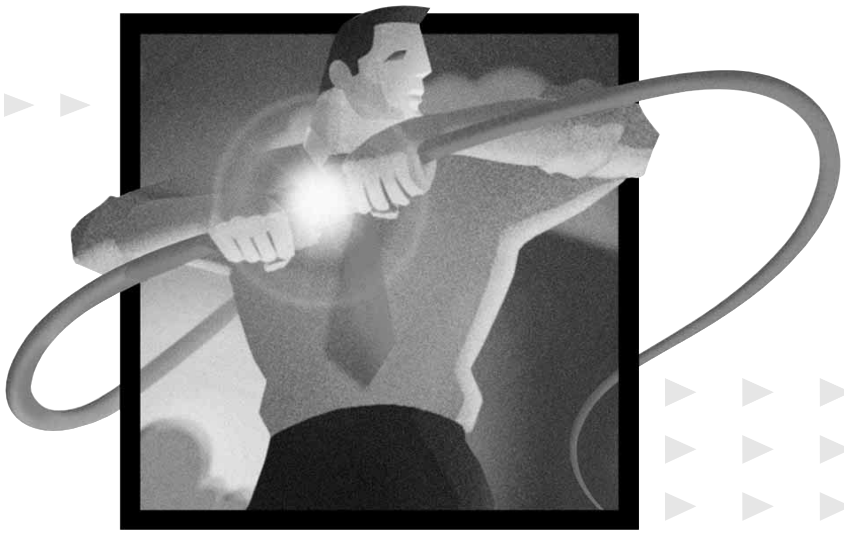


Multimedia Traffic Engineering for HFC Networks

A White Paper on Data, Voice, and Video over IP

Cisco Systems

Table of Contents

1 Introduction

1.1 Abstract

Traditional voice traffic engineering based on erlang traffic loading values and expected high day busy hour (HDBH, the year's busiest one-hour time period) peaks have been used for years in the telephony environment to build voice systems based upon customer loading. The concepts apply well to a private automated branch exchange (PABX) or central office (CO) switch, but do the existing modeling concepts also apply to hybrid fiber coaxial (HFC) plants?

HFC plants have far more diverse and irregular physical architectures that incorporate much more than bundles of twisted-pair wires. What constraints are imposed by a fiber node (FN) architecture? How does adherence to the Data over Cable System Interface Specification (DOCSIS) standard impact voice capacity? How is the ideal ratio of fiber nodes to DOCSIS cable modem termination system (CMTS) upstream ports calculated? What about allowing for redundancy?

Adding voice over IP (VoIP) functionality introduces additional complicating variables. What is the impact of encapsulating voice into packets versus a simple byte transport? How is Voice Activity Detection (VAD) accommodated? How is voice compression handled? What if the network runs a mixture of voice compression schemes? What is the impact of fax and analog modem traffic?

Data-enabling the network brings even more challenges. How do you satisfy both peak and average rate demands? How do you guarantee service? What about the emerging application of video streaming? Video streaming could supply more bandwidth into the internet than voice and data combined. How do unicast and multicast differ in their traffic models? How does an infinite amount of programming map to a finite amount of network bandwidth?

This paper addresses the basic elements of traffic engineering for HFC networks and proposes answers to each of the above questions. The approach is from a practical point of view, so that network designers will be able to ultimately answer the two most critical traffic engineering questions: "What equipment is needed, and how much is it going to cost?"

Updates to this paper as well as the supporting spreadsheets are available at ***http://www.cisco.com/cable***

1.2 General Approach

This paper explains how to size an HFC plant with data equipment to handle data, voice, and video connectivity. Classic traffic engineering principles are mixed with practical experience, and then related to actual equipment with real-world limitations. The assumed network configuration locates the CMTS at the hub location. The included calculations also work if the CMTS is located at the headend (HE), and the fiber node upstreams are backhauled to the headend using dense wave division multiplexing (DWDM) equipment at the hub.

The general approach described is the mapping of upstream and downstream connections between the fiber nodes and the CMTS. This mapping involves a round-up or round-down error, because there is always an integral number of fiber nodes and CMTSs. Round-off errors occur at each stage of the connection process, including the CMTS line cards, the hub backhaul, and the headend interconnection.

Equipment sizing is based upon a number of overlapping considerations. In order of priority, they include:

- Upstream location of CMTS
- Upstream combining ratio
- Upstream and downstream traffic

For low customer penetration rates, deployment of equipment will be determined by limits on upstream combining ratios. It will simply require a certain amount of equipment to connect all the upstreams. Service penetration rates would have to exceed some threshold before more equipment would be needed. For data only, this threshold may be 5 percent, depending upon the data rates and upstream combining ratios used.

It should be noted that although there may be a target number of households passed (HHP) per fiber node, no two fiber nodes are ever alike. The demographics of the population that the fiber node services will also vary from fiber node to fiber node. Thus a real installation could require sizing on a per-fiber node basis.

This paper pursues the equipment sizing problem from both theoretical and practical points of view and presents applicable engineering formulas. These formulas are followed by a concrete example that helps illustrate the procedure. Most of the formulas relate to the same running example. Results are rounded off to the relevant significant figure and often are rounded off to the practical limits of each piece of equipment.

1.3 Specification Template

The following is a template of the configuration parameters that are required to design a VoIP-over-HFC network. The values below are used in the example in this white paper. The values provided are generic and should be modified to each individual HFC network. The significance of each of these input values is covered in the text of this white paper. When variables are used that originate from this template, the reference format looks like *Variable Name (template)*.

Plant:	
HHP per Headend	100.000 HHP
Hubs per Headend	5 Hubs
Fiber Nodes per Hub	40 FN
HHP per Fiber Node	500 HHP
Upstream Combining of Fiber Nodes	
Max HHP per Upstream	2000 HHP
Max FN per Upstream	4 FN
DOCSIS Parameters	
Downstream QAM Level	64 QAM
Downstream Symbol Rate	5.057 Msps
Upstream QAM Level (4=QPSK)	4
Upstream Symbol Rate	1.280 Msps
MAC and PHY overhead u/s	15%
MAC and PHY overhead d/s	10%
Maintenance and Signaling OH per CM	1 kbps
Initial Maintenance Slot Size	2 ms
IP Network Parameters	
IP Network Signaling Overhead	2%
IP Packet Overhead	40 bytes
IP Framing Overhead	18 bytes

Voice:

Market Penetration	
First Line	20%
Second Line	15%
Third Line	10%
Fourth Line	5%
Codec Info	
Codec 1	64 kbps
Codec 2	64 kbps
VoIP Frame Interval	10 ms
Connections per 100% of Upstream	
Codec 1	50
Codec 2	23
Codec Usage	
First Line	codec #2
Second Line	codec #2
Third Line	codec #2
Fourth Line	codec #2
HDBH Traffic	
First Line	6 CCS
Second Line	5 CCS
Third Line	4 CCS
Fourth Line	3 CCS
ABSBH Traffic	
First Line	5 CCS
Second Line	4 CCS
Third Line	3 CCS
Fourth Line	2 CCS
HDBH Grade of Service	1.0%
ABSBH Grade of Service	0.5%
Use HDBH, ABSBH, or Worst Case	Worst case
Offnet Traffic	85%
Call Attempts	
Call Attempts per Hour per Line	2.78 calls
Call Holding Time	180 seconds
CMTS Signaling Packets per Call	20 packets
TGW Signaling Packets per Call	10 packets
DOCSIS PHS for Voice	Yes
Voice Region Allowance	10%
VAD	
VAD Enabled	No
VAD Activity Factor Two Way	40%
VAD Grade of Service	1%
VAD Region Allowance	0%

Data:

Market Penetration	
Data Penetration	10%
Overlap of Voice and Data	5%
Subscribers Activity Factor (relative to market penetration	
Average	20%
Peak	5%
Downstream Data Rate:	
Average	80 kbps
Peak	400 kbps
Max	2000 kbps
Upstream Data Rate:	
Average	32 kbps
Peak	100 kbps
Max	768 kbps
Average Packet Size	
Downstream	500 bytes
Upstream	64 bytes

IP Video Streaming (IPVS)

FN per IPVS Downstream Group	2
IPVS Market Penetration	40%
IPVS Activity Factor	60%
IPVS Sessions per Home	1.3
IPVS Unicast Usage	70%
IPVS Multicast Usage	30%
IPVS Multicast Derating	
IPVS Multicast Available	125 streams
IPVS Multicast Used Max	80%
IPVS Multicast Shape Factor	1
BW per IPVS Session	1500 kbps
Include with Other Downstream	No

Equipment Parameters:

CMTS

Downstreams per CMTS Domain	1 down
Upstreams per CMTS Domain	6 up
CMTS Domain per Line Card	1 domain
Line Cards per Chassis	4 line cards
N Max Redundancy Ratio	8:1 max

TGW

Voice Ports per Line Card	
G.711	192 ports
G.728	96 ports
G.729	144 ports
TGW Line Card per Chassis	20 line cards
Include with Other Downstream	No

CMS

Call Attempts per Hour per CMS	200,000 cat

DPT

Ring Bandwidth (for duplex traffic)	2400 Mbps

2 Background

2.1 Telephony Traffic Engineering Basics

Telephony connections have classically been divided between lines and trunks. Lines refer to the twisted pairs that actually terminate on plain old telephone service (POTS) phones. Trunks refer to the connections that backhaul the voice traffic into the network. A switch concentrates lines into trunks.

In the DOCSIS scenario, the cable modem (CM) is acting as this switch. It has POTS lines on one side and DOCSIS connections on the other side that act as trunk lines. So how many trunk lines are needed to support a given number of POTS lines? The answer lies in statistics and traffic engineering.

Enter the erlang. The erlang refers to traffic loading from a line or group of lines and has a value between 0 and 1 for a single line. If a line has an erlang value of 0.1, that means it is off hook 10 percent of the time. If there are 20 lines, each with 0.1 erlangs of traffic, then the total offered traffic is 2.0 erlangs.

If those 20 lines were switched to 20 trunks, no calls would ever get blocked. However, trunk utilization would still only be 10 percent, a level that would not be cost-effective. At the other extreme, if the traffic were evenly distributed in time among the 20 calls, all the calls could switch into two trunks with 100-percent utilization—very cost-effective, but not very practical. So the number of trunks required for this example is somewhere between 2 and 20.

Enter the blocking probability. Since the number of trunks will be less than the number of lines, there is a probability that a new call will be blocked. This phenomenon is known as the blocking probability or grade of service (GoS).

The formula that ties these factors together is the erlang B formula:

Formula (1)

$$P_r = \frac{\frac{A^N}{N!}}{\sum_{J=0}^{N} \frac{A^j}{j!}}$$

- A = Offered traffic load in erlangs
- N = Number of trunks
- Pr = Probability of blockage (GoS)

Given any two variables and a good spreadsheet, the third variable may be solved for. In the simple example with A = 2 erlangs and a typical value of Pr = 1 percent, the number of required trunks is 7. To express this formula solved to express N, the nomenclature that will be used in this white paper to reference the erlang B formula is as follows:

Formula (2) ***N = Erlang-B[A, Pr]***

Result: N = Erlang-B[2, 1%] = 7

A more conservative approach is to not believe that the calls will evenly distribute with large numbers, and pick a limit on the erlang B formula. An historic number used in telephony traffic engineering has been 90 percent. This white paper uses a nomenclature of “Erlang-BL” to indicate the erlang B formula with a 90-percent limit imposed.

The crossover point for the erlang B formula and the 90-percent limit and a GoS of 1 percent is 181 erlangs. So below 181 erlangs, the erlang B formula is used. Above 181 erlangs, the 90-percent limit formula is used.

Formula (3) ***N = Erlang-BL[N, Pr] = Max[N / 90%, Erlang-B[N, Pr]]***

Result: N = Max [181 / 90%, Erlang-B[181, 1%] = Max [202, 202] = 202

In this example, 202 trunks will be required if the total traffic loading is 181 erlangs.

The crossover point for the erlang B formula and the 90-percent limit is 181 erlangs. So below 181 erlangs, the erlang B formula is used; above 181 erlangs, the 90-percent-limit formula is used.

Appendix E: "Erlang B Tables" on page 70 contains an erlang table and Appendix F: "Excel Spreadsheet Erlang Macros" on page 73 provides Excel spreadsheet macros for calculating the erlang B formula.

So one erlang of traffic could be for one call that lasted one hour, or for ten calls, each lasting six minutes. This scenario introduces two more concepts that are intertwined, call attempts per hour and call holding time.

Given:

Erlangs = Call Attempts per Hour * Call Holding Time / 3600 seconds

Then:

Formula (4)	***Call Attempts per Hour = Erlangs * 3600 / Call Holding Time***
Result:	Call Attempts per Hour = 0.139 * 3600 / 180 = 2.78

Another method of expressing traffic is in terms of Calls per Centum-Seconds (CCS). A centum-second represents 100 seconds. There are 3600 seconds in an hour, so there are 36 centum-seconds in one hour. An erlang is usually averaged over an hour, so 1 erlang is equivalent to 36 CCS. The conversion from erlangs to CCS follows:

Formula (5)	***Erlang = CCS / 36***
Result:	0.167 erlang = 6 CCS
Result:	0.139 erlang = 5 CCS
Result:	0.111 erlang = 4 CCS
Result:	0.083 erlang = 3 CCS

The typical value used for CCS per line is between 3 and 6 CCS.

Another concept popular in telephony traffic engineering is the busy hour. The busy hour during a day is a one-hour period that contains the peak traffic. The phone system must work during this time. An interesting rule found in reference [1] is that the traffic during busy hour represents 15 to 17 percent of the traffic during a one-day period.

High Day Busy Hour (HDBH) is the busy hour during the busiest (highest) day of the year, usually the day after Thanksgiving. Average Busy Season Busy Hour (ABSBH), a subset of HDBH, excludes anomalies such as Mother's Day, Christmas, and natural disasters. The three months (not necessarily consecutive) with the highest average traffic in the busy hour are termed the "busy season." The busy hour traffic level averaged across the busy season is termed the "ABSBH load." ABSBH is always less than HDBH. Historically, HDBH is about 20 percent higher than ABSBH. Voice networks may be designed to either HDBH or ABSBH, although conservative networks usually use HDBH.

AT&T published some interesting statistics on its Web site at *http://www.att.com/network/netfacts98.html* that are useful assumptions for network design. To preserve their accuracy, they are quoted here in their original form.

Call Setup Time

- *Call set-up time is the period between the last number being dialed and the distant telephone ringing. On a typical point-to-point domestic call, call set-up time averages less than two seconds. It is approximately one-half second longer for a credit-card, toll-free, or "900" call.*

Calling Averages

- *The average business call is 2.5 minutes in length.*
- *The average consumer call is 8.0 minutes in length.*
- *Approximately 75 percent of all calls complete when dialed. The remaining 25 percent receive a busy signal or no answer.*

Calling Patterns

- *July and August are the busiest calling months of the year.*
- *Monday is the busiest long-distance calling day of the week, while Sunday generates the lowest number of calls.*
- *Monday after the four-day Thanksgiving weekend is the busiest day of the year for the network. However, Mondays during July and August also generate extremely heavy traffic.*
- *Sunday of the Memorial Day weekend generates the least number of calls for the network. The Fourth of July and Sunday of the Labor Day weekend are also low-volume days.*

Table 1 Peak Calling Hours - Eastern Time

Day	Time
Monday – Friday	10:00 a.m. to 12:00 Noon 2:30 p.m. to 4:30 p.m.
Saturday	10:00 a.m. to 2:00 p.m.
Sunday	7:00 p.m. to 9:00 p.m.

Holiday Calling

Certain holidays, such as Mother's Day and Christmas Day, are among the busiest long-distance calling days of the year in terms of overall network usage. Although the number of calls placed is approximately 50 to 60 percent of an average business day, the holiday calls tend to be longer than standard calls and are more evenly spread across the network. On a typical business day, the majority of calls travel from one major metropolitan business center to another; however, on holidays, network traffic is more evenly spread across the entire nation. Due to their popularity as retirement centers, the Sunbelt States of Arizona, California, Florida, and Texas generally receive a heavy volume of inward calls during holiday periods.

