EXPLORING THE IMPACT OF MIXING DELAY SENSITIVE AND DELAY INSENSITIVE TRAFFIC ON DOCSIS NETWORKS

Stuart Lipoff
IP Action Partners

Abstract

Results are presented, and discussed, of discrete event simulation and mathematical modeling exploring the quality of service (QoS) cross impacts of mixing various types of delay insensitive (e.g. web surfing) and delay sensitive (e.g. VoIP telephony) on a DOCSIS network. Of particular interest is the impact on over-the-top streaming traffic when the operator asserts QoS control over streaming services offered and managed by the cable operator.

INTRODUCTION

Historically most DOCSIS cableplant IP traffic was largely downstream intensive and undifferentiated– mainly consisting of delay insensitive web surfing and file downloads. However in today's situation we already have new applications (e.g. peer-to-peer) vastly increasing traffic intensity. As we look toward a future traffic mix of delay sensitive carrier managed services combined with delay sensitive 3rd party over-the-top services; there is the possibility of negative impacts on the user experience resulting from the new traffic mix.

The DOCSIS protocol was designed from the start to be future proof. By employing node splitting and/or exercising protocol options, the operators can scale the service to support increased traffic. DOCSIS also has hooks that allow highly flexible management of the shared up and down stream bandwidth to support service level guaranties, quality of service (QoS), and the overall user experience. Under the assumption that cable operators (MSO) act rationality, they will take advantage of the future proof capabilities of DOCSIS to add capacity and activate features (e.g. QoS) as the operator rolls out new services (e.g. voice over IP telephony) that require QoS and/or increased bandwidth. The flip side of a rational business decision however, is that an MSO would not be motivated to spend capital to add extra capacity to their cablemodem network simply to support a 3rd party service offering; especially when such a third party offering may require special treatment or this third party offering consumes an excessive share of the limited up and down stream system bandwidth without additional compensation to the MSO.

This study was designed to understand which types of services can gracefully coexist on the same plant versus those services that consume plant resources out of proportion to their economic value. In order to develop these insights, mathematical and discrete event modeling was employed to simulate the impact of adding various example services to a system that was rationally engineered to support a baseline web surfing traffic + MSO provided telephony load. The analysis explores the impacts on the user experience as a function of increasing traffic by adding new delay sensitive traffic to the legacy delay insensitive system designed for web surfing. In a similar fashion, the cross impact is studied between delay sensitive traffic provided by the MSO and delay sensitive traffic created by the subscriber adopting over-the-top delay sensitive services without the knowledge and support of the MSO.

Examples of the delay sensitive traffic include: streaming music, streaming video, voice telephony, and multiplayer games. Of particular interest are the current over-the-top third party service offerings of public switched telephone network (PSTN) voice telephony. The 3rd party providers of such services represent them to be comparable to legacy circuit switched PSTN services offered by local exchange carriers (LECs). The reason for my particular focus upon over-the-top PSTN voice telephony is the following:

- MSOs are now in the process of rolling out Packet Cable managed PSTN voice telephony services that are expected to consume significant up and downstream DOCSIS resource beyond today's legacy situation.

- Since the traffic and user expectations for PSTN voice telephony is well known, it is possible to understand the impacts of traffic mix changes on PSTN voice telephony using objective criteria that can be modeled and analyzed

- An MSO provided PSTN voice telephony service can take advantage of DOCSIS features to manage QoS and reduce the management message overhead of up and down stream resource allocation by means of Packet Cable features that closely integrate the PSTN packet switch with the CMTS. For example, RFC2748-COPS messages from the call management server to the CMTS to allocate Unsolicited Grant Service.

- An over-the-top PSTN voice telephony service can not take advantage of the MSO's efficient resource allocation mechanisms and instead must appear to the CMTS in the same way that other best effort traffic (e.g. web surfing) appears to the CMTS. Given the special needs of streaming delay insensitive voice telephony in both the up and down stream directions, it would suggest careful study to determine if a cable system designed for best effort delay insensitive web surfing would provide a satisfactory user experience for an over-the-top delay sensitive service offering.

An important question of interest to the cable industry MSO and technology providers is: what absolute capacity limits can a particular system provide for a particular mix of traffic. Because DOCSIS offers such a wide variety of configuration options taken together with HFC network partitioning variations, the infinite possibilities of service mix, subscriber adoption rates, and traffic intensity; the absolute capacity limits of DOCSIS is not a topic I address in this study. Instead the focus of this study is upon the shape of the operating characteristic that relates offered traffic to the user experience performance. To explain the meaning of shape in the context of this study, I refer to figure 1 below. Ideally you would like a communications system to have a type 1 linear operating characteristic where there is a linear relationship between the offered traffic and the system performance as experienced by the user. The system performance parameter(s) will vary somewhat depending on the type of traffic. For web surfing, peak down load speeds would be the primary system performance parameter.

For PSTN voice, call attempt blocking and packet loss during the call would be among the important performance parameters. For most multiplayer games, latency time would be the primary performance parameter of interest. Although a linear operating characteristic is ideal, most real multiple access systems, such as DOCSIS, have curves that look non-linear like type 2 where there are limits and thresholds. The objective of this study is to understand where the limits kick in and how steep is the curve's slope as the offered traffic mix is increased beyond the threshold.

