to be downloaded from the authoritative source into cache, and the next person who consumes the video accesses it from cache. No subsequent network transport is consumed upstream of the cache, and the access latency is shorter (by a few hundred milliseconds in worst case). [9]

Given the rapid migration from real time consumption to time shifted consumption, and the existence of the real time caching function, the benefits of caching derive similar benefits to multicast distribution at the edge, with the difference that consumers no longer need to consume content at the same real time timeline. The characteristic that one copy of the content (first user) is distributed across the backbone toward the edge is like multicast. The characteristic that subsequent consumers of the content generate no backbone traffic is also like multicast.

Three significant differences between web caching and real time video caching are identified in this paper. First the bandwidth and size consumed by "objects" is substantially higher. For MPEG-2 HD, the average bandwidth is about 15 Mbps – though it is reduced to approximately 8 Mbps for MPEG-4/AVC. Furthermore, the size of objects (the sum of all object segments) can be in the n*Gigabyte size range. Second the service level expectation of the consumer is higher than it is for web content. The jitter requirement in real time content delivery is much smaller than it is for web services and requires that the consumer experience no visible artifacts or delays. Third, there are multiple correlated object segments being delivered in real time video consumption – the 1x media stream, multiple fast forward renditions of the media stream, and multiple rewind renditions of the media stream.

Therefore additional characteristics are required within the cache delivery infrastructure for video. These are outlined below.

Tiered Hiearchies

A caching hierarchy is defined for real time content distribution. Figure 3 highlights several possible configurations. The number of tiers deployed is arbitrary – it can be minimal depending on the consumption characteristics of devices downstream (e.g. number of subscribers signed up for video service). As consumption demand grows, additional cache storage can be deployed either in parallel or in hierarchies in order to manage the tradeoffs between concurrent usage and latency and resource consumption.