Based upon the number of calls handled, the top six long-distance calling holidays are:

1. *Mother's Day*
2. *Christmas Day*
3. *Father's Day*
4. *Easter Sunday*
5. *New Year's Day*
6. *Thanksgiving*

2.1.1 Internet Traffic Reports

The Internet has some interesting sites that contain some level of traffic statistics. The most well known site is probably the Internet Traffic Report located at *http://www.internettrafficreport.com*. A copy of the daily report is included in Figure 1. Note the measured average response time of 310 ms and average packet loss of 2 percent. One technical detail is that these numbers are usually based upon trace-route numbers that use an Internet Control Message Protocol (ICMP) echo. This type of traffic is typically process switched in routers, meaning it is not switched through the fast path where the majority of data traffic is switched. ICMP echoes are the first traffic to get delayed or discarded when there is congestion. Thus these numbers are more worst case. Also, they are across the public Internet, which contains a mixture of engineered and not-so-engineered pathways. The variation in latency and loss should be lower and more contained in a private Intranet. Still, these numbers represent some good upper bounds to design to.

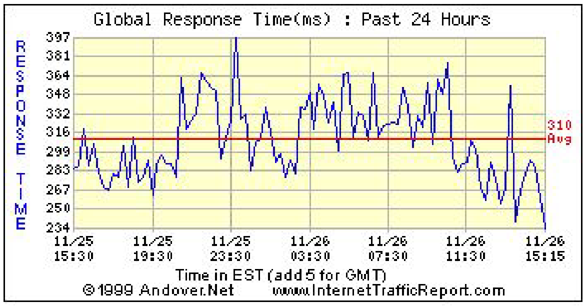
Other interesting sites on the Internet are the peering points between the major Internet service providers (ISPs). These are the points where traffic is taken off the network of one service provider and routed onto the network of another service provider. The largest peering points today are known as MAE West in San Jose, CA, and MAE East in Washington, DC. These two MAEs are operated by MCI/Worldcom. MAE is an acronym for Metropolitan Area Ethernet, although Internet folklore suggests that the name came from the 1930s screen goddess MAE West, and the acronym was a second priority. Information on MAE West and its sister sites can be found at *http://www.mae.net*. There are at least seven of these MAE peering points in the United States, with more world wide.

Figure 1 Internet Traffic Report

This graph shows the Global Traffic Index for the past 24 hours.

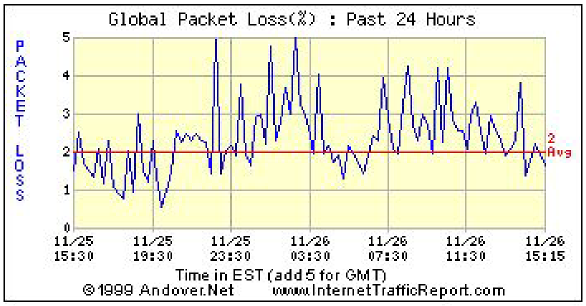
The “traffic index” is a score from 0 to 100, where 0 is “slow” and 100 is “fast.” It is determined by comparing the current response of a ping echo to all previous responses from the same router over the past 7 days. A score of 0 to 100 is then assigned to the current response depending on if this response is better or worse than all previous responses from that router.

This graph shows the Global Response Time for the past 24 hours.



Response Time in reference to Internet traffic is how long it takes for a chunk of data to travel from point A to point B and back (round trip). A typical response time on the Internet is 200 ms, which is 200 milliseconds (thousandths of a second) = 1/5 of one second.

This graph shows the Global Packet Loss for the past 24 hours.



Packet Loss measures the reliability of a connection. A known chunk of data is sent to the router and then the router is supposed to send the same data back unaltered (echo). If the packet returns intact, then there was 0 percent packet loss and that is good. If the packet returns with half of its data garbled or missing, then that is 50 percent packet loss, which is poor. A packet loss of 100 percent means the packet never returned.

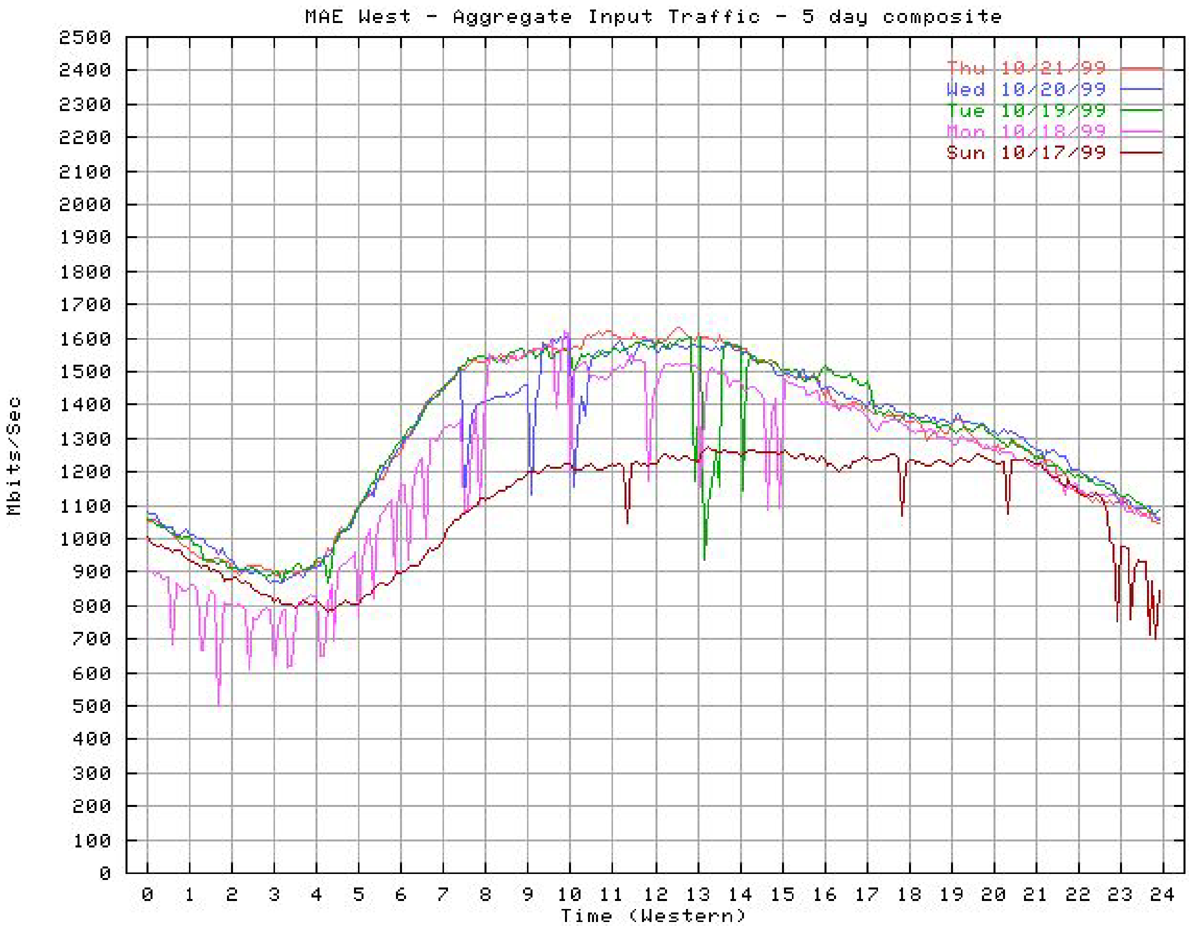
Figure 2 shows the total aggregate traffic through MAE West for five consecutive days in November 1999. Down spikes in the graph are due to corrupt or missing measurement data and can be ignored. Some interesting observations can be made.

- The peak traffic load within ~1 percent occurs between 10:00 a.m. and 2:00 p.m.
- The peak traffic load within ~10 percent variance occurs between 7:00 a.m. and 5:00 p.m. This is the entire workday.
- The traffic never seems to fall below 50 percent of peak.

So busy hour traffic for data seems to last all day. This isolated statistic represents Internet traffic, so it represents a combination of business and residential traffic. When the multisystem operators (MSOs) build their data backbone, if they continue to service their traditional residential market, the peak traffic is likely to shift into the evening.

The perspective one has to have is that there is no good, known set of traffic statistics for data. And with the Internet traffic doubling every six months or so and the mix of applications radically changing, anything that does exist is quickly outdated. The

Figure 2 MAE West Aggregate Input Traffic



only choice an operator has is to constantly monitor the traffic levels within a given network, monitor its growth rate, and make an effort to predict the future.

2.2 HFC Plant Layout

2.2.1 Headends, Hubs, and Fiber Nodes

In the beginning, cable plants were composed entirely of coax cable built into a large tree and branch networks. This scenario worked fine when there were many cable operators, all with very small plants and only a few downstream channels. As networks grew and operators began consolidating, fiber was introduced into the plant. Hence the name hybrid fiber coax (HFC) plant.

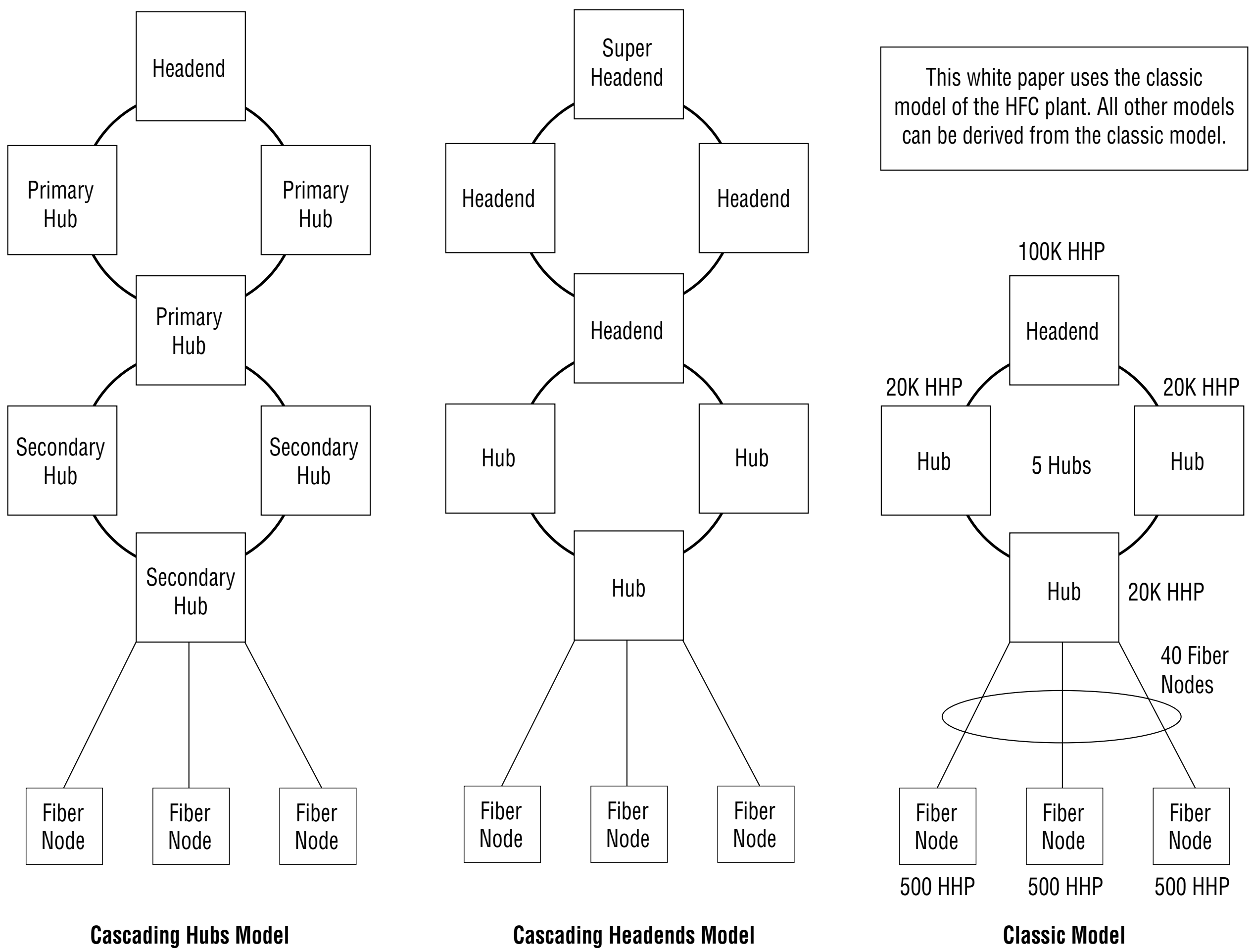
There are two types of fiber usage—analog and digital. The main hierarchy of components that are found in medium-sized HFC systems are the headend, the hub, and the fiber node. Video signals are delivered to the headend by land or by satellite, and may be in analog or digital format. In the downstream, the analog video channels are digitized and sent over a fiber ring to a series of hubs. The hubs convert the signals back to analog and send them down point-to-point fiber links to fiber nodes. The fiber node then receives the analog signal from the fiber and retransmits it onto a tree and branch coax network that ultimately connects to the subscriber.

One of the advantages to segmenting the coax network into 500 HHP sections is that the upstream path can be engineered to return signals from the subscriber's home, thus making the plant a two-way plant. The upstream signals that are received at the fiber node are retransmitted onto a point-to-point fiber optic link and sent to the hub. In our example system, there would be

40 fibers returning to the hub. The current digital fiber ring that connects the headend to the hubs does not have the capacity to carry all these upstream back to the headend, so the upstream path stops at the hub. Also, since before DOCSIS, there was not much significant information in the upstream to economically justify returning all 200 fibers from all five hubs all the way to the headend.

The classic model shown in Figure 3 applies to many cable companies, but there are those that are seemingly different as shown in Figure 3. As big cable companies have bought up smaller cable companies, a four-layer model has emerged. Using a four-layer model, some cable companies map out plants as fiber nodes connected to secondary hubs, secondary hubs connected to primary hubs, and primary hubs connected to the headend. Others label the four layers as fiber nodes, hubs, headends, and super headends. On top of these layers, MSOs can also have a regional headend. In practice, the headend also hosts one of the hub sites.

Figure 3 HFC Plant Layouts



A recent technology revolution in optics has made dense wave-division multiplexing (DWDM) a practical technology. DWDM allows multiple signals paths to be multiplexed onto one fiber with each signal path at a slightly different wavelength. By combining 16 or 32 wavelengths per fiber, all the return paths from a hub can be multiplexed onto two or three fibers. This manageable number of fibers allows the upstreams to be returned all the way to the headend. The advantage of this approach is that complex pieces of equipment like CMTSs are easier to manage and service if they are all colocated in one staffed facility instead of being located out in unstaffed hubs. Many of the larger MSOs are now adopting DWDM for their new network upgrades.

Another dimension to the plant layout model involves the placement of the CMTS. Today, a CMTS sits at the hub so that all the upstream fibers don't have to run up to the headend. Consider the required number of fibers in our example: each hub has 40 fibers from the fiber nodes. If fiber were carried through the hubs to the headends, the headends would receive 200 fibers (40 fibers from each hub with five hubs). Numbers in this range are too impractical to manage. Most operators place the CMTS at the hub so that fiber does not have to go beyond that layer.

Alternatively, some companies are employing DWDM to move the CMTS higher up in the hierarchy. With DWDM, light from multiple fibers can be multiplexed onto a single fiber. In our example, the 40 fiber nodes per hub could be consolidated onto 3 fibers that a maximum of 16 wavelengths per fiber. The five hubs aggregated, each with 3 upstream fibers, would require a total of 15 fibers back to the headend. These 15 fibers would carry the 200 wavelengths of light from the 200 fiber nodes in the system. Now all the light can be backhauled to the headend realistically. Equipment can be racked at the headend, simplifying the hubs and the overall manageability of the infrastructure. This methodology is employed today by some cable operators.