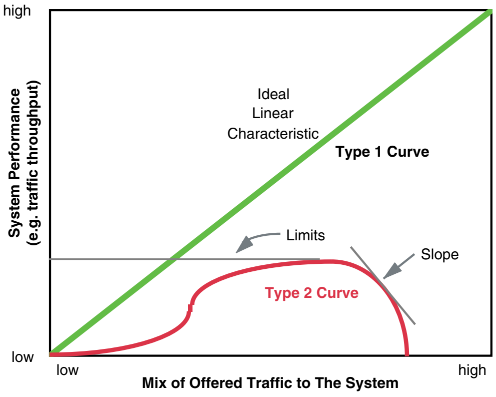
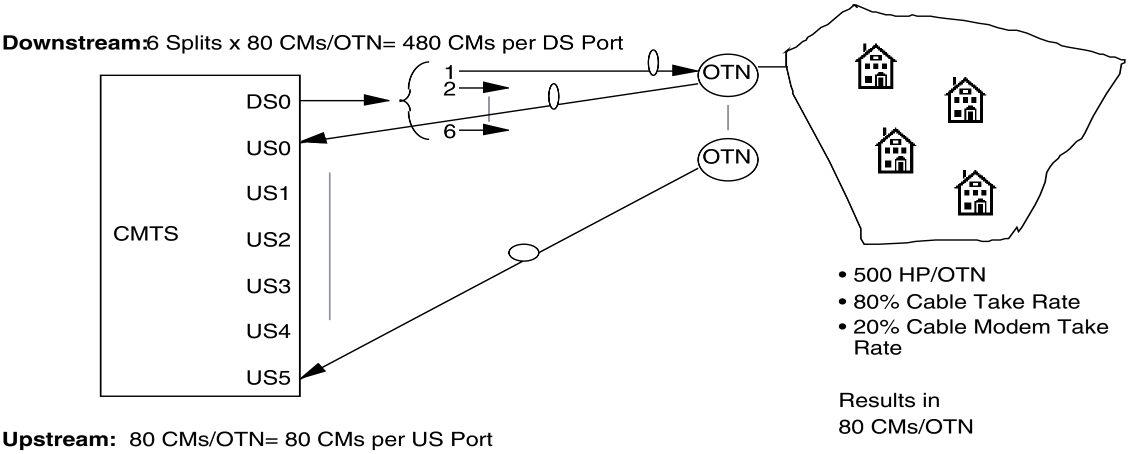


Figure 1. Illustrative Operating Characteristics

BASELINE DOCSIS SYSTEM ASSUMPTIONS

In parallel with rapidly growing consumer adoption of broadband cablemodem services we also have rapid growth in bandwidth hungry multimedia traffic. With that in mind, picking a single point system configuration as a baseline that is representative of all present and future cable systems is not possible. Non-the-less with so many possible variables, modeling requires some assumptions be made and a subset be held constant while other subsets are varied to develop insights. Since the stated objective of this study is not to develop a perspective on absolute traffic capacity of DOCSIS in general, but rather to understand the figure 1 nature of the shape of the operating characteristics, the choice of a particular baseline is somewhat less important as long at the assumptions are reasonable. For all the models analyzed in this study, the network layout shown in figure 2 is assumed. In the figure 2 system each optical transition node (OTN) has 500 homes passed (HP) with an 80% cable take rate of which 20% of the cable subscribers take DOCSIS cablemodem (CM) service resulting in 80 CMs/OTN. At the CMTS, each downstream port of the CMTS splits to serve 6 OTNs while there is a 1:1 mapping of one OTNs into one CMTS upstream port. The resulting load per CMTS port is 480 CMs/Port downstream and 80 CMs/Port upstream. As the source note in figure 2 indicates, these parameters are based on assumptions in a 2004 Cisco publication. However, in analyzing the Cisco assumptions in the light of today's typical traffic loads, for the purpose of this study I split the traffic on each CMTS by two times over Cisco's assumptions.



Downstream:6 Splits x 80 CMs/OTN= 480 CMs per DS Port
CMTS
DS0
US0
US1
US2
US3
US4
US5
1
2
6
OTN
OTN
• 500 HP/OTN
• 80% Cable Take Rate
• 20% Cable Modem Take Rate
Results in
80 CMs/OTN
Upstream: 80 CMs/OTN= 80 CMs per US Port

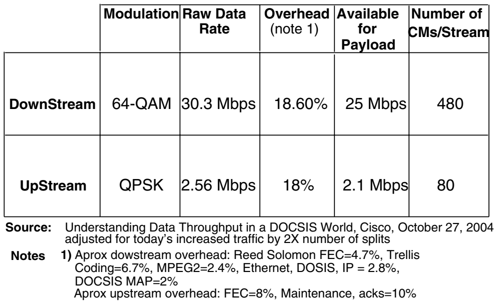
Source: Understanding Data Throughput in a DOCSIS World, Cisco, October 27, 2004 adjusted for today's increased traffic by 2X number of splits

Figure 2. DOCSIS Over HFC Network Layout Assumed for Modeling in this Study

This ability to split ports at the CMTS or HP/OTN demonstrate the ability of an HFC based DOCSIS system to scale to meet traffic needs. However a key assumption made in this study methodology is that MSOs acting rationally will invest in adding capacity to their system when there is an economic justification to do so. However when third party service providers promote over-the-top services that might consume significant transmission resources, no rational MSO should make an uneconomic investment in adding capacity to support such over-the-top services.

The figure 2 network layout is the first of three categories of basic system assumptions I employed in the modeling and analysis. Figure 3 shows the assumptions made regarding raw and useful payload data rates on each CMTS port.

In any actual DOCSIS system down and up stream raw data rates can be set and the overhead can vary somewhat, but for the purpose of modeling, figure 3 payload parameters will be assumed as available for the purpose of carrying the various mixes of traffic studied.

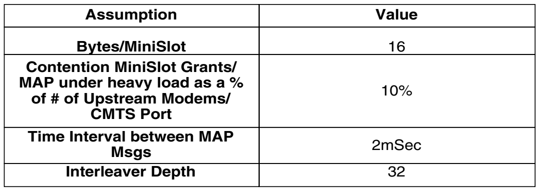


	Modulation	Raw Data Rate	Overhead (note 1)	Available for Payload	Number of CMs/Stream
DownStream	64-QAM	30.3 Mbps	18.60%	25 Mbps	480
UpStream	QPSK	2.56 Mbps	18%	2.1 Mbps	80

Source: Understanding Data Throughput in a DOCSIS World, Cisco, October 27, 2004 adjusted for today's increased traffic by 2X number of splits

Notes **1)** Aprox dowstream overhead: Reed Solomon FEC=4.7%, Trellis Coding=6.7%, MPEG2=2.4%, Ethernet, DOSIS, IP = 2.8%, DOCSIS MAP=2%
Aprox upstream overhead: FEC=8%, Maintenance, acks=10%