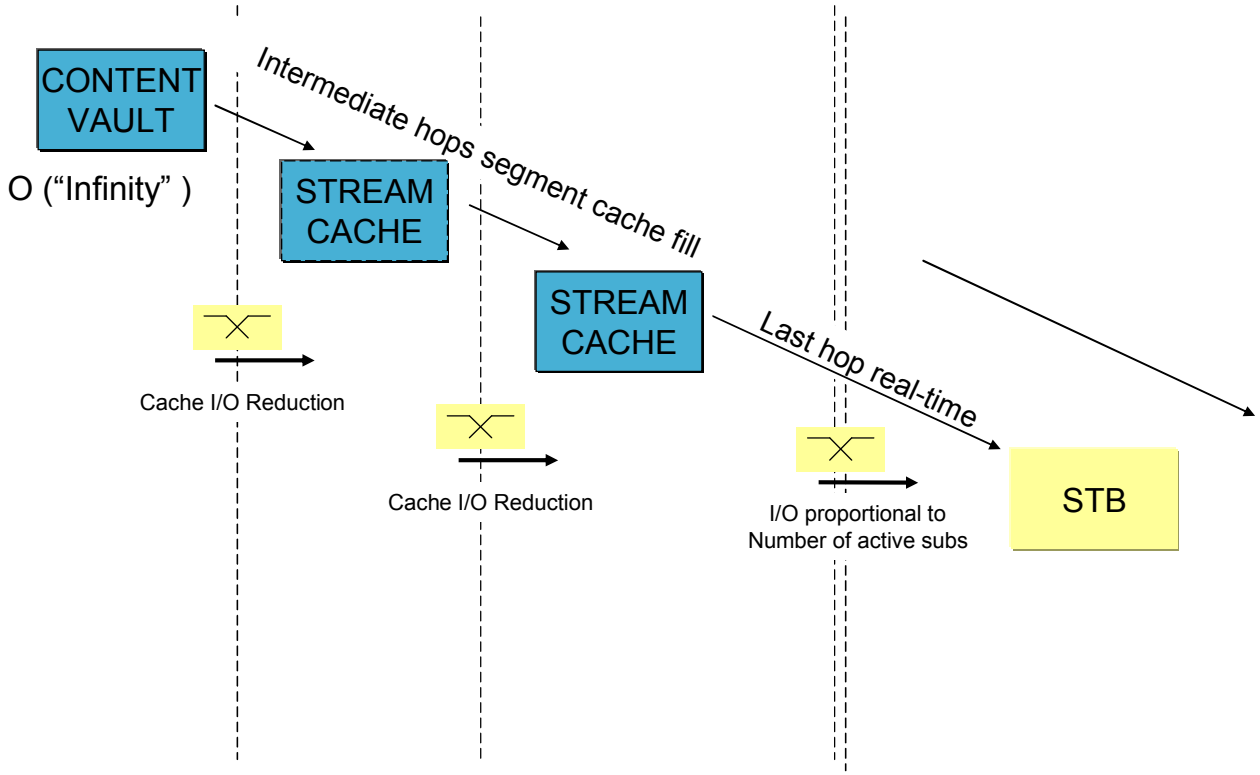

Figure 3 – Real time Caching Hierarchy

At each tier of the hierarchy the latency introduced is minimal – O(n*10ms). Also the ability to transition between consumption modes (1x, n*FF, n*REW, pause) with low latency – O(250ms) is enabled. A key requirement is that, with each of the transition modes, the awareness of frame by frame semantics of the content type is necessary so that the video segments within cache storage can be accurately managed.

Pull Versus Push Distribution

Caching object distribution protocols can be divided into two categories.

The pull category typified by web caching depends on the client that is issuing the requests and intelligently managing the transfer operations. No awareness of server state is communicated to the client. Usually no awareness of the bitrate of the content is communicated either (though it can be communicated by a separate control path).

In the push category, (typified by MPEG transport streams sent over either UDP or RTP transport) the content source is bandwidth aware, and maintains the rate of transfer in order to meet the bandwidth delivery characteristics of the content. The push model avoids significant bidirectional overhead (other than adaptation to network resource constraints) and enables assurance of stream rate for real time content objects.

Either style can be utilized. The pull category requires new mechanisms that assure that the real time caching servers (potentially different servers) being contacted for unpredictable mid-stream distribution operations have ways of learning that the previous operation has been canceled by the downstream cache, and initiate the abort procedure. The push category benefits from this mechanism since communication of client state already exists.

Segmented Object Distribution

Traditional video distribution models exhibit the characteristic that the entire video object must be distributed to the entity that streams the video toward the client. This generates several systemic deficiencies. First is that a significant delay is incurred while awaiting distribution. Second is that the percentage of content consumed is less than 100 percent (especially for long-tail where segments, e.g. famous scenes, are of primary interest). By distributing the entire object the cache is utilized in a non-optimized manner.

Therefore one of the characteristics of TV cache distribution is that object segments are transferred, on demand, if the correlated content segments are not already cached locally. This has the advantage of optimizing bandwidth consumption and cache storage consumption.

Another advantage of segmented object distribution is that it enables new services such as remixing, where arbitrary segments of content can be remixed into a new virtual asset. An example is all the goal shots of the world famous soccer star Pele combined into one segment. This can be achieved without distributing the dozens of entire full-game video objects to real time caching servers. Only the relevant scenes need to be real time cache filled. The object granularity needs to have the ability to identify frame level semantics in all cases of segmented object distribution. Methods for learning and communicating such semantics range from control plane extensions identifying offsets to embedded descriptors highlighted by standards such as TV Anytime [10].

Correlated Object Caching

The functionality of transitioning from 1x content to rewind and fast forward modes of consumption highlights another feature for real time caching that is not present in web caching. This is the capability to transition to whatever object distribution mode is being consumed by the client, assuming that the new object segments from the new mode are not yet cached on the caching entity. It should be noted that this is based on client behaviors driven by operations in the control protocol, e.g., RTSP.

In order to deliver correlated object caching, some structure, such as an indexing database, needs to be conveyed between the authoritative source and downstream caching entities so that all cache entities have accurate awareness of the object content segments contained within the cache storage.

Static-Object Verses Dynamic-Object

Two different types of objects are to be distributed by the caching infrastructure. The first category, here called static-object, pertains to a content item that has been completely ingested into the authoritative source prior to distribution towards the streaming server client. This is typically called VoD, but is not constrained to VoD objects. In this paradigm all ingest and other processing of the content object is completed prior to initiation of cache distribution. In one use case this object is not identified as available to clients until full ingest is complete.

The second category, here called dynamic-object, is a type of object that is dynamically created and ingested by the authoritative source, and is concurrently distributed into the caching infrastructure. In this paradigm the authoritative source is concurrently performing processing on the object (e.g., computing trick files, if required) and making the object available for concurrent distribution toward the destination. One well known use case for the dynamic-object paradigm is time-shift of linear content.

It should be noted that dynamic-object types have an impact on functionality of the correlated object caching indexing database, i.e., dynamic updates concurrent with ingest by the authoritative server.