The calculations in this paper can be applied to all these variations of three- and four-layer models, as well as CMTS placement at the hub or the headend. For the purpose of traffic engineering, the critical factors are the placement of the CMTS and the number of fiber nodes under the CMTS. Our calculations assume the CMTS is placed where the upstream wavelength stops, with a headend or a super headend above the CMTS, and some number of fiber nodes below the CMTS. In this fashion, all the models can be reduced to the three-tier model for the purposes of calculating traffic requirements.

The number of households passed per headend system in the basic HFC system in this white paper is always related to the number of hubs, fiber nodes, and households passed per fiber node. That relation is described in the following formula for the headend:

Formula (6) ***HHP per HE = HHP per FN * FN per Hub * Hubs per HE***

Result: HHP per HE = 100,000 = 500 * 40 * 5

and for the hub:

Formula (7) ***HHP per Hub = HHP per FN * FN per Hub***

Result: HHP per Hub = 500 * 40 = 20,000

2.2.2 CMTS-to-Fiber Node Connectivity

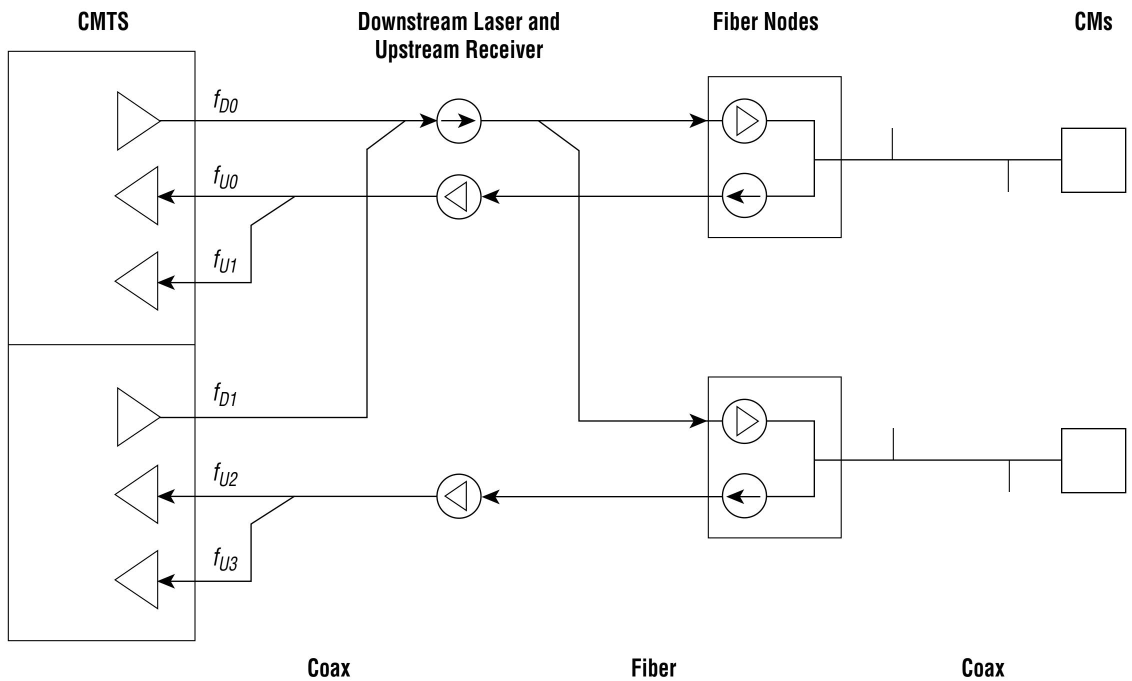
Since one of the main goals of traffic engineering is to connect fiber nodes to the CMTS, it is worth considering how the connectivity is actually done. Methods vary widely, depending on the age of the plant, cost of the solution, and the functionality that is trying to be achieved.

The CMTS coupled with an up converter generates a channel, which is combined with the existing video channels and coupled into a downstream laser. This laser is usually high quality and wide bandwidth, meaning it can be expensive. Since the downstream is a broadcast media, the fiber signal is optically split and sent to multiple fiber nodes. The output of the optical is referred to as the forward carrier path (FCP). The signal from the upstream fiber is received at the hub and is converted into an electrical signal to be sent over coax. This output of this receiver is referred to as the reverse carrier path (RCP). (See Figure 4.)

Because of the downstream fiber splitting, there is a natural grouping of fiber nodes. If the downstream path were split into four FCPs, then the CMTS would be presented with one downstream and four upstreams that it could not subdivide.

In addition, some fiber nodes are split internally with one downstream fiber feeding all the households passed, and two upstream fibers, each supporting half the number of households passed. For example, some fiber nodes are split internally with one downstream connection serving 1000 HHP and two upstream fibers, each serving 500 HHP. In this case, the FCP is larger than the RCP. The formulas in this paper assume the simpler model of having the same number of homes passed in the downstream and the upstream. This situation can be compensated by calculating the downstream and upstream with different size fiber nodes, or just using the upstream size for all calculations and dividing by two at the end.

Figure 4 CMTS-to-Fiber Node Connectivity



Another variation is to run two downstream fibers to the fiber node. One fiber carries the analog video from 54 to 550 MHz. This fiber is driven with a single, more expensive laser that uses an optical splitter to drive many other fiber nodes. The second fiber is driven directly with a less expensive laser with a bandwidth of 550 to 750 MHz. This setup allows the digital channels to service smaller groups for more local content without having to take the cost increase involved with duplicating the more expensive laser.

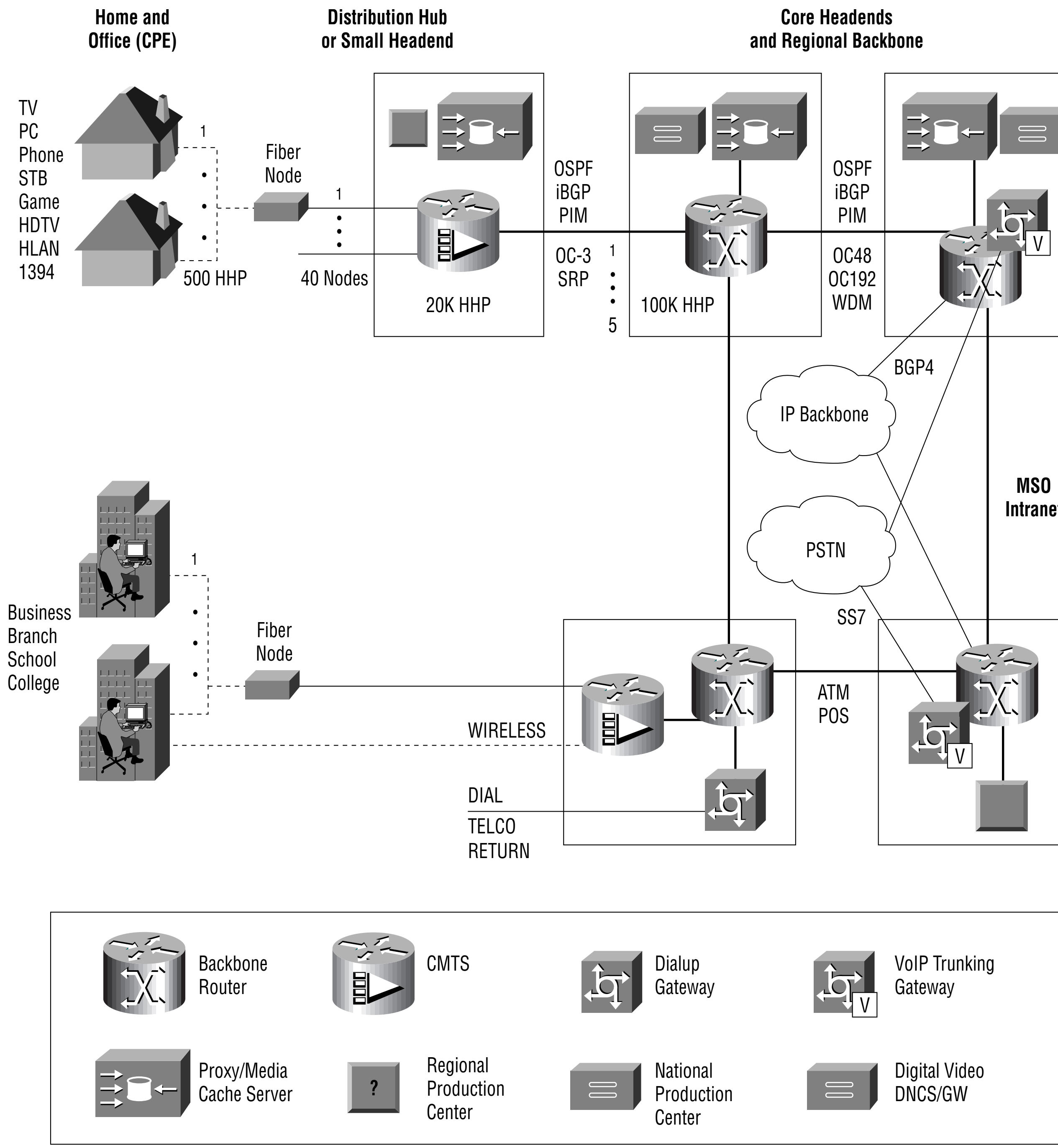
The RCP is connected to the receivers of the CMTS either by splitting, direct connecting, or by combining. When combining, three or four RCPs may connect to one CMTS receiver. For 500 HHP per fiber node, the result would be 2000 HHP per CMTS receiver. Direct connection would result in one RCP per CMTS receiver, or 500 HHP. Splitting would result in multiple CMTS receivers per RCP with perhaps only 125 HHP per CMTS receiver. Traffic engineering must figure out which one of these three scenarios to use.

The white paper assumes the most basic model of one FCP and one RCP per fiber node. The FCP is referred to as a downstream, and the RCP is referred to as an upstream.

2.3 Equipment Description

The calculations in this paper include some assumptions about the equipment used throughout the plant infrastructure. This section summarizes those equipment characteristics relevant to the calculations. A three-tier model retrofitted for data might look like the plant in Figure 5.

Figure 5 HFC Plant Layout



2.3.1 DOCSIS CMTS

A Cable Modem Termination System (CMTS) Media Access Control (MAC) domain, as it is defined, includes one or more downstream paths and one or more upstream paths. A CMTS contains one or more CMTS MAC domains. Depending upon the construction of the CMTS, the CMTS MAC domain may have its downstreams on one card with its upstreams on another card, or there could be one or more CMTS MAC domains per line card.

The basic model used in this white paper assumes one MAC domain per line, with variables including the number of downstreams per line card, number of upstreams per line card, and number of line cards per CMTS.

The raw bandwidth in each of the downstreams and upstreams is related to the symbol rate and the number of bits per symbol resulting from the modulation technique.

Formula (8)	***BW Raw = Bits per Symbol * Symbol Rate***
Result:	BW per Downstream Raw = 6 * 5.057 = 30 Mbps
Result:	BW per Upstream Raw = 2 * 1.28 = 2.56 Mbps

The usable bandwidth is equal to the raw bandwidth minus physical layer overhead.

Formula (9)	***BW Usable = BW Raw * (1 – PHY and MAC Overhead)***
Result:	BW per Downstream Usable = 30 * (1 – 10%) = 27 Mbps
Result:	BW per Upstream Usable = 2.56 * (1 – 15%) = 2.2 Mbps

2.3.2 DOCSIS Cable Modem

CMs reside at the customer's site and terminate the DOCSIS protocol. On the network side, they connect to coax. On the other side is a data port such as Ethernet, universal serial bus (USB), or Home Networking. If the CM has an embedded multimedia terminal adapter (MTA), it can provide data, voice, or video services.

If located outside the dwelling, the CM can serve as a network interface device (NID) that is network or battery powered. This configuration is being used for delivering primary voice services today. If located inside the dwelling, the cable modem can serve as a straight CM for data services, or as a set-top box for a variety of interactive and on-demand services.

2.3.3 Trunk Gateways

A voice trunk gateway interfaces between the Public Switched Telephone Network (PSTN) and a data network. On one side of the trunk gateway (TGW) is a PSTN connection such as a T1, DS3, or an OC-3 connection containing circuit-switched voice in 64-kbps DS0 time slots. On the other side of the trunking gateway is a data connection such as Ethernet, Packet over SONET, or Packet over ATM over SONET.

Thus a trunk gateway converts circuit-switched voice on a 64-kbps DS0 to IP packetized voice on a data media. Echo cancellation towards the PSTN is provided as well as voice compression on the data side. The important factors in sizing a trunking gateway are the number of DS0s or DS1s as well the number of VoIP connections. It is not uncommon for the VoIP connections to be dependent upon the compression algorithm, because the compression is usually done in a digital signal processor (DSP) farm.

2.3.4 Digital Video Solutions

Today, video sources are Motion Picture Experts Group (MPEG) encoded and then fed into an MPEG multiplexer that packs six to ten MPEG video streams into a 27-Mbps stream. This stream is uplinked to a satellite and then down linked to multiple headends that then distribute the prepackaged MPEG stream directly onto the HFC plant. This system is a very efficient and cost-effective for broadcast digital video from predetermined sources.

A complementary technology on IP networks called video streaming is emerging. Video streaming follows the same principles, but over an IP network instead of a satellite network. Streaming video originates from a video server or a real-time video encoder that encapsulates the video in IP. The video-over-IP packets are then sent over the IP network to the CMTS, which then passes them through on a DOCSIS downstream.

The application for this technology is to provide connectivity to video sources that are not real time or are not available with a conventional satellite distribution system. This unavailability could be because of the location of the video source, or the fact that it has subscriber specific content that is more applicable to unicast or multicast applications, rather than general broadcast.

2.4 Strategies for Scaling the Infrastructure

Good traffic engineering tries to provide an initial build-out of the plant that will cover deployment for the foreseeable future. In practice, the IP traffic (data, voice, and video) will not be evenly distributed across the plant. Some areas might require upgrading before others. Allowing for expansion on the initial equipment rollout eases subsequent upgrades. The planning options available include:

If upstream bandwidth to a fiber node runs out:

- Leave a spare receiver(s) per line card initially or empty slots so that the receiver may be added in later. This setup allows for easy upgrades, but may increase the amount of original equipment.
- Move the upstream to a new line card where more receivers are available. This setup is cost-effective because equipment is added on an as-needed basis. However, it requires reprovisioning of the cable modems on that line card, a process that could be laborious.
- Replace the line card with a higher-density line card and reuse the lower-density line card in another system. This solution is very practical, unless the line card is already as dense as is available.

If the downstream bandwidth runs out:

- Split the fiber node to minimize the number of downstream channels used. Also, with that much bandwidth usage, the market penetration is high, so a fiber node split may be justified.
- Add another downstream channel. If the increase of bandwidth is uniform across the fiber node because of high-bandwidth applications such as Internet streaming video, then everyone may just need more bandwidth.

If the backbone bandwidth runs out:

- Backbone routers with Synchronous Optical Network (SONET) interfaces in the backbone will allow upgrades from OC-48 to OC-192 without downtime.
- Transition SONET rings to dynamic packet transport (DPT) to increase ring efficiency.

3 HFC Traffic Sizing

3.1 General Approach

A DOCSIS plant has four fundamental groups of traffic to be sized: data, voice, video, and signaling and maintenance. Traffic engineering should be done for each of these applications, as well as separately for the upstream and downstream channels. Because the optimized voice-density calculations in Appendix D: "Voice-over-IP Call Density Tables" on page 68 include physical (PHY) level overhead, the other applications are individually mapped to the PHY layer as well. This section calculates the contribution from each application, along with any limitations that the application may contribute.

The fundamental approach is to establish how many fiber nodes are serviced by how many upstream receivers. This physical connectivity will cause a rounding error that may result in a requirement for extra resources. For example, if 1.5 upstream receivers are required to service one fiber node, then two upstream receivers are required, resulting in 33-percent extra resources.

3.2 Data

In general, voice traffic is symmetrical, has constant bandwidth, and has a low tolerance to latency and jitter. Data traffic is the complete opposite. Data traffic is asymmetrical and bursty, and is highly tolerant to latency and jitter. Traffic engineering must use statistics that reflect these differences in order to arrive at an efficient and economical result.