Figure 3. Data Rate Settings for Downstream and Upstream CMTS Ports



Assumption	Value
Bytes/MiniSlot	16
Contention MiniSlot Grants/ MAP under heavy load as a % of # of Upstream Modems/ CMTS Port	10%
Time Interval between MAP Msgs	2mSec
Interleaver Depth	32

Figure 4. CMTS Parameters That Impact Modeling and Limit Transmission Rates in this Study

Figure 4 contains CMTS parameters that are the third category of assumptions that are needed for the modeling I performed. None of the figure 4 parameters are fixed by DOCSIS but instead can be set by the operator and dynamically adjusted, if desired, by scheduling and resource management software running on the CMTS. In the simulation and modeling in this study, I picked the parameters in figure 4 to represent numbers that one might find during a period of high traffic intensity. Using these parameters as fixed in the simulations is justified since I am only interested in the behavior of the system when it is heavily loaded, so the operating characteristic the models predict should be valid during the heavily loaded system time of interest. Each CMTS supplier is free under DOCSIS to develop proprietary scheduling algorithms that attempt to maximize the user experience as a function of the traffic. During period of light traffic load, such a scheduling algorithm will adjust the ratio of contention to reservation minislots such that there are more contention slots available during light traffic loads to minimize latency in servicing best efforts contention upstream bandwidth requests, but during periods of heavy traffic loads the scheduler will assign more of the upstream transmission minislots to carry reserved payloads rather than be used for upstream bandwidth requests. The number of contention minislots for each MAP interval in this high traffic load modeling is set to 10% of the number of upstream modems per CMTS so as to balance efficient use of the upstream bandwidth but still leave a reasonable number of contention slots so that idle CMs can have opportunities to request bandwidth. A typical 2mSec is used in the modeling for the length of each MAP interval. This 2mSec inter MAP internal taken together with the 32 interleaver depth results in any one cablemodem needing to wait for every other MAP interval (i.e. 4mSec) to request bandwidth if the first contention request is not received. A interleaver depth of greater than 32 would increase latency and could require a cablemodem having to wait three or more MAP intervals before making a second request for bandwidth.

OFFERED TRAFFIC ASSUMPTIONS, QUALITY OF SERVICE, AND USER EXPERIENCE EXPECTATIONS

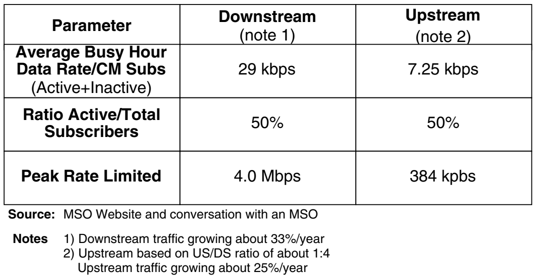
This section identifies the possible mixes of traffic on the DOCSIS system, the likely traffic intensity, and the different quality of service criteria that would impact the user experience. It is important to note that each traffic type will generally have very different QoS criteria and very different thresholds for acceptable versus unacceptable user experiences. For example some traffic types (e.g. web surfing) will maximize the user experience by providing very fast peak downstream data rates while the user will be generally insensitive to brief interruptions in the data stream. Multiplayer, *shoot first* games will care most about latency while at the other extreme public switched telephone network voice will care about the probability of blocking, latency, and packet loss.

Websurfing and Other Best Effort Data Traffic

The bulk of today's operational DOCSIS systems have been engineered for Websurfing traffic and that is the first traffic stream considered in this study. Although many general characteristics of this traffic are the same today as when systems were first launched, it is also the case that the magnitude of this traffic is rapidly growing. The general characteristics of websurfing traffic include:

- It is downstream intensive resulting in an asymmetrical traffic load between the down and up stream paths

- It is delay insensitive in that brief interruptions of the packet stream or short delays in serving upstream best-efforts contention requests for bandwidth will not be noticeable to the end user



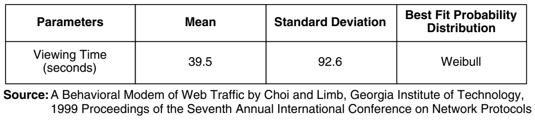
Parameter	**Downstream** (note 1)	**Upstream** (note 2)
Average Busy Hour Data Rate/CM Subs (Active+Inactive)	29 kbps	7.25 kbps
Ratio Active/Total Subscribers	50%	50%
Peak Rate Limited	4.0 Mbps	384 kpbs

Source: MSO Website and conversation with an MSO

Notes 1) Downstream traffic growing about 33%/year
2) Upstream based on US/DS ratio of about 1:4
Upstream traffic growing about 25%/year

Figure 5. Websurfing Average Traffic Parameters

Figure 5 shows the assumptions employed in the modeling for this baseline web surfing load. Of particular note in figure 5 is that traffic is growing between 25% to 33% per year and that 50% of subscribers are assuming to be actively surfing during any measurement interval. The MSO interviewed indicated there is no well defined busy hour but instead this is an afternoon to early evening busy period of 4-6 hours in duration. With this in mind, the modeling assumes that the voice telephony busy hour will fall somewhere within the web surfing busy period assuring a 100% overlap in traffic. The rapid growth rate of web surfing traffic also implies a regular need to scale the system to accommodate growth. The average data rate shown in figure 5 was reported to me by an MSO based on actual system measurements. The inter-arrival time between web pages is given by the viewing time of a web page before the next page is requested and that parameter is shown in figure 6. Of note in figure 6 is a page viewing time of about 40 seconds but notice the very large standard deviation suggesting there is considerable variation about the 40 second average. The impact of this 40 second time for websurfing is that on the average of every 40 seconds, a subscriber in best efforts mode will need access to a contention minislot for the purpose of requesting bandwidth. As will be shown, this 40 second time is a comfortable many orders of magnitude larger than a much shorter interval for best effort telephony traffic.