Source & Sink State Synchronization

The web service caching model exhibits a lack of state synchronization between the client and the server. Each side estimates the projected behavior of the other side. Neither side is aware of any average or instantaneous bottlenecks or constraints of the other side. The real time cache fill protocol should support a method of communicating instantaneous load and state change of both source and client.

One example of state synchronization is the awareness of bandwidth. Each source and sink has a finite aggregate I/O bandwidth limit. Examples of such bandwidth constraints are on-board bus bandwidth, bandwidth to associated storage, and bandwidth between memory and adapters. The real time object caching service should exhibit bilateral awareness of I/O constraints of the source and sink so that unnecessarily high latencies or jitter behaviors are not introduced.

Ingest Overrun Avoidance

Each caching node in the distribution path from the authoritative source to the client has finite bandwidth ingest constraints. Content distribution must not overrun the ingest bandwidth with the aggregate maximum number of active session. This implies the ability to maintain tight tolerances on smooth delivery of the stream, and avoidance of unnecessary bandwidth bursts.

Network Bandwidth Optimization

The path from the content source to downstream caches has finite bandwidth constraints. The caching protocol must be designed so as not to induce either packet loss or excessive buffering jitter in the aggregate number of streams being concurrently delivered.

Opportunistic Resource Utilization

The cache fill protocol should be aware of resources of the source and sink, and also be capable of adapting transfer behavior in response to dynamically changing resource behaviors. If for example, the source, sink, and intervening path are lightly loaded from the perspective of resources, then an optimization is to enable content to be transferred at higher rates opportunistically. Such transfers are not directly correlated to the actual play out state of the content with respect to the subscriber.

Elasticity Assurance

The real time cache fill protocol should also optimize management of cache fill buffer elasticity, while maintaining a short maximum latency for stream event transitions. This requires a distribution mode where content is initially transferred at a rate higher than stream rate, and then, after a short time window, settles down to transfer at stream rate. The transition to higher rates occurs at any point that new content is transferred and short streaming startup latency is required. Examples include session start, splice points (different content objects), interactive transitions to new content, and trick mode transitions (also where different content is transferred). The elasticity buffer accommodates reasonably bounded jitter behavior and retransmission of dropped packets without disrupting streaming behavior to the subscriber.

N-Screen Initiative

A system architecture for enabling N-screen delivery is required. Key enabling characteristics include decoupling of the awareness of delivery infrastructure from the

application layer, and embedding all distribution and resource management into the delivery infrastructure.

Because the model is (a) inherently a "cache-on-demand" model, (b) separates delivery from metadata, and (c) enables "real time ingest" from external storage, sparsely populated media consumption formats (format transcoding) may be generated on-demand, or opportunistically. As with the demands on the central storage systems themselves, this load scales only with the unique assets being consumed, not with the number of streams being watched.

The characteristics of the application layer are expected to be like web services in nature. Figure 4 shows a sample configuration. Subscriber interfaces will be provided for navigation, business logic (purchases, rentals), service configuration, entitlements, etc. Each device type will have control and transport interfaces that are specific to the device, but which are not seen by the application layer.