Note that data may be passed over a DOCSIS HFC plant in two fundamental ways. The first method is the default DOCSIS data mode where Layer 3 packets such as IP are sent natively over the DOCSIS link. This mode of transport is the preferred one, and is addressed in this section. The other mode for sending data is within a VoIP stream as modem tones. In this scenario, the data traffic characteristics are converted with a great deal of waste of bandwidth into voice traffic characteristics. If analog or fax modem traffic is carried directly in a VoIP connection, then the erlang traffic and call holding times should be kept separate and calculated separately from the voice traffic. This mode is covered in the section "Support for Legacy Analog Modem and Fax" on page 49.

Bandwidth can be further described with the following categories:

- ***Average bandwidth:*** Since data is bursty, the bursts from different users will be statistically multiplexed together when sent on a common media. Over some extended period of time, the user will perceive an average data rate. For example, the performance of File Transfer Protocol (FTP) which could easily be transferring data for minutes, will be the most impacted by average bandwidth. In order to compare measured rates between systems, a reference time from is needed. For reference, a time interval of 5 minutes is used herein.
- ***Peak bandwidth:*** This bandwidth is that which the user perceives when the user is actually transferring data. This bandwidth would be important to applications such as Web page downloading. The number of actual users transferring data during some short interval determines the peak bandwidth. For reference, a time interval of 1 second is used herein.
- ***Max bandwidth:*** The user experiences this bandwidth during the actual burst. This limit is really intended for configuring any rate limiting the network might have, and is not used directly in traffic engineering.
- ***Media bandwidth:*** This bandwidth is the basic bandwidth of the media. If users are rate limited, they will not achieve media speeds. The media bandwidth is more significant to determine how large a group of users can be supported based upon their average bandwidth and peak bandwidth requirements.

For example, the performance requirements in the downstream may specify an average bandwidth of 80 kbps, a peak bandwidth of 400 kbps, rate limiting set at 2 Mbps per user, over a 27-Mbps media.

A percentage of the households will have subscribed to the data service will be passed. This is known as the data market penetration. Of that group, only a certain group of those subscribers will be sitting at their computers and logged on. This group is referred to as active subscribers, whose size is measured by the data average activity. Of that group, yet a smaller group of the subscribers will actually be transferring data. This group is referred to as peak subscribers, whose size is measured by the data peak activity factor.

The significance of the time measurement intervals used is that during the 5-minute measurement window for active subscribers, all the active subscribers will be provided with the average data rate. During the 1-second measurement window for peak bandwidth, a much smaller number of subscribers will be receiving peak data rates.

For example, in a particular serving area, 10 percent of the customers may have data service. Of those, 20 percent may be logged on during busy hour. Of those 20 percent, only 25 percent may be downloading data, so the peak demand is 20 percent * 25 percent = 5 percent.

When data is transferred, it might be in only one direction. For example, e-mail programs may download new e-mail for 30 seconds every 10 minutes throughout the day, with concentrated but random uploads when the user replies or sends out new mail. Web surfing may be a download every 30 seconds. More typically, applications will change over time, as will the usage pattern of each application. The asymmetry of the downstream to the upstream for these two applications could be 10:1 or even 20:1.

The two examples of e-mail and Web surfing are also good examples of asymmetrical data applications. As an approximation, assume that the upstream traffic is much, much less than the downstream traffic. Then assume that the traffic is IP, and that one IP acknowledgment (ACK) packet will be sent in the upstream for every two IP packets received in the downstream. The worst-case ratio of downstream to upstream traffic would be:

Formula (10) ***Data Asymmetry Worst Case = (2 * Average Downstream Packet Size) / (Average Upstream Packet Size)***

Result: Data Asymmetry Worst Case = (2 * 500) / 64 = 16:1

To deal with this variability, data applications use the metrics of bandwidth and switching performance. Bandwidth calculations deal with bearer-like capacity of ingress and egress. Switching performance uses packet-per-second (pps) calculations that deal with packet-switching performance of a particular node. These calculations are performed separately for each direction.

The number of active data users receiving average data rates follows:

Formula (11) ***Active Subscribers = HHP * Data Penetration * Data Average Activity Factor***

Result: Active Subscribers per HE = 100,000 * 10% * 20% = 2000

Result: Active Subscribers per FN = 500 * 10% * 20% = 10

The number of data users who are simultaneously passing data and receiving a peak data rate follows:

Formula (12) ***Peak Subscribers = HHP * Data Penetration * Data Peak Activity Factor***

Result: Peak Subscribers per FN = 500 * 10% * 5% = 2.5

Result: Peak Subscribers per HE = 100,000 * 10% * 5% = 500

Next, the required bandwidths for average and peak traffic are both calculated. The maximum of these two numbers is then carried forward. Since the desired bandwidth numbers are classically referenced to an Ethernet packet, an allowance is made for DOCSIS MAC and PHY overhead.

The average bandwidth required for both downstream and upstream during data busy hour follows:

Formula (13) ***BW Data Average = (Active Subscribers * Bandwidth per Subscriber) * (1 + MAC and PHY Overhead)***

Result: BW Data Downstream Average = (10 * 80 kbps) * (1 + 0.1) = 880 kbps

Result: BW Data Upstream Average = (10 * 32 kbps) * (1 +0.15) = 368 kbps

The peak bandwidth required for both downstream and upstream during data busy hour follows:

Formula (14) ***BW Data Peak = (Peak Subscribers * Bandwidth per Subscriber) * (1 + MAC and PHY Overhead)***

Result: BW Data Downstream Peak = (2.5 * 400 kbps) * (1 + 0.1) = 1100 kbps

Result: BW Data Upstream Peak = (2.5 * 100 kbps) * (1 + 0.15) = 288 kbps

The larger of the average and peak requirements is taken forward.

Formula (15) *BW Data = Max[BW Data Average, BW Data Peak]*

Result: BW Data Downstream = Max[880, 1100] = 1100 kbps

Result: BW Data Upstream = Max[368, 288] = 368 kbps

3.3 Voice

3.3.1 Derivation of Subscriber Density Numbers

The useful metrics for aggregated voice calculations follow:

- Market penetration
- Lines per home
- Traffic per line

The first issue to consider is that customers may have data, voice, or both. To calculate traffic densities, independent voice and data percentages need to be known. However, if we assume that a cable modem can handle one or both services, the combined percentages must also be known to calculate the number of modems needed. The raw field data usually is the separate voice and data market penetrations. The customers who share voice and data must also be known.

The most useful approach is to have the combined percentages provided (as in the Specification), and derive the separate percentages.

Formula (16) *Total Market Penetration = Σ Separate Market Penetrations - Overlap*

Result: Total Market Penetration Voice = 15% Voice Only + 5% Voice/Data = 20%

Result: Total Market Penetration Data = 5% Data Only + 5% Voice/Data = 10%

Result: Total Market Penetration Total = 20% Voice + 10% Data - 5% Voice/Data = 25%

The next issue to consider is that each home may have a different number of telephone lines and each line may have a different amount of traffic. This parameter needs to be averaged per customer and per line. Table 2 shows an example of the raw data that a MSO might assemble for their customer base.

Table 2 Example Multiline Market Penetrations

Line	Market Penetration	Relative Penetration	HDBH Traffic (CCS)
1	20%	100%	6
2	15%	75%	5
3	10%	50%	4
4	5%	25%	3

The market penetration number will also vary by year. The m1arket penetration numbers should be taken for the year when it is acceptable to upgrade the equipment. For example, if the installation will not be upgraded for two years, then traffic densities for two years from installation should be used.

Since every voice customer has at least one line, the market penetration number for the first line equals the market penetration for all voice customers. The relative penetration number is calculated by comparing the subsequent line penetration to the first line penetration.

Formula (17) **Nth Line Relative Penetration = Line Penetration / First Line Penetration**

Result: First Line Relative Penetration = 0.20 /0.20 = 100%
Result: Second Line Relative Penetration = 0.15 / 0.20 = 75%
Result: Third Line Relative Penetration = 0.10 / 0.20 = 50%
Result: Fourth Line Relative Penetration = 0.05 / 0.20 = 25%

If this table described an area of 100 HHP, intuitively there would be 50 phone lines (20 + 15 + 10 + 5) spread across 20 homes. Thus to calculate the average lines per home,

Formula (18) ***Lines per Home = Σ Line Market Penetration / First Line Penetration***

Result: Lines per Home = (0.20 + 0.15 + 0.10 + 0.05) / 0.20 = 2.5

The same result can be also be achieved by summing the *Nth Line Relative Penetration (formula 17)* numbers.

Formula (19) ***Lines per Home = Σ Line Relative Penetration / 100%***

Result: Lines per Home = (100% + 75% + 50% + 25%) / 100% = 2.5

Following the intuitive approach, the first 20 of those 50 phones would have 6 CCS of traffic; the next 15 would have 5 CCS, and so on, for a total of 250 CCS. This is spread across 50 lines. To calculate the average *Traffic per Line (formula 20)*, the traffic is summed with a weighted average.

Formula (20) ***Traffic per Line = Σ (Traffic per Line * Market Penetration) / (First Line Penetration * Lines per Home)***

Result: CCS per Line HDBH = (6*0.20 + 5*0.15 + 4*0.10 + 3*0.05) / (0.20 * 2.5) = 5 CCS
Result: Erlangs per Line HDBH = 5 / 36 = 0.139 erlangs
Result: CCS per Line ABSBH = (5*0.20 + 4*0.15 + 3*0.10 + 2*0.04) / (0.20 * 2.5) = 4 CCS
Result: Erlangs per Line ABSBH = 4 / 36 = 0.11 erlangs

The same result can be calculated using *Nth Line Relative Penetration (formula 17)*.

Formula (21) ***Traffic per Line = Σ (Traffic per Line * Relative Penetration) / Lines per Home***

Result: Traffic per Line = (6 * 100% + 5 * 75% + 4 * 50% + 3 * 25%) / 2.5 = 5 CCS

If the traffic has to deal with different codec types, the results from this formula could be split up. In an isolated example, if lines one and two were G.728 (16 kbps) and lines three and four were G.711 (64 kbps), then the percentage of traffic attributed to each codec type follows:

Formula (22) ***Codec Share of Traffic = Σ (Traffic per Line * Relative Penetration per Codec) / (Lines per HHP * Traffic per Line)***

Result: G.728 Share of Traffic = (6 * 100% + 5 * 75%) / (2.5 * 5) = 3.9/(2.5 * 5) = 78%
Result: G.711 Share of Traffic = (4 * 50% + 3 * 25%) / (2.5 * 5) = 1.1/(2.5 * 5) = 22%

3.3.2 Connectivity

The total number of lines of voice follows:

Formula (23) ***Lines per Node = HHP per Node * Market Penetration of Voice * Lines per HHP***

Result: Lines per HE = 100,000 * 20% * 2.5 = 50,000
Result: Lines per Hub = 20,000 * 20% * 2.5 = 10,000
Result: Lines per FN = 500 * 20% * 2.5 = 250

The total network traffic follows:

Formula (24) ***Traffic per Node = Lines per Node * Traffic per Line***

Result: Erlangs per HE HDBH = 50,000 * 0.139 = 6944
Result: Erlangs per Hub HDBH = 20,000 * 0.139 = 1389
Result: Erlangs per FN HDBH = 250 * 0.139 = 35
Result: Erlangs per HE ABSBH = 50,000 * 0.11 = 5556
Result: Erlangs per Hub ABSBH = 20,000 * 0.11 = 1111
Result: Erlangs per FN ABSBH = 250 * 0.11 = 28

This voice traffic comes from VoIP ports on DOCSIS cable modems that are the equivalent to subscriber lines. The unsolicited grant service (UGS) scheduled connection slots in the DOCSIS upstream act as trunks in which the subscriber-line traffic may be concentrated into. The number of connections required is calculated by the erlang B formula for both HDBH and ABSBH.[1]

Formula (25) ***Connections per FN = Erlang-BL (Erlangs per Node, Grade of Service)***

Result: Connections per FN HDBH = Erlang-BL(35, 1%) = 47
Result: Connections per FN ABSBH = Erlang-BL(28, 0.5%) = 41

One of these two connection numbers must be carried forward. HDBH is usually higher than ABSBH. A conservative design will use HDBH, whereas a more relaxed design will use ABSBH. Technically, both should be evaluated, and the worst case carried forward. In this example, the worst case will be carried forward.

As a sanity check, the ratio of the number of lines of voice to the number of trunk connections can be calculated:

Formula (26) ***Line-to-Trunk Ratio = Lines per FN / Connections per FN***

Result: Line-to-Trunk Ratio = 250 / 47 = 5.3

3.3.3 Multiple Codecs

The table in Appendix D: “Voice-over-IP Call Density Tables” on page 68 provides the optimized number of connections supported in an upstream. These values have been calculated assuming all PHY and MAC overhead is included. The concept that this white paper uses is to start with the theoretical maximum number of voice connections, and then de-rate this number to allow for data and signaling traffic. If only one codec type is active, such as G.711, then these numbers can be used directly. If multiple codec types are active, then the connection numbers need to be scaled in proportion to their usage.

For example, a CM might have a limitation that it can support compression only on the first two lines. Thus lines 1 and 2 might be G.728, and lines 3 and 4 might be G.711. Another operation assumption may be that line 3 is always G.711 fax or G.711 analog modem.

Formula (27) ***Connections per Upstream Max = Floor [∑ (Connections per Upstream per Codec * Codec Share of Traffic)]***

Result: Connections per Upstream = Floor [23 * (100%)] = 23

The formula allows for multiple codecs, although the example is showing a calculation for a G.711 system only.

The required raw upstream bandwidth required to support the maximum number of voice connections follows:

Formula (28) ***BW Voice = Bandwidth per Upstream * Connections per FN / Connections per Upstream***

Result: BW Voice = 2.56 Mbps * 47 / 23 = 5.1 Mbps

1. The use of the erlang B formula assumes that the bandwidth reserved for voice in the DOCSIS upstream is 100-percent available. If these assumptions are not true, an additional correction factor or an alternative probability distribution may have to be applied.

Note that for data calculations, the formulas use an approximation for the PHY and MAC overhead that converts raw bandwidth to usable bandwidth, and then traffic formulas are applied to the results. For voice, the packet sizes are constant and so the PHY and MAC overhead is very predictable. So for the best optimization, the voice calculations work with raw bandwidth, and the PHY and MAC overhead are accounted for in the connection density calculations done in Appendix D: "Voice-over-IP Call Density Tables" on page 68.

As an isolated example, these formulas will resolve when different codecs are being used. For a 1280-ksps, 2.5-Mbps, 1.6-MHz, QPSK upstream running header suppression and G.711, the number of connections that can be mapped into an upstream is 23. For G.728, the number of connections is 50. As an isolated example, if the first two lines were G.728 and the last two lines were G.711, the connection number would be as follows:

Result: Connections per Upstream = Floor [50 * 78% + 23 * 22%] = Floor [39 + 4.06] = 43

Result: BW Voice = 2.56 Mbps * 47 * (78% / 50 + 22% / 23) = 2.5 * 1.2 = 3 Mbps

For now, we will assume the downstream has a similar percentage of PHY overhead, so this number is used for both the upstream and downstream.

3.4 Video

The HFC network has many new as well as many existing video sources. After all, the HFC network was originally developed to carry video. This section examines the current and upcoming video sources and how they impact network design and planning for video, with a focus on IP-based video services.

3.4.1 Analog Video

Analog video is the earliest and still most common way of transporting video down an HFC plant. In the United States, by convention, analog video is placed in the spectrum of 54 to 550 MHz. This range is broken up into radio frequency (RF) channels of 6 MHz each, with a channel designation from 2 to 78. These channels are all broadcast channels.

3.4.2 Digital Video

Digital video is typically MPEG encoded with 6 to 12 MPEG streams that are packed into one 6-MHz RF channel spacing. In the United States, by convention, the RF channels that carry digital video are placed in the 550–750 MHz spectrum. Within this 200-MHz spectrum, there is room for 33 RF channels. At 9 MPEG streams per RF channel, this spectrum has the capacity to handle about 300 digital video channels.