Parameters	Mean	Standard Deviation	Best Fit Probability Distribution
Viewing Time (seconds)	39.5	92.6	Weibull

Source: A Behavioral Modem of Web Traffic by Choi and Limb, Georgia Institute of Technology, 1999 Proceedings of the Seventh Annual International Conference on Network Protocols

Figure 6. Websurfing Traffic Interarrival Times

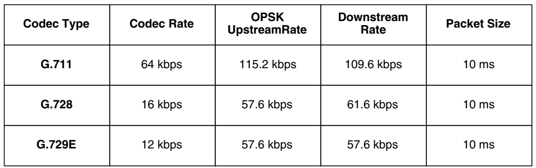
As noted in the introduction to this section, the main performance measure impacting the websurfing user experience will be the peak limited downstream data rate. Of secondary interest to websurfing would be latency associated with sending upstream page requests that exceed on the order of a second. For email and file up and down loads, the peak data rates in both the up and down stream direction would most impact the user experience. Unlike websurfing however, latency would generally be unimportant. For user consumption of real time audio, video, or multimedia streaming downstream average data rates of less than 100 kbps are typical so the peak data rate is a less important factor; however, for this streaming content the key user experience impacting parameter is an interruption of the downstream longer than the buffer size (settable for most players generally in the range of 10-30 seconds).

Multiplayer games are an emerging important category of best efforts traffic. It was reported that one game, CounterStrike, was the third largest generator of UDP traffic on the internet behind only DNS and RealAudio traffic[1]. There are many variations of multiplayer online games including: peer-to-peer and client server variants. CounterStrike is a client server game and as the reference paper notes, the players consume an average of 40 kbps during the sessions that can last up to about 2 hours in duration. CounterStrike is in the popular category of a *shoot first* game which means that the key user experience impacting parameter will be response time. Packets are typically small on the order of 50mSec indicating that response times should be on the same order. In the case of a DOCSIS system, the mechanisms that could likely control response time would a lower bound on the interleaver depth and propagation delays to an upper bound set by congestion in upstream contention minislots.

Although best efforts data traffic on DOCSIS networks is more than just web traffic, since there were no measuring tools available to separate this best efforts traffic into categories such as web page downloads, email, file uploads, over-the-top multimedia, etc; for modeling purposes the traffic is assumed to be 100% web surfing in terms of average data rate. Also for accessing the user experience acceptable threshold, only latency delays associated with web page requests at inter-arrival time is considered further in this study.

PSTN Voice

As will be shown, the findings of interest in this study relate to interactions between best efforts websurfing discussed above with PSTN voice traffic. The PSTN voice traffic of interest is of two types: 1) QoS managed voice traffic that is provided by MSOs and 2) Best efforts voice traffic provided by third party service providers on an over-the-top basis. Assuming that the target user experience is the same for type 1) and 2) PSTN voice services, the payload data rates and user experience acceptability criteria will be the same. Figure 7 shows a variety of waveform and low bit rate source codecs and their raw data rates along with the resulting up and down stream payload data rates on the network.



Codec Type	Codec Rate	OPSK UpstreamRate	Downstream Rate	Packet Size
G.711	64 kbps	115.2 kbps	109.6 kbps	10 ms
G.728	16 kbps	57.6 kbps	61.6 kbps	10 ms
G.729E	12 kbps	57.6 kbps	57.6 kbps	10 ms

Source: White Paper: Engineering CMTS and HFC for VoIP with Capital and Operating Expense in Mind, Strater & Nikola, Motorola Broadband Communications Sector, December 2004

Figure 7. PSTN Voice Data Rates

Also shown is a typical number used in the modeling for the packet size. Longer packet sizes can provide more efficient use of the bandwidth resource but add latency and therefore numbers on the order of 10mSec are assumed as an acceptable tradeoff. In recent years the voice quality of low bit rate codecs has improved to the point that they are more than acceptable for voice only services. However, since the PSTN voice services of interest are those comparable and fully competitive with those offered by wireline local exchange carriers (LECs), for this study I assume the G.711 waveform codec which would assure quality comparable to an end office last mile wireline circuit and be compatible with fax and analog data modem traffic. As figure 7 shows, this implies payload rates on the DOCSIS system of 115.2 kbps upstream and 109.6 kbps downstream with 10mSec packet sizes.

While type 1) and 2) PSTN voice traffic have different QoS management and resource allocation mechanisms, for a comparable user experience they share the same underlying mechanisms that impact the user experience. Under the constraint of the same maximum acceptable round trip path delay of 300mSec, the user experience criteria fall into two categories[2]:

- Blocking of call attempts

  —This is typically managed by admission control in which a maximum number of voice circuits are reserved in order to keep the probability of blocking by a subscriber to <1.0%

  —For mathematical modeling purposes the Erlang B criteria was employed to compute blocking probability

  —Voice activity detection can be employed to improve system efficiency but was not assumed in this modeling

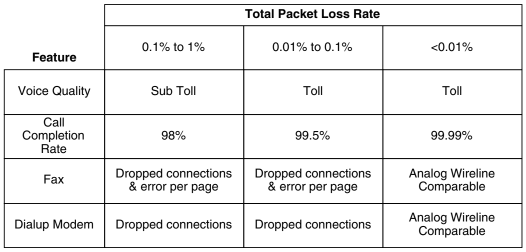
- Packet loss following call establishment

  —Once a call is established, excessive loss of packets will impair the channel. In the case of voice the conversation is interrupted and for fax and analog modem data, bits are lost

  —Packet loss can be compensated for by buffering and retransmission but this adds latency impacting path delay

  —Toll quality services typically aim to keep round trip path delay under 300mSec (150mSec end-to-end) and this was assumed as limit in the modeling

Figure 8 reveals the impact of packet loss rate on the user experience. Consistent with the <1% blocking criteria for a service offering comparable to a wireline LEC, I have chosen a maximum tolerable packet loss rate of <0.01% in order to provide toll quality voice while maintaining fax and analog data modem compatibility.