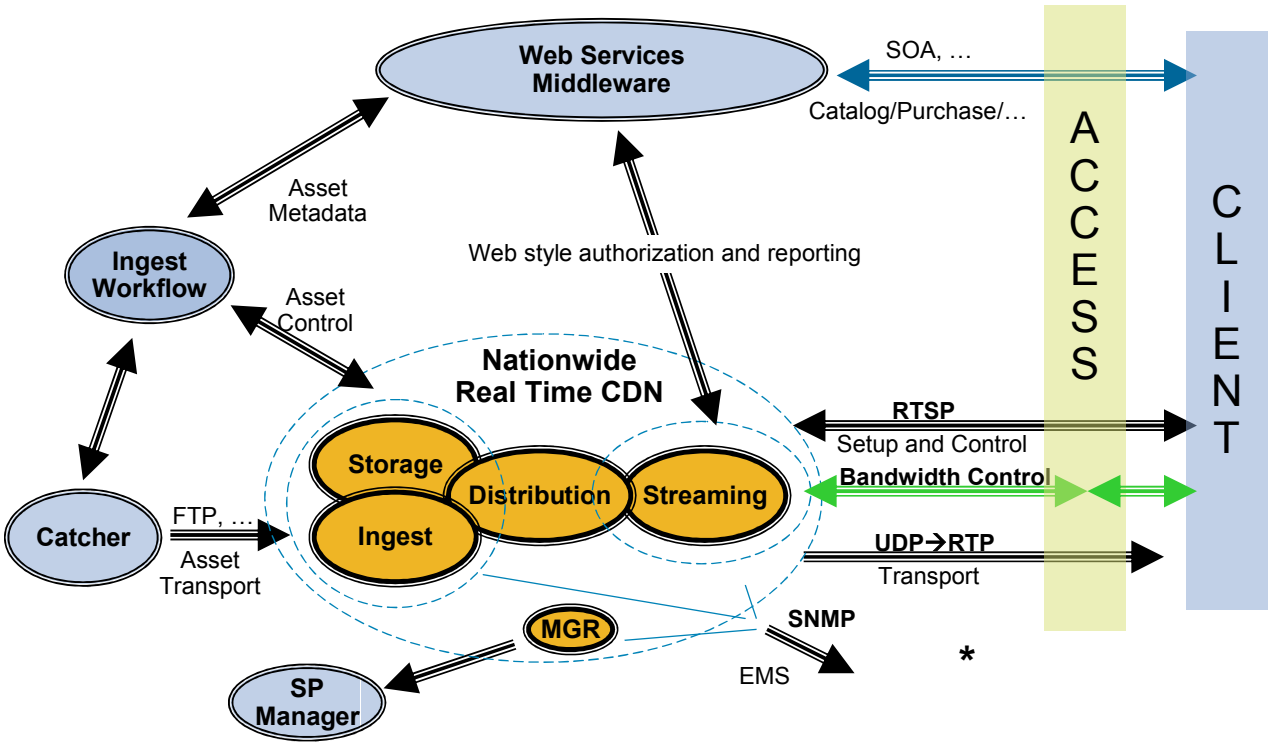

Figure 4 – "n-screen" Convergence

Real Time Streaming Protocol (RTSP) is the dominant control plane signaling for session management. Capabilities discovery and on-path session resource management will be utilized to identify the appropriate version (encoding resolution, bit rate, codec type) of an asset to stream to the device, and is appropriate for more advanced media control. Systems should also expect to provide simplified HTTP semantics (with potentially degraded performance and feature characteristics) to enable the broadest class of device consumption.

The Web services style of interfaces can be defined to allow all streaming infrastructures to consult the authoritative business logic of the application layer. This logic should be authorization oriented, not authentication oriented, as the criteria for playback may be user-centric (commercial entitlements or user

sharing permissions) or publisher-centric (rights management or web availability).

CONCLUSION

This paper identifies the rapid shift in user consumption behavior from traditional consume-on-broadcast-timeline (on TV only) to consume-on-subscriber-timeline with the ability to pause, rewind, and fast forward content (on any device). Infrastructure convergence toward all-IP, wideband edge network transport, and unicast enabled real time cache distribution paradigms are highlighted.

The benefit of having authoritative sources for the content (permanent library storage), as exhibited in the web object distribution model, plus the insertion of real time enabled caching servers in the path between the authoritative source and the destination client, enables the service to be scaled to an unlimited number of consumers, consuming an unlimited library of content (both on-demand and time-shifted live), while preserving the user expectation of DVR-like consumption delivered by the network.

Also highlighted is the decoupling of the application layer from the real time content delivery layer. No specific protocols are detailed in this paper. The primary focus is establishing a framework for scaling to the world of infinite content and infinite number of subscribers.

References

[1] Real Time Video Services & Distributed Architectures: Irreconcilable Differences or a Marriage Made in Heaven, John R. Pickens, SCTE 2006

[2] Netflix 2007 press release http://www.netflix.com/MediaCenter?id=5384

[3] Netflix Current selection - http://www.netflix.com/BrowseSelection

[4] Comcast Project Infinity press release 2008 CES - http://www.comcast.com/ces/infinity_hd.aspx?section=hd

[5] An Open Architecture for Switched Digital Services in HFC Networks, Luis Rovira, Lorenzo Bombelli, SCTE 2006 Conference on Emerging Technologies.

[6] Web Cache Wiki article - http://en.wikipedia.org/wiki/Web_cache

[7] A Survey of Web Caching Schemes for the Internet, Jia Wang, ACM SIGCOMM Computer Communication Review, Volume 29 , Issue 5 (October 1999)

[8] ZIPF Wiki article - http://en.wikipedia.org/wiki/Zipf's_law

[9] *VOD Servers - Equations and Solutions*, Glen Hardin, W. Paul Sherer, NCTA 2005

[10] Metadata - the role of the TV-Anytime specification, Morecraft, C. Storage and Home Networks Seminar, 2004. The IEEE, Volume , Issue , 3 Nov. 2004.