A key design feature of digital video is the number of MPEG streams that can be packed into a 6-MHz channel. An example system design would comprise 9 MPEG video encoders, which then fed a multiplexer that packed the channels into a 6-MHz channel bandwidth. With the introduction of statistical multiplexing (statmux), multiple video programs sharing the fixed network bandwidth each get their own portion of bandwidth based on the content of each channel.

There are two fundamental modes of operation. The first is a "closed-loop" scenario where the MPEG multiplexer monitors and allocates the appropriate bandwidth to every participant by feeding back signaling to the MPEG encoders to increase or decrease their coding rate based upon available channel bandwidth. This model can be viewed as a "controlled" statistical multiplexing model since the MPEG multiplexer essentially controls the rate of packet generation from each different input source so that no packets are dropped and no "extra" NULL packet can be generated. This model is preferred for fixed broadcast systems.

The second mode of operation is an "open-loop" scenario where the statmux cannot feed signaling back to the encoders. This scenario occurs when the encoders and multiplexers have different proprietary interfaces, or they are not colocated. In this mode, the MPEG multiplexer must be more intelligent and transcode the signals, a process that involves re-coding and re-time-stamping the video signals that require adjustments to their bandwidths.

In addition to just combining the MPEG signals, the MPEG multiplexer acts as a "quality equalizer." In other words, when all the incoming video streams have highly dynamic content, overall quality is reduced. On the other hand, when they all have static scenes, overall quality can be very good. However, there will be neither packet drops nor extra inserted NULL packets.

Since digital video uses QAM modulation techniques similar to those that DOCSIS uses, this area of spectrum is a good place to put the DOCSIS downstreams. From a viewpoint, this spectrum will need to be divided between digital video and DOCSIS data/voice/video services. Again, this service is broadcast only.

3.4.3 MPEG Video on Demand

Instead of dedicating all the digital video channels to broadcast, they could be used to deliver personalized content or video on demand (VOD) to a consumer. The transmission and spectrum usage is similar to standard digital video. With a larger number of video sources to choose from and more unique multiplexing nodes, the preferred operation model is the "open-loop" model described previously.

From a traffic engineering viewpoint, the packing density of MPEG streams per RF channel may be different for MPEG VOD than for MPEG digital video since the equipment may also be different. The amount of RF spectrum required to support a VOD system could be calculated using the same approach that was described for IP unicast video streaming.

MPEG VOD is intended to be a point-to-point delivery system, although it may be broadcast in implementation on some links because of the granularity of the network. It also requires a two-way media so that users can select their programming material.

3.4.4 Video Telephony and Videoconferencing

Video telephony has long been considered, but slow to be adopted. The main obstacle has been lack of bandwidth, a problem soon to be solved with broadband data-over-HFC plants. Traffic engineering for video telephony should use a similar approach as VoIP traffic engineering, with adjusted penetration rates and with appropriate codecs, such as H.263 at 384 kbps. Call holding times for video telephony may initially be longer than standard telephony calls because of the uniqueness of the call, but should converge since the application of one-to-one conversation is the same.

Videoconferencing is usually performed by bringing all video sessions back to a common video conference server. From an access network viewpoint, the same amount and type of bandwidth is required. There may be extra data bandwidth required to support a "white board" application where all users share a view to a common data application. From a backbone network viewpoint, there will be heavier concentrations of video calls at the videoconference server rather than a more random distribution across the network. The networks surrounding the videoconference server should be engineered to meet or exceed the videoconference server requirements. Call holding times will be longer than standard video telephony calls because videoconferences are usually associated with meetings that last one or more hours.

3.4.5 IP Video Streaming

IP video streaming (abbreviated in this white paper as IPVS) became popular with the emergence of personal computer applications such as Real Player, Windows Media Player, QuickTime, and a host of other programs. IP video streaming can be any of the following:

- Real time or non-real time
- Constant bit rate (CBR) or variable bit rate (VBR)
- Unicast or multicast

A current example of video streaming was the Net Aid benefit concert that Cisco Systems and the United Nations hosted at http://www.netaid.org in 1999. The Net Aid Web site handled over 2.4 million simultaneous video streams on its inaugural day.

3.4.5.1 Real Time vs. Non-Real Time

In real time, video is recorded, compressed, and immediately sent out in IP. In non-real time, video is recorded, compressed, and stored as a file on a video server. When it is played back, it is converted to IP packets. The transmission formats for real time and non-real time do not have to be different if the viewing bit rate is the same. This choice only impacts network traffic engineering if the subscription rates differ, although it will impact the choice and location of video servers.

3.4.5.2 CBR vs. VBR

Simple video transport and uncompressed video can be CBR, but most if not all video-compression schemes produce a VBR output. For example, when a sports anchor on a news cast is talking, very few of the pixels on the screen are changing. However, when the anchor cuts over to a scoring goal in a football game, all the pixels on the screen are changing, and changing rapidly. Uncompressed video would transport the pixels at a constant rate, whereas video compression would transmit changes only in pixels, resulting in a bandwidth that is variable.

The challenge comes in combining these streams. What if 10 IP video streams are combined into one channel group and each has a scoring goal; will there be packet drops? What if all of them have news anchors; will there be wasted bandwidth?

With standard digital MPEG video, this problem is addressed by combining video streams within a group that have different anticipated action levels (such as the Garden Channel and the Movie Channel) and then using a MPEG multiplexer that adjusts the compression if the combined peaks of the streams exceed the channel bandwidth. If simultaneous peaks last too long, then there will be some sort of reduction of quality to the video signals introduced in order to get all the signals through.

With IP video streaming, the receiver has a deeper video buffer that allows for the IP video packets to tolerate more jitter in their transport. Thus when there is a simultaneous peak across video streams, the packets will get naturally statistically multiplexed, some will get delayed, and then they will get through. This scenario will allow short video peaks to not get clipped. For longer, sustained peaks of video bandwidth on multiple channels, this method is insufficient, so the two choices of system design would be as follows:

- A video multiplexer operating in an “open-loop” transcoding mode would efficiently pack the video streams in the same manner as MPEG VOD is managed.
- Use a less dense packing of IP video streams so that all simultaneous peaks in bandwidth would get through, and then mix in lower-priority traffic such as IP data that can make use of the unused bandwidth when there are no simultaneous peaks of bandwidth demands.

The trick in both cases is to not overload the bandwidth of a channel group with video streams. With IP video streaming, IP data can be placed on the same channel at a lower priority. The IP video stream loading on the transport channel is kept limited so that simultaneous peaks cause data to be delayed, but they themselves never oversubscribe the channel. Then video streams are neither dropped nor delayed.

For traffic planning purposes, it is convenient to use a typical bit rate to calculate densities of video streams within a channel. For quality of service (QoS) calculations, it may be prudent to limit the number of video streams per channel to prevent packet drops. As with MPEG digital video, the key traffic parameter is how many IP video streams will fit into a RF channel.

IP video streaming is a two-way application, with video flowing in one direction and control information flowing in the other direction. It may also exist in either direction on the HFC plant. Downstream applications would include broadcast viewing or VOD. Upstream applications include users who wish to originate content into the Internet such as high school soccer games or even “nanny cams” in the home.

3.4.5.3 Unicast and Multicast

Unicast refers to the generation and delivery of a video stream from one source to one destination. Unicast video streaming simulates VOD in an IP environment.

Multicast refers to an IP feature for the generation and delivery of a video stream from one source to multiple destinations. Multicast video streaming simulates broadcast video in an IP environment.

The significance of multicast video is that packets are duplicated in the network only when they need to traverse separate paths. For example, if ten subscribers across two different HFC downstreams are watching the same multicast session, one video stream will leave the video server. The upstream router from the two target CMTSs will duplicate the stream and send it to both CMTSs. Each CMTS will send only one copy of the video packet downstream that will then be received by each of the five subscribers.

3.4.5.4 Encoding Rates

Low-bit-rate video streaming on the PC works but has poor quality at bits rates from 16 to 128 kbps. These speeds were chosen to map into analog modems and ISDN lines. In HFC environments, as long as the video server and the video terminal will handle it, the rates are likely to be higher. Small windows on a PC may typically run at 384 kbps using H.263 type encoding. Since the bandwidth used is ultimately related to the video-compression technology and not the transport technology, full-screen, broadcast-quality IP video streaming can be 4 to 7 Mbps, just as it is with standard digital video signals. When evaluating video codecs, it is important to look not only at bit rate, but at the frame rate and resolution as well. Some typical video compression technologies and their compression rates are listed in Table 3.

Table 3 Typical Video Compression Rates

Video Compression Technology	Application	Typical Bit Rates
Real Video, Windows Media Player, QuickTime	Proprietary formats for the PC and Apple environments; bandwidths have been chosen to map to available analog modem and ISDN bandwidths; generally poor quality with choppy video but very popular; initially appropriate for small windows only, but likely to improve as bandwidth expands	From 16 to 128 kbps, as well as up to 1 Mbps
H.263	1/16 screen resolution	128 kbps
	Quarter screen resolution	384 kbps
MPEG 2 Encoding	Quarter screen resolution	1.5 Mbps
	Half screen resolution	3.0 Mbps
	Full screen resolution	3.8 to 7 Mbps
MPEG 4	The frame size for MPEG4 is CIF (320x240), 240x180, or 160x120 (or QCIF).	Up to 3 Mbps

It is worth noting that the MPEG standard has both an MPEG video-encoding layer and a separate MPEG-TS transport layer, meaning that systems can mix and match layers, based upon the technical and economic needs of the system. With MPEG and IP, the combinations with typical examples follow:

- MPEG video over MPEG transport (digital video)
- MPEG video over IP transport (video streaming)
- IP data over MPEG transport (DOCSIS)
- IP data over IP transport (Ethernet)

All four of these technologies are deployed today. Since the technologies are layered, these technologies should be interchangeable. So a source that starts off as MPEG video over MPEG transport could be converted to MPEG over IP, transported across an IP intranet, converted back to an MPEG transport across the HFC plant, and delivered to an MPEG terminal. At that point, it might even get converted back to MPEG over IP for distribution within a home network.

From a practical point of view, these choices represent technological choices. The final choice is based what can be built and deployed with acceptable economics, quality, and time frames. The choice of using transport depends upon the availability and cost of the video servers, coders, transport facilities, and the choice and design of the endpoint.

3.4.5.5 Calculations

The following formulas calculate the bandwidth required to support a network with a mix of unicast and multicast video streaming.

The first step is to calculate the number of users in a serving area, which is taken to be an integral number of *Fiber Nodes per IPVS Downstream Group (template)* times the *HHP per Fiber Node (template)*. The *IPVS Market Penetration (template)* indicates the number of HHP subscribed and the *IPVS Activity Factor (template)* indicates the number of subscribed households that are currently watching video streaming. A user in this context is a single video stream. If one person is using two video streams, this context refers to that as two users. There may be multiple video streams per household.

Formula (29) ***IPVS Users = HHP per FN * FN per IPVS Downstream Group* IPVS Market Penetration * IPVS Activity Factor * IPVS Sessions per Home***

Result: IPVS Users = 500 * 2 * 40% * 60% * 1.3 = 312

Of the 100-percent video streams, there will be a split in usage between multicast and unicast. For unicast, there will be one session per user.

Formula (30) ***IPVS Unicast Sessions = IPVS Users * IPVS Unicast Usage * 1 User per Unicast Session***

Result: IPVS Unicast = 312 * 70% * 1 = 218

For multicast, there should be fewer sessions than multicast users. Before applying de-rating, the number of *IPVS Multicast Users (formula 31)* is calculated.

Formula (31) ***IPVS Multicast Users = IPVS Users * IPVS Multicast Usage***

Result: IPVS Multicast Users = 312 * 30% = 94

The number of available multicast sessions will appear as program selections. For a large program selection, it is likely that all the programs would not be simultaneously used within a large serving area. For example, if there were one United States news channel, it could be heavily subscribed in a United States neighborhood. However, it there were one news channel for every single country in the world, say 200 news channels, then it is likely that within a reasonably large area, only five to ten news channels would ever get used at once.

So the *IPVS Multicast Used Max (template)* parameter sets an upper limit on the maximum realistic usable multicast sessions in a large geographical area (the size of this area and the following de-rating formula are discussed in the section "Derivation of Multicast De-rating Factor" on page 29), and allows the available number of multicast sessions to be much higher. In practical terms, the *IPVS Multicast Sessions Max (formula 32)* is for traffic engineering purposes, and *IPVS Multicast Available (template)* is for programming and marketing purposes.

Formula (32) ***IPVS Multicast Sessions Max = IPVS Multicast Available * IPVS Multicast Used Max***

Result: IPVS Multicast Used Max = 125 * 80% = 100

If the serving area is less than ten times the size of *IPVS Multicast Sessions Max (formula 32)*, than a further de-rating can be applied.

With:

- U = *IPVS Multicast Users (formula 31)*
- M = *IPVS Multicast Sessions Max (formula 32)*
- F = *IPVS Multicast Shaping Factor (template)*

Formula (33)

$$\textit{IPVS Multicast De-rating Factor} = \frac{Min\left[U, Ceiling\left[M \times \left\{1 - exp\left(-\frac{U}{M} \times F\right)\right\}\right]\right]}{M}$$

Result:

$$\text{IPVS Multicast De-rating Factor} = \frac{\text{Min}\left[94, \text{Ceiling}\left[100 \times \left\{1 - \exp\left(-\frac{94}{100} \times 1\right)\right\}\right]\right]}{100}$$

Now with the final de-rating factor, the number of *IPVS Multicast Sessions (formula 34)* can be calculated.

Formula (34) ***IPVS Multicast Sessions = IPVS Multicast Sessions Max * IPVS Multicast De-rating Factor***

Result: IPVS Multicast Sessions = 100 * 61% = 61

The separate *IPVS Unicast Sessions (formula 30)* and *IPVS Multicast Sessions (formula 34)* are now summed to get the total number of simultaneous *IPVS Sessions (formula 35)*.

Formula (35) ***IPVS Sessions = Ceiling [IPVS Unicast Sessions + IPVS Multicast Sessions]***

Result: IPVS = Ceiling[218 + 61] = 279

For backhaul purposes, that number of sessions can be directly translated to bandwidth.

Formula (36) ***IPVS Network BW = IPVS Sessions * BW per IPVS Stream***

Result: BW IPVS = 279 * 1.5 Mbps = 419 Mbps

But for calculating the number of downstreams needed, round-off errors with fitting video streams into RF channels may occur, so first the video streaming capacity of the RF channel is calculated.

Formula (37) ***IPVS per RF Channel = Floor [BW per RF Channel / BW per IPVS Session]***

Result: IPVS per Channel = Floor [27/1.5] = 18

Then the overall number of sessions requiring support is divided by the number of sessions supported per channel to determine the number of RF channels required.