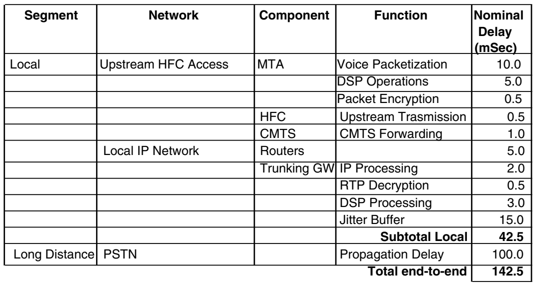
Feature	Total Packet Loss Rate		
	0.1% to 1%	0.01% to 0.1%	<0.01%
Voice Quality	Sub Toll	Toll	Toll
Call Completion Rate	98%	99.5%	99.99%
Fax	Dropped connections & error per page	Dropped connections & error per page	Analog Wireline Comparable
Dialup Modem	Dropped connections	Dropped connections	Analog Wireline Comparable

Source: White Paper: Engineering CMTS and HFC for VoIP with Capital and Operating Expense in Mind, Strater & Nikola, Motorola Broadband Communications Sector, December 2004

Notes Jitter is another important factor that could impact QoS in addition to packet loss rate

Figure 8. Impact of Packet Loss Rate on PSTN Voice End User Experience

In building the simulation model for the impact of packet loss on best effort over-the-top services, I also needed to understand the maximum contribution to latency from the uncertainty of timely grants of payload packet transmission requests in a contention minislot. Note that this is not an issue with respect to QoS managed MSO provided services since, the assumption is that the MSO provided services operate on an Unsolicited Grant Service basis where once the call is setup, the bandwidth is granted to the voice call for the entire duration without having to resort to using contention minislots for a bandwidth request. Figure 9 below shows a typical allocation of latency in a VoIP network by the contribution of each network segment. If the end-to-end delay limit is set to 105mSec, then the maximum allowable delay due to grant uncertainty would be 105-142.5=7.5mSec. As noted earlier, for the choice of interleaver depth and 2mSec MAP interval, a particular CM would be able to request bandwidth on a contention minislot only every other MAP interval resulting in 4mSec between requests. This in turn implies that if the delay contribution from grant uncertainty is held to 7.5mSec, then no more than two contention requests can fail before a packet will be lost.



Segment	Network	Component	Function	Nominal Delay (mSec)
Local	Upstream HFC Access	MTA	Voice Packetization	10.0
			DSP Operations	5.0
			Packet Encryption	0.5
		HFC	Upstream Trasmission	0.5
		CMTS	CMTS Forwarding	1.0
	Local IP Network	Routers		5.0
		Trunking GW	IP Processing	2.0
			RTP Decryption	0.5
			DSP Processing	3.0
			Jitter Buffer	15.0
			Subtotal Local	**42.5**
Long Distance	PSTN		Propagation Delay	100.0
			Total end-to-end	**142.5**

Source: White Paper: Engineering CMTS and HFC for VoIP with Capital and Operating Expense in Mind, Strater & Nikola, Motorola Broadband Communications Sector, December 2004

Figure 9. Typical Contribution to VoIP Path Delay by Function and Network Segment

MODELING AND ANALYSIS FINDINGS

Included within the scope of this study is two separate analysis which share in common the impacts of other traffic on best-efforts over-the-top PSTN voice services. As indicated above, if the goal is to provide an over-the-top PSTN voice service that is comparable to wireline LEC and MSO provided QoS managed VoIP, then the over-the-top service must not exceed the following two limits:

- The probability of blocking during a busy hour call attempt must be <1.0%

- Under a constraint of a maximum 150mSec end-to-end delay (e.g. 7.5mSec grant uncertainty contribution) the packet loss during a call must be <0.01%

Blocking Analysis

The blocking analysis was performed using a mathematical model based on Erlang B traffic theory. This theory is based on the statistical probability of finding available bandwidth (for DOCSIS reservation minislots) available given a finite number of voice circuits (i.e. those minislots assigned to a user for the call), a finite number of subscribers, and information on traffic intensity (e.g. Erlangs) during the busy hour of interest. The Erlang B theory is an approximation since it assumes infinite sources, Poisson arrivals, exponential holding times, and blocked calls cleared. The mathematics of this theory outputs the number of voice circuits required to maintain the level of blocking probability under 1% given the random arrivals of subscriber call attempts.

The Erlang B theory is employed in study as follows:

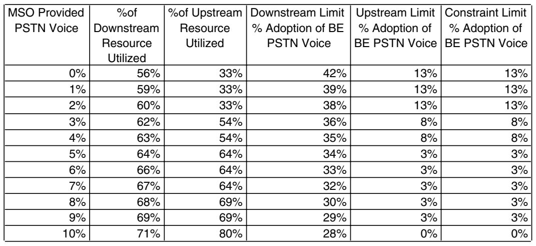
- Unlike websurfing that has an asymmetrical traffic load, voice telephony has a nearly symmetrical up/down stream data rate; therefore the upstream load of 80 CMs/ CMTS port is used as the tight constraint

- The baseline load of figure 5 best efforts websurfing and other data traffic is assumed as being consumed and therefore unavailable to PSTN voice telephony users

- A busy hour traffic intensity of 0.1 Erlangs/ subscriber was assumed[3] for computation of The Erlang B blocking at the 1% QoS

- In examining the impact of mixing type 1) MSO provided QoS managed telephony with type 2) third party provided over-the-top best-efforts telephony; the modeling assumes the QoS management system will give priority to MSO telephony. The resulting effect is that the combination of best efforts websurfing plus MSO QoS managed telephony traffic will consume upstream resources that will not be available for the over-the-top best efforts telephony

Employing the points above, what was studied was the impact of type 1) MSO telephony upon type 2) 3rd party telephony to see if they can both co-exist on a system that a rational MSO has engineered to be of a size that supports MSO provided services. Figure 10 provides the results of this analysis.



MSO Provided PSTN Voice	%of Downstream Resource Utilized	%of Upstream Resource Utilized	Downstream Limit % Adoption of BE PSTN Voice	Upstream Limit % Adoption of BE PSTN Voice	Constraint Limit % Adoption of BE PSTN Voice
0%	56%	33%	42%	13%	13%
1%	59%	33%	39%	13%	13%
2%	60%	33%	38%	13%	13%
3%	62%	54%	36%	8%	8%
4%	63%	54%	35%	8%	8%
5%	64%	64%	34%	3%	3%
6%	66%	64%	33%	3%	3%
7%	67%	64%	32%	3%	3%
8%	68%	69%	30%	3%	3%
9%	69%	69%	29%	3%	3%
10%	71%	80%	28%	0%	0%

Figure 10. DOCSIS Capacity for Other-the-Top versus MSO Provided PSTN Voice at 1% Blocking

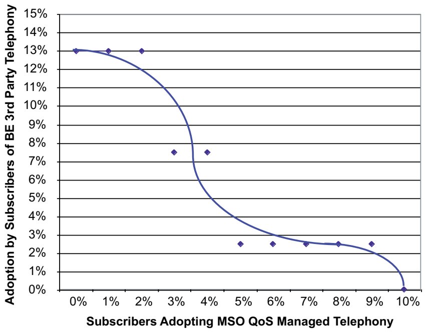


Figure 11. Graph of DOCSIS Capacity for Other-the-Top versus MSO Provided PSTN Voice at 1% Blocking

The left most column in figure 10 shows the percent adoption by the 80 subscribers on the CMTS port of MSO provided PSTN voice services. Notice that at zero percent adoption, 33% of the upstream resource is already utilized by just best efforts websurfing and other data traffic. At this starting level of no MSO provided PSTN voice service, the system will accommodate up to 13% adoption of over-the-top PSTN voice (13% computes to 10 subscribers on this 80 CM/CMTS port system). As the number of MSO subscribers grows from zero the assumption is that the CTMS and Packet Cable QoS mechanisms will give priority to the MSO traffic and the number of voice circuits available for the 3rd party over-the-top services will need to be limited. Figure 10 shows that as the MSO provided PSTN voice adoption grows to 10% (10% computes to 8 subscribers), there are not enough voice circuits available to support a 1% blocking rate for the over-the-top PSTN voice service provider. Figure 11 shows the shape of the operating characteristic by fitting a curve to the figure 10 data. Note the steep slope of the curve at the threshold as the MSO provided service adoption approaches 9%.

Lost Packet Analysis Theory

The blocking analysis above is but one of two impairments considered in this study. The second analysis examined the impact of lost packets in over-the-top PSTN voice services from 3rd party providers. Unlike the QoS managed PSTN voice services from MSO providers, the mechanism for requesting and receiving bandwidth grants for over-the-top providers is assumed to be best-efforts, as available, bandwidth. Unlike the previous blocking analysis which examined the limited up and down stream reservation minislots carrying payload traffic, this lost packet analysis looks instead at the bandwidth allocation process as the primary mechanism leading to lost payload packets.

For this analysis, the assumption is that the DOCSIS network is heavily loaded and the minimum number of contention minislots possible is allocated to maximize payload carrying reservation minislots. A well designed CMTS scheduling algorithm would reduce the actual number of contention minislots to the lowest level suitable to support the normal best-efforts data (mostly websurfing) traffic.

I further assume that MSO provided voice telephony for subscriber originated calls occurs at the very low rate of 0.75 call attempts during the busy hour.[4] As figure 6 shows, the mean time between best-efforts requests for bandwidth on contention minislots for websurfing is on the order of 40 seconds. On the other hand, figure 7 shows the packet size for PSTN voice is 10mSec. Assuming that this over-the-top 3rd party PSTN voice is treated just like other best-efforts data traffic such as websurfing it will require timely access to a contention minislot on the order of every 10mSec to avoid packet loss. While lost bandwidth requests made on a contention minislot can be repeated there is a limit on acceptable latency that would imply a finite buffer size so that only a limited number of bandwidth request repeats are possible prior to losing a packet. It is further assumed that no attempt will be made to give over-the-top best efforts PSTN voice any priority mechanisms (e.g. will disable piggyback requests in request/transmission policy) over less frequent websurfing traffic.

To arrive at an estimate for lost over-the-top PSTN voice lost packets my model reduces to looking at how many of the upstream contention minislot requests for bandwidth are received without corruption from contention for the minislot. Since there is a maximum end-to-end latency limit of 150mSec, there will also be some limit on how many sequential contention minislot bandwidth requests can be corrupted prior to losing the packet. To understand how many requests in sequence can be lost, I examined the total contribution to end-to-end latency as shown in figure 9.

Of the maximum limit of 150mSec end-to-end delay there is a maximum of 7.5mSec available for grant uncertainty delay in the contention minislot mechanism. In this system, using the figure 4 parameters, a missed contention request must wait for every other 2mSec

MAP interval; therefore only two sequential contention minislot grant requests can be corrupted before the packet is lost.

Lost Packet Analysis Model

The Extend™[5] discrete event modeling software package was employed to build a simulation model to determine the traffic intensity that would cause a specific over-the-top best-efforts PSTN voice subscriber under test (SUT) to experience a packet loss in excess of the toll quality limits of 0.01%.

This modeling was performed at a macro level in order to reduce the computational complexity so that alternative parameters could easily be explored. In this context a "macro level" means that there was not an attempt to model the performance of the system at the detailed level of operation of each logical elements and decision process in the real system. Instead a number of lower level logical elements are aggregated into large macro blocks that are designed to provide a good estimate of the actual system.

The analysis was based on the following:

- Only the upstream was modeled assuming 80 cablemodems per OTN per CMTS port
- Minislot size of 16 bytes
- Time interval between MAP messages of 2mSec
- With interleave depth of 32, a subscriber under test (SUT) would have best effort access to every other MAP at 4mSec intervals
- The over-the-top VoIP service would make a best effort (BE) request for each packet