Formula (38) ***IPVS RF Channels = Ceiling [IPVS Sessions / IPVS per RF Channel]***

Result: IPVS Downstream Channels = Ceiling [279 / 18] = 16

3.4.5.6 Derivation of Multicast De-rating Factor

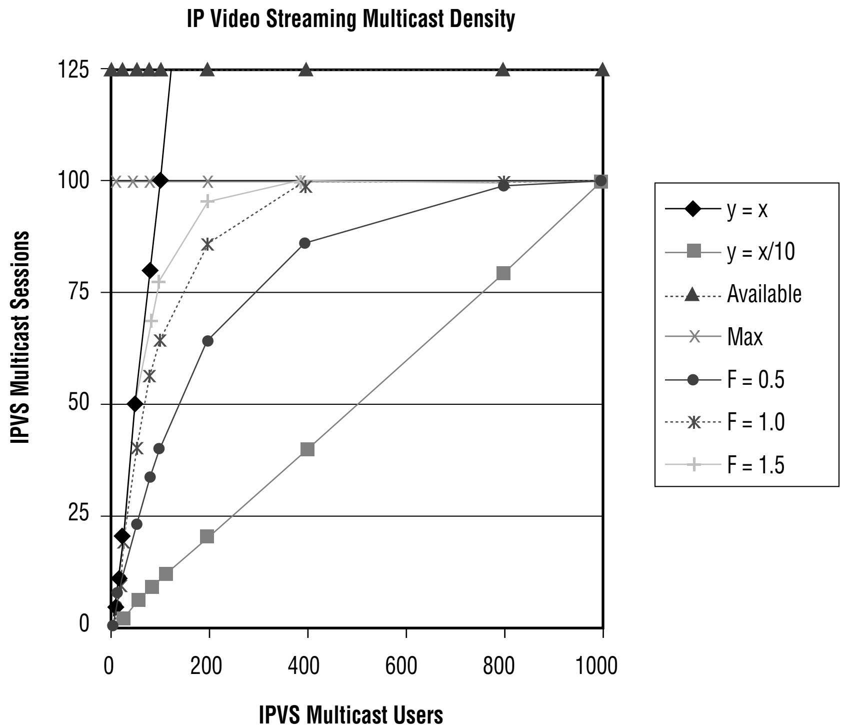
Subscribers must join and leave multicast sessions. The acts of joining and leaving cause the multicast session to be reserved across the network, an occurrence that is significant to traffic engineering. Say a network has 100 multicast sessions available. If there were only one user on the network, only one multicast session would be transported. If there were 10 users, and some watched the same channel, then maybe only 7 sessions may be active. If there were 1000 users, most of them might be tuned to the top ten channels, with some of them watching the other channels, so there might only be 40 multicast sessions, even though all users have a choice to watch all 100 channels. In a classic broadcast system, the 100 channels would always be broadcast and always be using network bandwidth, even if no one was watching.

The differences are interesting. From one point of view, if the subscriber traffic remains constant, then the number of choices of multicast video streams can be increased without having to linearly increase the network bandwidth. The only impact that would be seen is that the additional video streams may be interesting enough to reduce the simultaneous viewing of the top ten channels and thus increase the number of simultaneous video streams. To continue with the previous example, if 100 available multicast streams generated 40 multicast sessions, then 120 available multicast sessions might generate only 45 multicast sessions.

These observations can be defined and placed on a graph of *IPVS Multicast Sessions (formula 34)* on the y-axis versus *IPVS Multicast Users (formula 31)* on the x-axis. Refer to Figure 6.

- A fixed number of multicast sessions are available. This number is referred to in this white paper as *IPVS Multicast Available (template)*. In this example, that line is y = 125.
- A maximum number of channels are anticipated to be used at any one time. This situation is referred to in this white paper as *IPVS Multicast Sessions Max (formula 32)*. In this example, that line is y = 100.
- The number of sessions can never exceed the number of users. That line is y = x.
- The law of large numbers: For a low number of users and a low number of sessions, the number of users will be close to the number of sessions. For a large number of users and a large number of sessions, there will be more usage of favorite channels, and the number of sessions will be much less than the number of users. That line is *IPVS Multicast Sessions (formula 35)*, and it has an exponential shape.
- If the number of users far exceeds the number of the maximum number of sessions, then the number of multicast sessions approaches *IPVS Multicast Sessions Max (formula 32)*.

Figure 6 Multicast De-rating Curve



The formula for this line follows:

$$\textit{IPVS Multicast De-rating Factor} = \frac{Min\left[U, Ceiling\left[M \times \left\{1 - exp\left(-\frac{U}{M} \times F\right)\right\}\right]\right]}{M}$$

Where:

- U = *IPVS Multicast Users (formula 31)*
- M = *IPVS Multicast Sessions Max (formula 32)*
- F = *IPVS Multicast Shaping Factor (template)*

The ceiling function enforces the fact that there are never fractional sessions. It also forces the exponential to the maximum value in a fixed amount of time instead of an infinite amount of time.

The factors that determine the shape of this line are the shaping factor (F) and the maximum number of sessions (M). Figure 6 shows the relationship of F, U, and M. With F = 1, the curve is a classic exponential with a ceiling function applied and has the following properties:

- When the number of users (U) equals the maximum number of sessions (M), the number of sessions is 63 percent of M.
- When the number of users (U) is greater than 4.6 times the maximum number of sessions (M), the number of sessions required is equal to the maximum number of sessions (M).

To further understand Table 4, take the example of a network design with F = 5 and a serving area that had 1000 users (U).

- If 1000 users had 1000 maximum usable video streams that they could watch, the network would need a capacity of 390 sessions (from formula 34).
- A 1000 users would require 1000 sessions if the maximum usable video streams that they could watch exceeded 9.2 x 1000 = 9200.

By keeping track of the usage statistics in a network, appropriate values of F and M may be chosen. A single network statistic that measures the number of sessions open can then be matched with the maximum sessions available (M) and the number of users (U) to determine the shaping factor (F). Values of F < 1 would be considered an aggressive network design, and values of F > 1 would be considered a conservative network design.

Table 4 Relationship between F, U, and M

F	At U = M, the De-rating Factor is:	At a De-rating of 100%, U / M is:
0.5	39%	9.2
0.6	45%	7.7
0.7	50%	6.6
0.8	55%	5.8
0.9	59%	5.1
1.0	63%	4.6
1.1	67%	4.2
1.2	70%	3.8
1.3	73%	3.5
1.4	75%	3.3
1.5	78%	3.1
1.6	80%	2.9
1.7	82%	2.7
1.8	83%	2.6
1.9	85%	2.4
2.0	86%	2.3

3.5 Maintenance and Signaling

The maintenance and signaling budget covers overhead from DOCSIS, PacketCable, and other signaling protocols. Note that this number is per cable modem. Voice cable modems have more signaling messages than data-only cable modems, but data-only cable modems have contention overhead, while voice-only cable modems do not. More detailed budgets could be done by considering actual voice and data activity. However, an approximate estimate is sufficient.

The bandwidth required follows:

Formula (39) ***BW M&S = HHP per FN * Total Penetration * Overhead per Modem***

Result: BW M&S per FN = 500 * 25% * 1 kbps = 125 kbps

4 HFC Equipment Sizing

4.1 Equipment Sizing Based upon Plant Topology

If the upstream fiber from the fiber node is terminated at the hub, then the CMTS must be located at the hub. The minimum number of CMTSs is thus equal to the number of hubs:

Formula (40) ***CMTS per HE Min = Hubs per HE * 1 CMTS per Hub***

Result: CMTS per HE Min = 5 * 1 = 5

The maximum households passed per CMTS as limited by the plant follows:

Formula (41) ***HHP per CMTS Max = HHP per HE / CMTS per HE Min***

Result: HHP per CMTS Max = 100,000 / 5 = 20,000

In practice, the number of HHP, fiber nodes, and hubs varies with geography, so calculations must be done per hub. This paper assumes that the number of homes passed per fiber node and hub remains consistent to get an approximate view of a larger network.

4.2 Equipment Sizing Based upon Upstream Combining

The previous section established that at least one CMTS is required per hub. However, it did not address how many line cards are required to support the upstream connectivity.

When looking at upstream density, the two factors consist of:

- The number of households passed
- The number of fiber nodes

If the number of households passed per fiber node is constant, than the number of fiber nodes becomes the dominant concern. If the fiber node sizes vary considerably, then the number of households passed becomes the dominant concern. Further, the more the number of HHP, the more ingress noise from the plant may exist. However, the more the fiber nodes combined, the higher the noise floor will be from the upstream lasers.

The maximum combining ratio for fiber nodes is typically 3 to 4, while the maximum number of households passed per fiber node is about 2000. However, these limits will vary from plant to plant, and should be considered variables. In practice, the HFC plant should be analyzed on a per-fiber node basis with a spectrum analyzer and be combined with other fiber nodes in combinations that meet the noise floor requirements set forth by the HFC plant operator.

Further, 16 QAM upstream modulation requires a quieter plant with less noise than does QPSK, so 16 QAM systems will have a much lower combining tolerance than QPSK systems.

The natural topology migration of 500 households passed per fiber node system follows three physical topologies, each with bandwidth expansion options under software control. Example numbers are included.

- *Low penetration using combined upstreams:* Build the plant with 4:1 upstream combining of fiber node per upstream receiver on a CMTS. The system runs a baseline case of 1280 ksps, QPSK, 2.5 Mbps, and 1.6 MHz per upstream. As bandwidth demands increase, there is a software option to double the bandwidth of the upstream channels (1280 ksps vs. 2560 ksps), an option that would not require a physical plant upgrade.
- *Moderate penetration using direct connection:* The next implementation step is to reduce combining down to one fiber node per upstream. At this point, the number of HHP should be low enough to allow a software upgrade from QPSK to 16 QAM for bandwidth doubling, in addition to the bandwidth doubling that can be gained by doubling the symbol rate.
- *High penetration using split upstreams:* The final stage would be a 1:3 or similar splitting of upstreams across CMTS receivers. Each CMTS receiver would operate at a different frequency within the one upstream. As before, each receiver could have its bandwidth quadrupled under software control.

For the system with low customer penetration, upstream combining ratios—not network traffic—will dictate the amount of equipment required. Noise limits the upstream combining ratios. Two limits are involved in upstream combining. The first is ingress noise. On a clean plant, ingress noise comes primarily from subscribers' homes, and is, therefore, proportional to homes passed. The maximum number of homes passed divided by the homes passed per fiber node will indicate the maximum combining ratio. The second limitation is noise floor of the upstream lasers. Depending upon the quality of the lasers, there will be a limit to how many can be combined.

The following formula calculates the maximum number of fiber nodes per upstream receiver, first considering the noise floor from the HHP, and then considering the noise floor from the fiber node lasers.

Formula (42) ***FN per Receiver Max 1 = Min [Floor (HHP per Upstream Limit /HHP per FN), FN per Upstream Limit]***

Result: FN per Upstream Max 1 = Min [Floor (2,000/500), 4] = 4

4.3 Equipment Sizing Based upon Bandwidth

4.3.1 Upstream Connectivity—A Theoretical Approach

The fundamental approach is to establish how many fiber nodes are serviced by how many upstream receivers. This physical connectivity will cause a rounding error. For example, if 1.5 upstream receivers are required to service one fiber node, then two upstream receivers are required, resulting in 33-percent extra resources. The fact that no two fiber nodes are alike makes the modeling accuracy even worse.

One approach would be to take the individual bandwidth contributions from each source, sum them, and divide by the bandwidth per upstream. If applications such as video are handled on separate downstreams or upstreams, they may be calculated separately.

Formula (43) ***BW Total = BW Data + BW Voice + BW Video + BW M&S***

Result: BW Total Downstream = 1.1 + 5.1 + 0 + 0.125 = 6.5 Mbps downstream per FN

Result: BW Total Upstream = 0.37 + 5.1 + 0 + 0.125 = 5.7 Mbps upstream per FN

In the high market penetration scenario, the bandwidth required is greater than the provisioned upstream receiver bandwidth, and multiple upstream receivers will be required per fiber node. The upstream path is thus split across multiple upstream receivers. In the context used here, the term "upstream" alone is synonymous with "upstream receiver."

- For fiber node splitting

Formula (44) ***Receivers per FN Min = Ceiling(BW Total Upstream / BW per Receiver)***

Result: Upstreams per FN Min = Ceiling(5.7 / 2.5) = Ceiling (2.3) = 3

In the low market penetration scenario, the bandwidth required is much less than the provisioned upstream receiver bandwidth, so combining may be used to allow multiple fiber nodes per upstream receiver. This combining must not exceed the maximum allowed households passed per fiber node for noise considerations. For example, if data and signaling traffic were the only consideration for the upstream path, then:

- For fiber node combining

Formula (45) ***FN per Receiver Max 2 = Floor (BW per Receiver / BW Total Upstream)***

Result: FN per Upstream Max 2 = Floor(2.5 / 0.5) = 5

Now taking the minimum of the two limits,

- For fiber node combining

Formula (46) ***FN per Receiver = Min [FN per Receiver Max 1, FN per Receiver Max 2]***

Result: FN per Upstream = Min [4, 5] = 4

While good in theory, this approach overlooks the coexistence of voice and data. Since a CM only has one upstream transmitter, if it supports both voice and data services, then it must do so on the same upstream channel, meaning that data and voice connections cannot be broken up and assigned to different physical resources, such as different upstreams and downstreams.

Therefore, as far as balancing voice and data connections across multiple upstreams is concerned, one can be optimized (most likely voice), while the other will be unoptimized. This point is a significant one. *If voice calls from a single fiber node are evenly distributed across multiple upstream receivers, data calls will not be. If data calls are evenly distributed, the voice calls will not be.* The rounding up in formula 44 in part covers for this inefficiency.

4.3.2 Upstream Connectivity—A Practical Approach

A more practical approach is for each upstream and downstream to have a reserved amount of bandwidth for voice, and a reserved amount of bandwidth for data.

4.3.2.1 Maximum Voice Density

A minimum channel bandwidth can be determined for data. If voice is assumed to be constant, then in order to be able to meet the performance requirement, there has to be enough bandwidth in the rest of the channel to handle the maximum data rate. The phrase “data region” refers to the percentage of the upstream bandwidth that is used for all data.

Formula (47)	***Data Region Min = Max Data Rate / Usable Channel BW***
Result:	Data Region Min Upstream = 0.768 / (2.2) = 35%
Result:	Data Region Min Downstream = 2 / 27 = 7.4%

The DOCSIS protocol contains an initial maintenance slot that is used for ranging new cable modems. This slot is, by specification, 2 ms long. If the upstream were 100-percent provisioned for voice, then the scheduling of a maintenance slot would add 2 ms of jitter. To avoid this additional jitter, voice should be provisioned less than 100 percent. If we assume that the voice samples are repeated on a frame basis, where the frame time is equal to the packet recording length, the amount of bandwidth required for this maintenance slot follows:

Formula (48)	***Maintenance Region = Maintenance Time / Voice Frame Time***
Result:	Maintenance Region = 2 ms / 10 ms = 20% (for 10 ms frame time)
Result:	Maintenance Region = 2 ms / 20 ms = 10% (for 20 ms frame time)

The voice density must be limited to allow enough room for data and initial maintenance slots. Assuming that the bandwidth for data and initial maintenance overlap, the room left for voice follows:

Formula (49)	***Voice Region Target = 100% – Max (Data Region Min, Maintenance Region)***
Result:	Voice Region = 100% – Max(35%, 20%) = 65%

The phrase “voice region” refers to the target percentage of the upstream bandwidth that is used for all voice traffic. Each system design should choose its own maximum voice region, depending upon the VoIP frame size and relative number of voice and data users. The maximum voice region is chosen to guarantee that data will not become bandwidth starved.

4.3.2.2 Matching Fiber Nodes to the CMTS Receivers

A strict operating range would be to use the voice region as an absolute maximum. A more practical maximum would be to add plus or minus 10 percent error margin to the voice region. This margin recognizes that the numbers used are all based upon assumptions and averages, allowing for real-world compromises when mapping fiber nodes to the CMTS. By allowing for a range of error, cost and business consideration can be factored in.

With a subset of the upstream identified for voice, we can now map the number of required connections into it.

Formula (50) ***Receivers per FN = Ceiling [Connections per FN / (Connections per Upstream * Voice Region Target * (1 + Voice Region Allowance))]***

Result: Receivers per FN Nominal = Ceiling [47 / (23 * 65%*(100% – 0%))] = Ceiling (3.1) = 4

Result: Receivers per FN Adjusted = Ceiling [47 / (23 * 65%*(100% + 10%))] = Ceiling (2.7) = 3

A judgment call is required here. One of the factors to now consider if multiple upstream receivers are required is in what multiple are the upstreams available, and how does it impact cost and packing densities.

Another option to consider is increasing the upstream symbol rate or modulation rate (bits per symbol). Increasing the symbol rate increases the size of the RF spectrum, making the upstream spectrum more susceptible to overlapping a noise source. Increasing the modulation rate increases the noise sensitivity of the signal and thus the need for a higher carrier-to-noise ratio. Either of these options also increases the packets-per-second performance requirements of the CMTS. If the RF performance and the CMTS switching performance can be accomplished, the more cost-effective decision would be to increase the receiver bandwidth instead of using more than one receiver.