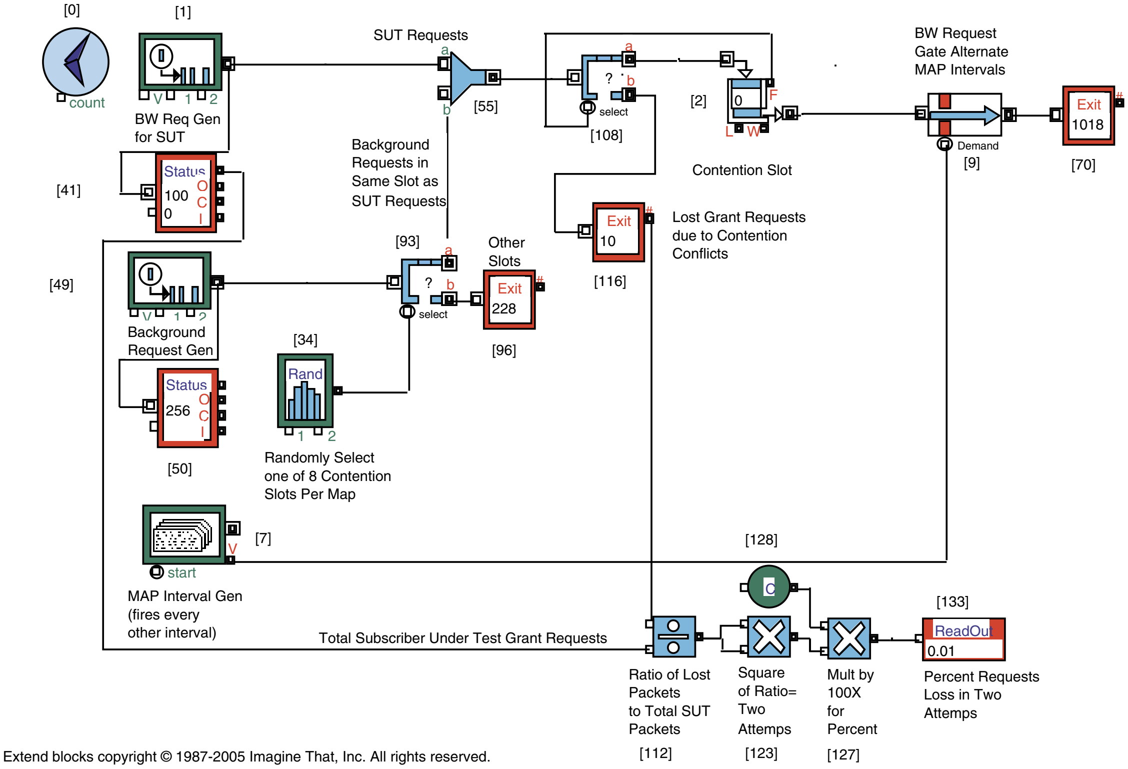

[0]
count
[1]
V
1
2
BW Req Gen for SUT
SUT Requests
a
b
[55]
a
?
b
select
[108]
[2]
0
F
L
W
Contention Slot
BW Request Gate Alternate MAP Intervals
Demand
[9]
Exit
1018
#
[70]
[41]
Status
100
0
O
C
I
Background Requests in Same Slot as SUT Requests
Exit
10
#
[116]
Lost Grant Requests due to Contention Conflicts
[49]
V
1
2
Background Request Gen
[93]
a
?
b
select
Other Slots
Exit
228
#
[96]
[34]
Rand
1
2
Status
256
O
C
I
[50]
Randomly Select one of 8 Contention Slots Per Map
[7]
V
start
MAP Interval Gen (fires every other interval)
[128]
C
Total Subscriber Under Test Grant Requests
[133]
ReadOut
0.01
Ratio of Lost Packets to Total SUT Packets
[112]
Square of Ratio= Two Attemps
[123]
Mult by 100X for Percent
[127]
Percent Requests Loss in Two Attemps

Figure 12. Discrete Event Model to Estimate Lost Packet Rate in Over-the-Top PSTN Voice Service

- The MSO provided VoIP service would be accommodated using unsolicited grant service with 8 minutes between call attempts and would therefore would have minimal impact on traffic in the bandwidth request contention slots

- The busy hour traffic is 0.1 Erlangs with 8 minute average call duration

The model architecture consisted of:

- Examination of a single SUT to determine the impact of other BE traffic upon the packet loss rate to the SUT

- There is contention between one SUT traffic generator and a background traffic generator

- A maximum of two bandwidth request attempts is permitted before the packet is lost in order to maintain low latency

Figure 12 shows a block diagram of the simulation model. This model operates as follows:

- Block [1] represents grant requests every 10mSec by a single subscriber under test (SUT) for bandwidth on a single contention minislot represented by a queue of length one in block [2]. The 10mSec interval is given by the 10mSec packet length and for best-efforts service the bandwidth must be requested after each packet.

- Block [49] represents background contention minislot requests by other over-the-top subscribers who have active calls that overlap with the SUT. The arrival rate varies as will be explained based on the number of other active calls.

- Block [7] generates the timing for 4mSec map intervals and controls the gate block [9]. If there is only one non-corrupted request in the contention minislot, the bandwidth request will get through the gate to the exit block [70]

- Any background requests from block [49] that make their way through block [55] to the contention slot in block [2] will block the SUT request resulting in the SUT request switched by block [108] to the lost requests to block [116] simulation exit.

- Note that not all background bandwidth requests from block [49] make their way to contend with the SUT. Since there are 8 contention minislots per map interval, on an equal random basis, one in eight background grant requests will fall into other slots. The random generator block [34] working with the switch block [93] configures the switch to route only one in eight background grant requests to the same contention slot as the SUT.

- The lost SUT grant requests in block [116] are divided by the total SUT grant requests from block [41] to compute the ratio of lost grant requests to total grant requests. To compute lost packets the loss grant ratio is squared to account for a maximum of two requests by block [123] and converted from a decimal to a percentage by block [127] for readout of lost percent packets in block [133].

All of the parameters necessary to exercise the model have been stated with one exception. The remaining parameter is the block [49] background traffic contention grant request arrival rate. This parameter is in turn controlled by the expected number of simultaneous over-the-top calls that overlap with the SUT. Figure 13 shows the results of estimating these overlaps as a function of the number of over-the-top subscribers on the upstream CMTS port. Recall that this system supports a maximum of 80 CM subscribers, so figure 13 explore the overlaps from 2% to 100% (i.e. 1 to 80 subs). By employing the previously stated assumptions of 0.1Erlangs/sub during the busy hour, and a mean call length of 8 minutes; a single background subscriber active during the 60 minute busy hour would present a mean 0.11 number of average simultaneous background calls to the SUT. The background grant request rate parameter entered into the model block [49] background traffic generator is shown in the next to last column in of figure 13. This number is computed as twice the 10mSec packet rate divided by the average simultaneous calls. The basis for the factor of two, is that the SUT MAP interval is 4mSec (i.e. every other MAP) but the background traffic will be evenly distributed between every 2mSec MAP interval. Therefore the SUT will never be corrupted during half of the MAP intervals.