4.3.3 Downstream Connectivity

Downstream connectivity can be approached in the same manner. By definition, a DOCSIS domain may contain not only multiple upstreams, but also multiple downstreams. The number of fiber nodes that a CMTS transmitter can support is equal to the bandwidth of the transmitter divided by the bandwidth required.

Formula (51) ***FN per Transmitter Max = Floor(BW per Downstream Raw / BW Total Downstream)***

Result: FN per Transmitter Max = Floor(30 Mbps / 6.5 Mbps) = 4

4.3.3.1 Adjustment for Voice Activity Detection

Voice Activity Detection (VAD), also known as silence suppression, significantly reduces the traffic associated with a connection. VAD does this by transmitting voice packets only when the energy of the voice signal is above a set threshold. If the receiving end does not have any packets to play out, it substitutes comfort noise based upon measured background noise. During the silence interval, the transmitting end will send out keepalive Real-Time Transport Protocol (RTP) packets to help keep the connection alive.

If we assume that only one of the parties is talking at one time, then the total bidirectional traffic is already reduced by 50 percent. And with a pause between both parties talking, a more realistic VAD activity level drops about 40 percent. That means that one person is talking for more than 40 percent of the time and is silent 60 percent of the time.

This value of 40 percent is for a two-way conversation. The DOCSIS upstream represents a collection of one-way conversations. Since these conversations do not collide, they are far more likely to overlap. So for a low number of one-way conversations, the activity factor will be much higher than 40 percent. As the number of voice conversations approaches large numbers, the activity level will decrease and get close but not equal to the *VAD Activity Factor Two Way (template)*.

The VAD activity level assumes a connection without significant background noise. If there is significant noise such as background music, an earthquake, or even background telephony noise, there may never be silence to suppress in a particular direction.

The trick is to provision VAD without introducing further blocking. Two criteria are important:

1. The maximum number of VAD calls needs to fit within the maximum allowable bandwidth to prevent call blocking under extreme circumstances, although data blocking may occur.
2. The nominal number of VAD calls needs to fit within the target voice region to statistically allow enough bandwidth for data services.

Criteria 1

With all VAD calls active, a basic amount of bandwidth should be reserved to allow signaling and maintenance. The DOCSIS maintenance slot that occurs for 2 ms every 2 seconds is a good baseline to choose. A further *VAD Region Allowance (template)* has been included in the formula as a tuning factor to increase or decrease the *VAD Region Max Target (formula 52)*. For a non-blocking system, the allowance should be 0 percent.

Formula (52) **_VAD Region Max Target = (100% - Maintenance Region) * (1 + VAD Region Allowance)_**

Result: VAD Region Max Target = (100% - 20%) * (1 + 0%) = 80%

The maximum number of VAD calls, *VAD Connections Max*, is equal to the offered load from the fiber node, which is *Connections per Fiber Node (formula 25)*.

Formula (53) **_VAD Connections Max = Connections per FN_**

Result: VAD Connections Max = 47

Just as was done without VAD, the maximum number of connections is compared to the connections available per upstream receiver modified by the target voice region for VAD.

Formula (54) **_VAD Receivers per FN = Ceiling (Connections per FN / (Connections per Upstream * VAD Region Max Target))_**

Result: VAD Receivers per FN = Ceiling(47 / (23 * 80%) = Ceiling (2.5) = 3

Criteria 2

Given the maximum number of VAD connections, the nominal number of VAD connections has to be calculated. As mentioned earlier, it is not valid to directly use the *VAD Activity Factor Two Way (template)* because it describes activity within a two-way conversation, and the upstream is a collection of one-way conversations. However, the collection of conversations is an excellent application for erlang statistics.

For example, if each VAD connection has a *VAD Activity Factor Two Way (template)* of 40 percent, then the amount of traffic it will have is 0.4 erlangs. If there were 100 VAD full-time conversations, then they would comprise 40 erlangs of traffic, an amount that would require 53 full-time connections, if we assume a 1-percent *VAD Grade of Service (template)*. In this case, the nominal number of connections is 53 percent of the maximum number of connections. This percentage will increase as the maximum number of VAD calls decreases, and will decrease as the number of VAD calls increases, so a simple halving of the VAD maximum conversations is not accurate.

At low numbers, the erlang model predicts a value higher than the actual number of VAD conversations, so a minimum limit is imposed in the formula. At high numbers, the erlang formula becomes too aggressive, so the 90-percent limit is observed as before.

- If *VAD Connections Max = Connections per FN (formula 25)*, then

Formula (55) **_VAD Connections Nominal = Min [VAD Connections Max, Erlang-BL [VAD GoS, VAD Connections Max * VAD Activity Level Two Way]]_**

Result: VAD Connections Nominal = Min [47, Erlang-BL [1%, 47 * 40%]] = 29

As a sanity check, this calculation results in a one-way activity factor of:

Formula (56) **_VAD Activity Factor One Way = VAD Connections Nominal / VAD Connections Max_**

Result: VAD Activity Factor One Way = 29 / 47 = 62%

Since the maximum number of VAD connections does not occupy the entire channel, the *VAD Activity Factor One Way (formula 56)* is combined with the *VAD Region Max Target (formula 52)* to get the *VAD Region Nominal Target (formula 57)*.

Formula (57) ***VAD Region Nominal Target = VAD Region Max Target * VAD Activity Factor One Way***

Result: VAD Region Nominal Target = 80% * 62% = 50%

This target goal represents full-time connections, and has to be less than the non-VAD voice region in order to statistically support the required data bandwidth.

Formula (58) ***VAD Region Nominal Target ≤ Voice Region Target***

Result: 50% ≤ 65%

If this last formula is false, then the *VAD Region Allowance (template)* can be decreased to bring it in line. Conversely, if it is felt that the maximum number of VAD calls occurs below some acceptable threshold, then the *VAD Voice Region Allowance (template)* can be increased while keeping this formula 58 true.

VAD will impact switching performance as calculated in the section "Voice" on page 45 because there are simply fewer packets to switch. The switching calculations should use the *VAD Connection Nominal (formula 55)* value for switching performance, and not the *VAD Connections Max (formula 53)*. No separate de-rating is necessary. The trunking gateway will have one port dedicated for each VAD connection, so a separate de-rating should be applied. Since the trunking gateway carries both directions of the call and the port density is relatively high, the *VAD Activity Factor Two-Way (template)* can be used directly.

4.3.3.2 Making an Optimized Decision

Upstreams are usually available in groups of two and equipment limitations usually require that fiber nodes be totally connected within a line card. An example of the optimized choices of upstreams given a particular line card size is given in Table 5.

Table 5 Optimized Upstream Groupings

Line Card, Downstream x Upstream	Optimum Upstream Groupings
1 x 8	1, 2, 4, 8
1 x 6	1, 2, 3, 6
1 x 4	1, 2, 4
1 x 2	1, 2
1 x 1	1

The more fiber nodes that can be connected per line card, the less CMTS equipment is required, potentially translating to lower costs and less rack space utilization.

The number of fiber nodes per line card, having accounted for the *Voice Region Allowance (template)*, is as follows:

Formula (59) ***FN per CMTS Domain = Floor (Receivers per CMTS Domain / Receivers per FN Min)***

Result: FN per LC = Floor(6 / 3) = 2

The number of fiber nodes supported per CMTS chassis follows:

Formula (60) ***FN per CMTS = FN per CMTS Domain * CMTS Domains per Line Card * Line Cards per CMTS***

Result: FN per CMTS = 2 * 1 * 4 = 8

A useful metric is the number of households passed per CMTS. Note that the HHP number is extremely dependent upon penetration rates, and could vary from 1000 to 50,000, depending upon the circumstances.

Formula (61) ***HHP per CMTS = FN per CMTS * HHP per FN***

Result: HHP per CMTS = 8 * 500 = 4000

Consider an example of an economic versus a technical trade-off where four upstreams are required per fiber node and the line card has six upstream receivers. The solutions follow:

- Use a 1 x 6 line card with one fiber node on four upstreams; the other two upstreams are spare. This solution is the most expensive solution; it has excess bandwidth, but it does have room for growth.
- Use a 1 x 4 line card with one fiber node on all upstreams. The line card cost is now optimized, but there are only four fiber nodes per chassis.
- Use a 1 x 6 line card and one fiber node on three upstreams, resulting in eight fiber nodes per chassis. This setup gives the least-expensive price as well as half the rack density at a slightly reduced performance level. That reduction is not a full 25-percent penalty, however, because of round-off error since the four was chosen by rounding up, and three was chosen by rounding down. This trade-off may be the best cost/space/performance trade-off.

Some cases of rounding may result in receivers not being used. For example, a 1 x 8 line card with three receivers required per fiber node would use six receivers to support a total of two fiber nodes, and would have two receivers unused. The mapping efficiency of a particular solution can be measured by the following formula:

Formula (62) ***Line Card Utilization = (Receivers per FN * FN per Domain / Receivers per Domain)***

Result: Line Card Utilization = (3 * 2 / 6) = 100%

These equations are assuming that upstreams and downstreams are on the same line card. If they are not, then a group of upstreams and downstreams across a group of line cards should be used in these formulas in place of the single line card variable.

Another consideration is if the equipment will allow sharing of fiber nodes across line cards, where each line card has its own downstream. For example, if three receivers were needed, and the line cards were arranged as 1 x 4, could a fiber node be connected across two line cards? This would require the CMTS to allow sharing of fiber nodes across line cards, and would likely require the fiber node to carry the downstream frequency from both line cards. This scenario is interesting only if there is enough traffic to actually occupy two downstream RF channels. Downstream channels represent revenue on a cable system, and under-utilized channels means lost revenue. Again, this scenario can be modeled by treating the group of line cards as one larger line card.

4.3.3.3 The Final Metrics

Now that the mapping has been done, the actual number of connections available is known:

Formula (63) ***Connections per FN Max = Connections per Upstream Max * Upstreams per FN***

Result: Connections per FN Max = 23 * 3 = 69

and the actual size of the voice region can be calculated.

Formula (64) ***Voice Region Actual = Connections per FN / Connections per FN Max***

Result: Voice Region Actual = 47 / 69 = 68%

If VAD is active, the metrics are similar but renamed. The formula for the actual maximum VAD connections is the same as the formula for the non-VAD voice region.

Formula (65) ***VAD Region Max Actual = Connections per FN / Connections per FN Max***

Result: VAD Region Max Actual = 47 / 69 = 68%

The nominal VAD voice region is calculated by accounting for the one-way activation factor.

Formula (66) **VAD Region Nominal Actual = VAD Region Max Actual * VAD Activity Factor One Way**

Result: VAD Region Nominal Actual = 68% * 62% = 42%

In this example, both the theoretical approach of the last section and the practical approach of this section concluded three fiber nodes per upstream. This result will not always occur. Note that the theoretical approach rounded up to three fiber nodes, while the practical approach rounded down to three fiber nodes.

Up to this point, all example calculations have been targeted at the ideal 500 HHP fiber node. In a real system, the fiber node sizes on an average 500 HHP system might actually vary from 20 to 2000. The market penetration rates will also vary across nodes, resulting in lots of variety. For a fiber node of 250 HHP, the results for a 500 HHP fiber node cannot be halved because the erlang calculation is not linear. All the formulas and judgment calls have to be redone for each node size.

4.4 Equipment Sizing Based upon Performance

Of the different performance metrics to be considered, this section focuses on the packet-per-second switching capacity of the CMTS. Packets per second is a fundamental performance metric for packet-based switching systems.

4.4.1 Data

Unless an explicit packet-per-second performance value is given, one will have to be derived by taking average bandwidth numbers and assuming an average certain packet size. Packet sizes tend to follow the Ethernet sizes, which are 64 to 1522 bytes (this value includes the Ethernet encapsulation).

For data applications, the dominant protocol is TCP/IP. TCP generates an ACK for every packet received. Most of today's implementations combine two ACKs together into one packet. ACK packets are 64 bytes in size and travel in the opposite direction of the original packet. The typical data traffic today in the downstream will be Web traffic, file transfer, or streaming media, meaning big packets in the downstream and small packets in the upstream.

The average downstream packet rate is determined by:

Formula (67) **PPS Data per Direction = Data Rate / 8 bits per byte / Packet Size**

Result: PPS Data per Downstream Average = 80,000 / 8 / 500 = 20 pps

Result: PPS Data per Upstream Average = 32,000 / 8 / 64 = 63 pps

In formula 67, there are several times more packets in the upstream than in the downstream. Yet, the traffic model is assuming that large packets dominate the downstream and the upstream is dominated by ACKs. With two ACKs per packet, a more realistic model would be to set the packet-per-second rate in the upstream to be half of that in the downstream.

Formula (68) **PPS Data per CM = 1.5 (Bandwidth per Subscriber / 8 bits per byte / Packet Size)**

Result: PPS Data per CM Average = 1.5 * 80,000 / 8 / 500 = 30 pps

Result: PPS Data per CM Peak = 1.5 * 400,000 / 8 / 500 = 150 pps

Formula 68 has some practical advantages. It recognizes that when one direction of traffic is saturated, the other direction will not be. This is because the asymmetry of the data transfer is different from the asymmetry of the media.

There is no significant signaling for data other than the signaling for DOCSIS that has been included in a previous budget.

The data traffic loading on the CMTS follows:

Formula (69) **PPS per CMTS = PPS per CM * Active Subscribers per FN * FN per CMTS**

Result: PPS per CMTS Average = 30 * 10 * 8 = 2,400 pps

Result: PPS per CMTS Peak = 150 * 2.5 * 8 = 3,000 pps

For voice and data systems with many upstreams per fiber node, the data traffic will appear low. For data-only systems with many fiber nodes per upstream, the data traffic will appear higher. For example, a voice and data system may have three receivers per fiber node, whereas a data-only system may have four fiber nodes per receiver. That number is 3 * 4 = 12 times the density of fiber nodes and hence twelve times as many data customers per CMTS.

4.4.2 Voice

The packets per second of traffic due to the voice bearer path are related to the VoIP frame interval, the amount of time that it takes to record a voice packet. Typical times are 10 ms and 20 ms. The packet rate in each of the two directions are summed together. The traffic per voice bearer path follows:

Formula (70) ***PPS Voice per Connection = 2 Directions * 1 / VoIP Frame Interval***

Result: PPS Voice per Connection = 2 * 1 / 0.01 sec = 200 pps for 10-ms frames

Result: PPS Voice per Connection = 2 * 1 / 0.02 sec = 100 pps for 20-ms frames

Using the connection results from *Connections per FN (formula 25)*, the voice traffic per fiber node follows:

Formula (71) ***PPS Voice per FN = PPS Voice per Connection * Connections per FN * VAD Activity Level Two Way***

Result: PPS Voice per FN = 200 pps * 47 * 100% = 9.4 kpps

Next is the voice bearer traffic per CMTS.

Formula (72) ***PPS Voice per CMTS = Voice PPS per FN * FN per CMTS***

Result: PPS Voice per CMTS = 9.4 * 8 = 75.2 kpps

4.4.3 Signaling

The voice signaling traffic is related to the call attempts per line per hour. Note that call attempts are per line and not per connection.

Formula (73) ***CAT per CMTS = CAT per Line * Lines per FN * FN per CMTS***

Result: CAT per CMTS = 2.78 * 250 * 8 = 5,560 CAT per hour

Each call attempt has associated signaling packets. The signaling packets such as H.323, Media Gateway Control Protocol (MGCP), and Session Invitation Protocol (SIP) that transit the CMTS will flow in the data path and do not need to be separately accounted for. Protocols such as DOCSIS 1.1, Resource Reservation Protocol (RSVP), Common Open Policy Service (COPS), and billing which terminate or originate on the CMTS, do need to be accounted for by assuming an average number of signaling packets per call attempt. Therefore, the voice signaling packet rate follows:

Formula (74) ***PPS Signaling per CMTS = Call Attempts per CMTS * CMTS Packets per CAT / 3600 seconds per hour***

Result: PPS Signaling per CMTS = 5560 * 20 / 3600 = 31 pps

Since the typical result is several orders of magnitude lower than the data rate, it can be ignored unless the signaling packets are handled in a separate signaling path from the bearer packet switching. This assumption can be further verified by calculating the ratio between the number of packets generated by a single bearer channel in a call versus the number of packets generated by the signaling for the same call.