Number of BE VoIP Subs	Total Number of Background Calls Placed During the Busy Hour	Average Simultaneous Calls	Calculated Call Arrival Rate per 4mSec Interval	Subscriber Adoption Percentage
1	1	0.11	182	2%
6	4	0.51	39	7%
27	20	1.69	12	33%
40	30	2.92	7	50%
53	40	4.15	5	67%
67	50	5.18	4	83%
80	60	5.83	3	100%

Figure 13. Estimation of Average Number of Simultaneous Over-the-Top PSTN Voice Calls During the Busy Hour

Recall that the background traffic contention grant requests are not only spread randomly between each of the two 2mSec SUT MAP intervals, but they can also be expected to be uniformly distributed between the eight contention minislots during each MAP interval. This means that for the call arrival rate from background traffic to become comparable to the SUT traffic in the contention minislot, the overlapping background subscribers need to number about 16 relative to the SUT.

Lost Packet Analysis Findings

The model was exercised using the figure 13 parameters at four levels of over-the-top subscriber adoption. Since the percent packet loss threshold is a small number, on the order of 0.01%, the model was set to run for ten thousand steps. It appeared to converge at about to three thousand steps to the stable findings as shown in figure 14.

% of 80 Subscribers who adopt over-the-top PSTN Voice	Average Number of Active Calls Overlaping with SUT	Background Traffic Contention Request Rate per MAP in msec	% Probability of Packet Loss for SUT After TwoContention Attempts to Request Bandwidth
7	0.51	39	0.01
33	1.69	12	0.07
50	2.92	7	0.53
100	5.83	3	3.31

Figure 14. Results of Discrete Event Modeling to Estimate Lost Packet Rate in Over-the-Top PSTN Voice Services

The model results suggest that the packet loss rate approaches the figure 8 toll quality limit (i.e. OK for voice but unsuitable for fax and analog modem data) at about 7% adoption of over-the-top PSTN voice. This 7% adoption level corresponds to a mean number of about 0.51 background calls that overlap with the SUT. Just above 50% adoption by over-the-top subscribers, the packet loss rate shown in figure 14 approaches the 1% level of noticeable voice quality impairments.

OPPORTUNITIES FOR FURTHER WORK

This study focused upon the interactions between best efforts data traffic, MSO QoS managed PSTN voice, and third party over-the-top best-efforts PSTN voice. As mentioned the best efforts data traffic will be a mix of delay insensitive websurfing and other delay sensitive and insensitive other data traffic (e.g. streaming music, videoconferencing, file uploads, multiplayer games). An obvious extension of this study would be to attempt to understand the nature and impact of other than websurfing best efforts data traffic. It can be expected that there may well be significant impairments to best-efforts PSTN voice due to large peak downstream data bursts in addition to the upstream resource limitation impacts exposed by this study.

Other areas for additional study would involve refinement of the discrete event model to increase the level of complexity from the high macro level to one where the random nature of some of the input parameters are explored. In particular, the parameter employed in the model for simultaneous background calls was based on the offline (i.e. outside of the model) computation of the mean number of simultaneous calls. A model refinement would use not just the mean but would explore the probability distribution to yield a packet loss rate estimate at a specific probability (e.g. to address what level of adoption would keep packet loss rates under 0.01% for 90% of the time).

CONCLUSIONS

There is significant promotion by over-the-top providers of PSTN voice services who suggest these services can be used in a best efforts mode on DOCSIS cablemodem networks. At the same time these over-the-top providers are promoting their services, MSOs are rapidly rolling out QoS managed PSTN services. Both the MSO and over-the-top providers are representing their services as toll quality comparable to the wireline services now offered by incumbent local exchange carriers (LEC).

However, if a rational MSO economically dimensions their DOCSIS service to meet their service obligations for websurfing and MSO provided PSTN voice, it is unlikely that third party over-the-top PSTN providers will be able to maintain LEC comparable quality of service.

There appears to be two mechanisms that will impair the quality of over-the-top PSTN voice. One mechanism derives from limited upstream reservation (i.e. traffic payload) minislots. For this mechanism, as MSO provided PSTN voice grows, there will not be enough voice circuits for over-the-top PSTN voice to meet the <1% probability of blocking QoS criteria. The analysis findings for the system examined in this study suggests that there will be a steep decline in over-the-top PSTN voice service when the MSO traffic levels reach adoption levels as low as 9% of cablemodem subscribers.

The second mechanism impairing over-the-top PSTN voice is independent of the growth of MSO provided PSTN voice. Unlike the limited reservation minislot mechanism above, this second mechanism is controlled by limited contention minislots available for a best-efforts bandwidth grant requests. For the example system modeled, it appears that an over-the-top PSTN voice service in best efforts mode will experience packet loss that degrades voice quality when there are about 3 simultaneous other active calls competing with one subscriber under test. Given typical residential subscriber call arrival rates and holding time, the average adoption rate at which the impairment becomes noticeable occurs at about 50% adoption of over-the-top PSTN voice by cablemodem subscribers.

ENDNOTES

[1] "A Traffic Characterization of Popular On-Line Games" by Feng et al, IEEE/ACM Transactions on Networking, Vol 13, No. 3, June 2005

[2] The source for these parameters is the same as the source note shown in figure 8.

[3] "Multimedia Traffic Engineering for HFC Networks", Cisco Systems, 1999 states typical traffic intensity between 3-6 CCS, so a value of 0.1Erlang was selected at the conservative low end between 3-4 CCS

[4] ibid, "Multimedia Traffic Engineering for HFC Networks", Based on of 0.1 Erlangs of traffic during the busy hour and mean residential call duration of 8 minutes. 0.1 Erlangs is average 6 minutes of busy hour traffic/subscriber of 8 minutes duration is 0.75=(6/8)

[5] Extend™ is a trademark for discrete event modeling software claimed by the firm of ImagineThat.