Formula (75) ***Ratio Bearer:Signaling = (Call Holding Time * Voice Traffic per Connection) / CMTS Signaling Packets per Call***

Result: Ratio Bearer:Signaling = 180 seconds * 200 pps / 20 packets = 1800

This result validates the assumption that the transmission switching path can ignore the impact of signaling packets.

4.4.4 Total CMTS Traffic Loading

Adding the packet loading from the various sources provides the total loading:

Formula (76) ***PPS CMTS = PPS Data per CMTS + PPS Voice per CMTS + PPS Signaling per CMTS***

Result: PPS CMTS Avg = 2400 + 75,200 + 31 = 77,631

Result: PPS CMTS Peak = 3000 + 75,200 + 31 = 78,231

If the calculated performance requirements for the CMTS exceed its capabilities, than the voice connection density should be reduced until the performance levels can be guaranteed.

4.4.5 The Brute Force Approach to Performance

All this theory and modeling is nice, but sometimes there is a simple elegance to a brute force approach.

Without too much trouble, a CMTS can easily dump out the following statistics:

- The number of cable modems attached to an upstream receiver or downstream transmitter
- The number of packets per second on that upstream or downstream
- The number of bytes per second on that upstream or downstream

As it turns out, these are the only statistics that are really needed. The number of attached cable modems represents that total combined market penetration. The number of packets per second represents the total combined traffic performance requirements. The number of bytes per second represents the total combined bandwidth requirements.

If these numbers are further broken down into traffic types, active versus nonactive, peak versus average, and so on, a greater understanding of the network will be achieved. However, it may be at the expense of less accuracy due to the numerous assumptions and additional variables that get added as part of the process.

For example, in data-only DOCSIS environments, it currently has been observed that data traffic tends to be 3 to 5 packets per second per cable modem. (In older dialup modem banks, the number tended to be one packet per second per analog modem)[1]. Of course, not all cable modems are logged on, so the ones that are active are using more than that. But that doesn't matter so much. What matters is if there are 1000 cable modems on the upstream, then the upstream receiver has to handle 3000 to 5000 packets per second of performance.

So why bother with the details if the end results are all that matters? The answer lies in another question. Why is the performance level 5 packets per second per cable modem, and what is it going to be a year from now? The only way to answer that question is to understand the various sources of traffic, what their individual market penetrations are, and what their individual growth factors are.

Once again, theory and practice work together. The brute force approach should be used to check that the models are accurate. The models should be used to predict the brute force results and to predict what it will become as time goes on.

4.5 Cable Modem Termination System Sizing

Up to this point, redundancy has not been accounted for. This section explains and accounts for the impact of adding redundant CMTSs. The Working CMTS is the CMTS that normally carries traffic. The Protect CMTS is the redundant CMTS that picks up the traffic load when the working CMTS fails.

4.5.1 Without Redundancy

Since line cards are usually priced separately from CMTSs and not all CMTSs may be fully populated, the quantity of both CMTS chassis and line cards are worked out separately. The number of *Working Line Cards per Hub (formula 77)* required follows:

Formula (77) ***Working Line Cards per Hub = Ceiling(FN per Hub / (FN per CMTS Domain * CMTS Domains per Line Card)***

Result: Working Line Cards per Hub = Ceiling(40 / (2 * 1) = 20

1. These numbers have been derived by observing "show interface" statistics from Cisco Systems products at various customer sites.

The number of *Working CMTSs per Hub (formula 78)* follows:

Formula (78) **Working CMTS per Hub = Ceiling(Working Line Cards per Hub / Line Cards per CMTS)**

Result: Working CMTS per Hub = Ceiling(20 / 4) = 5

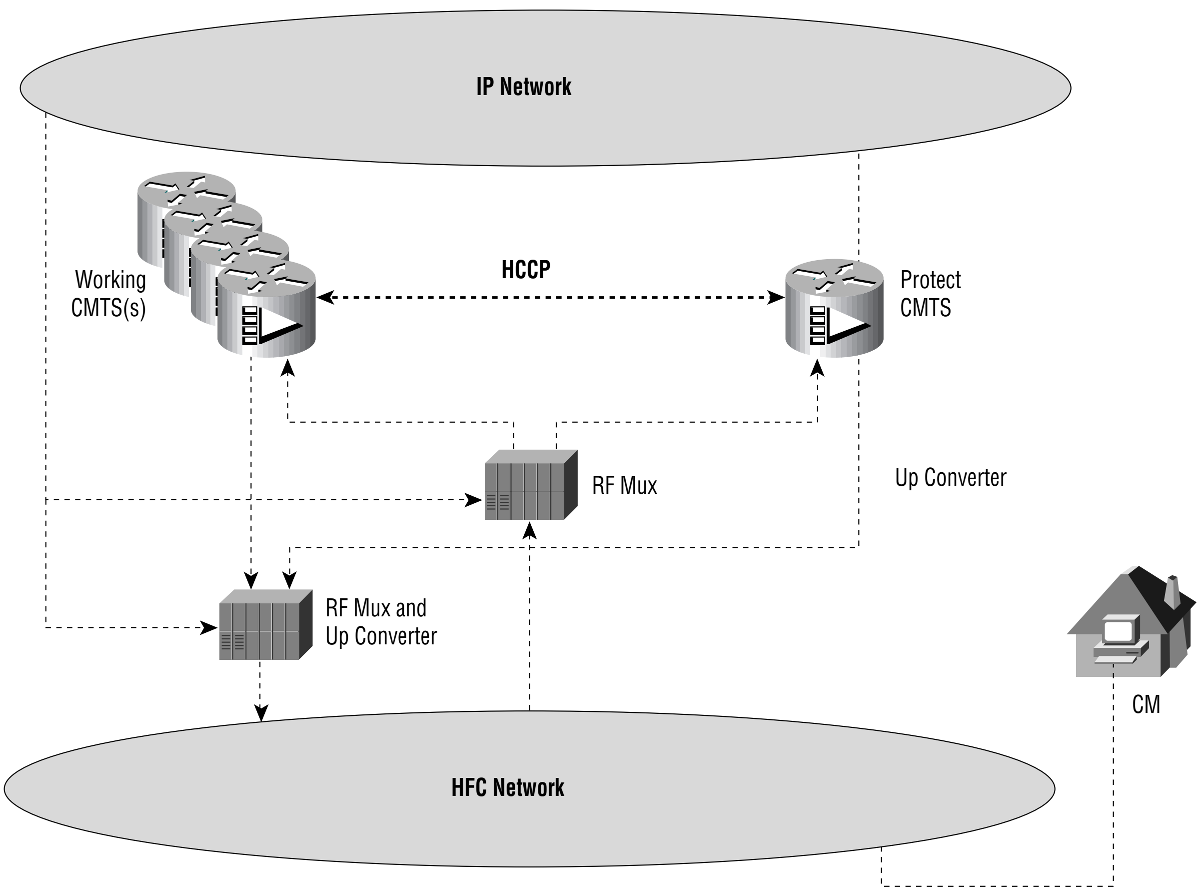
4.5.2 Adjustment for Redundancy

To ensure a highly available system, a certain amount of redundancy must be added. For large system components, such as backbone routers, redundancy is often done within the component, referred to as redundancy at the FRU (field replaceable unit) level. For smaller system components such as access routers, redundancy is achieved by having more than one router, referred to as redundancy at the chassis level A CMTS may fit either one of these profiles.

For both profiles, redundancy can be implemented as 1+1 or N+1. In 1+1 redundancy, for every system element, such as a line card, there is a backup element. In N+1 redundancy, for every N number of system elements, there is one backup element. 1+1 redundancy is a subset of N+1 redundancy.

Figure 7 shows a N+1 CMTS configuration with redundancy at the chassis level. The redundant CMTS is referred to as the Protect CMTS. The main CMTSs are referred to as Working CMTSs.

Figure 7 Redundant CMTS System



The value of N is chosen by separate availability metrics.

The number of Protect CMTSs required follows:

Formula (79) ***Protect CMTS per Hub = Ceiling(Working CMTS per Hub / N Max)***

Result: Protect CMTS per Hub = Ceiling(5 / 8) = 1

The number of additional line cards required follows:

Formula (80) ***Protect Line Cards per Hub = Protect CMTS per Hub * Line Cards per CMTS***

Result: Protect Line Cards per Hub = 1 * 4 = 4

Since each hub location that has redundancy must have at least one redundant component, the *N Max Redundancy Ratio (template)* may not be reached. The formula can be reversed to check the actual value of N.

Formula (81) ***N Actual = Working CMTS per Hub / Protect CMTS per Hub***

Result: N Actual = 5 / 1 = 5

4.5.3 The Result

By adding the working and protect system components together, the CMTS sizing can be completed.

The number of total line cards per hub follows:

Formula (82) ***Line Cards per Hub = Working Line Cards per Hub + Protect Line Cards per Hub***

Result: Line Cards per Hub = 20 + 4 = 24

The number of total line cards per headend system follows:

Formula (83) ***Line Cards per HE = Line Cards per Hub * Hubs per HE***

Result: Line Cards per HE = 24 * 5 = 120

The total number of CMTSs per hub follows:

Formula (84) ***CMTS per Hub = Working CMTS per Hub + Protect CMTS per Hub***

Result: CMTSs per Hub = 5 + 1 = 6

The total number of CMTSs per headend system follows:

Formula (85) ***CMTS per HE = CMTS per Hub * Hubs per HE***

Result: CMTS per HE = 6 * 5 = 30

4.6 Cable Modem Sizing

There are three general configurations of cable modems: voice only, data only, and voice and data. Then typically there may be two or four lines of voice. An aggressive voice market plan that allows for later service expansions may always assume a voice CM with data, regardless of the customer. The total number of cable modems in a network depends upon the overall market penetration.

Formula (86) ***CM per HE = HHP per HE * Market Penetration Total***

Result: CM per HE = 100,000 * 25% = 25,000

4.7 Call Management Server Sizing

The centralized call management server (CMS), be it a MGCP) call agent or a SIP proxy, manages signaling traffic and does not see bearer traffic. Thus a CMS is sized by its ability to handle signaling traffic. That signaling traffic is related to the number of calls generated within the serving area of the call management server.

The number of call attempts in a serving follows:

Formula (87) ***Call Attempts per HE = Call Attempts per Line * Lines per HE***

Result: Call Attempts per HE = 2.78 * 50,000 = 139,000

The number of call management servers based upon this metric follows:

Formula (88) ***CMS per HE = Call Attempts per HE / Call Attempts per CMS Max***

Result: CMS per HE = 139,000 / 200,000 = 69.5%

The CMS value is not rounded up since CMS servers are geographically independent and may be shared among headend systems. Call attempts are usually measured per hour, although other units of time may be used if consistency is maintained.

Other CMS sizing factors may also include provisioning and Operations Support System (OSS) concerns, but they are not addressed here.

4.8 Trunk Gateway Sizing

4.8.1 Bandwidth

Voice traffic that remains on the IP network for an end-to-end call is referred to as on-net traffic. Voice traffic that must go off the IP network and onto the PSTN is known as off-net traffic. Off-net traffic will go through a trunk gateway.

Some assumptions must be made about the percentage of calls that will remain on net versus off net. Note that even with high market penetration, if the market penetration is strictly residential, then all calls to businesses will go off net. Therefore, if 50 percent of the calls to and from a residence were with businesses, than even 100-percent penetration of the residential market would still result in 50 percent off-net calls. Further, if market penetration was 25 percent and 25 percent of residential calls were on net, then 75 percent of the remaining 50 percent of the calls will also be off net. The result is 85- to 90-percent off-net calls, even with decent market penetration of residential voice.

Since the VoIP traffic concentration already occurs on the IP network, the trunking gateway does not provide concentration. That means that one voice connection on the IP network is mapped to one DS0 time slot on the PSTN network.

The trunk gateway receives an aggregation of traffic from the headend system. To size a trunk gateway, the calculation is done by looking at the bulk traffic on the network, multiplying by the percentage of off-net traffic, and then doing an erlang calculation to see how many trunks or ports are required to support it. Since the number of traffic erlangs is large, the number of trunking gateway ports is usually determined by the 90 percent that this white paper has imposed on the erlang B formula.

Formula (89) ***TGW Ports per HE = Ceiling(Erlang-BL(Erlangs per HE * off net, Grade of Service))***

Result: TGW Ports = Erlang-BL(6944 * 85%,1%) = 6559

The number of physical boxes involved relies on the number of ports per box. If the boxes are located at different locations, then there will be a round-off error not included here. The number of ports is usually rounded off to the nearest T1 or E1.

Trunking gateways usually use DSP farms to perform compression. Thus the number of ports per trunking gateway is usually directly related to the compression algorithm being run. The number of ports supported will typically be higher for G.711 and one-half to one-quarter for G.729 or G.728.

Formula (90) ***TGW Line Cards per HE = Ceiling (TGW Ports per HE / TGW Ports per Line Card)***

Result: TGW Line Cards per HE = Ceiling (6559 / 192) = 35

The number of chassis required to contain these line cards follows:

Formula (91) ***TGW Chassis per HE = Ceiling (TGW Line Cards per HE / TGW Line Cards per Chassis Max)***

Result: TGW per HE = 6559 / 4032 = Ceiling (35/20) = 2

The number of line cards in each chassis follows:

Formula (92) ***TGW Line Cards per Chassis = Ceiling (TGW Line Cards per HE / TGW Chassis per HE)***

Result: TGW Line Cards per Chassis = ceiling (35 / 2) = 18

4.8.2 Performance

Of the different performance metrics to be considered, this section focuses on the packet-per-second switching capacity of the trunk gateway.

4.8.2.1 Voice

The bearer path switching performance follows:

Formula (93) ***TGW Bearer PPS = PPS per Connection * TGW Ports per Line Card * TGW Line Cards per Chassis * VAD Activity Level Two Way***

Result: TGW Bearer PPS = 200 * 192 * 18 * 100% = 691 kpps

4.8.2.2 Signaling

The number of call attempts per hour follows:

Formula (94) ***TGW Call Attempts = Call Attempts per Line * Line-to-Trunk Ratio * TGW Ports per Line Card * TGW Line Cards per Chassis***

Result: TGW Call Attempts = 2.78 * 5.3 * 192 * 18 = 51,000

The TGW will primarily originate and terminate H.323 or MGCP messages, but may also use RSVP or other Layer 3 reservation events.

Formula (95) ***TGW Signaling PPS = TGW Call Attempts / 3600 seconds per hour * Packets per CAT***

Result: TGW Signaling PPS = 51,000 / 3,600 * 10 = 142 pps

Again we see that the required bearer path switching performance completely overshadows the signaling performance requirements. However, a different processor than the one handling the bearer packets may handle the signaling packets. For example, the central CPU might handle all signaling packets, while a dedicated hardware switching matrix might handle the constant bearer packets. Thus, both performance limits must be met.

If the calculated performance requirements for the trunk gateway exceed its capabilities, than the voice connection density should be reduced until the performance levels can be guaranteed.

5 IP Network Sizing

5.1 CMTS Backhaul Bandwidth

If most of the traffic that a CMTS handles is data, the backhaul data bandwidth will be the sum of the data from the fiber node, if we assume that there is very little local traffic. If most of the traffic in a CMTS is voice, then the CMTS is acting as a concentration point, and the amount of backhaul bandwidth required becomes a traffic calculation. One modeling approach treats the CMTS as a group of fiber nodes, and calculates the voice and data loading from those fiber nodes.

The data loading per CMTS is the sum of the data loading from each fiber node. Upstream and downstream are done separately, because the higher of the two numbers will dictate the bandwidth of any duplex media used for backhaul. The bandwidth on the CMTS backhaul in the downstream direction follows:

Formula (96) ***BW Data per CMTS in the Downstream = BW Data Downstream * FN per CMTS***

Result: BW Data per CMTS in the Downstream = 1100 * 8 = 8800 kbps

And in the upstream direction:

Formula (97) ***BW Data per CMTS in the Upstream = BW Data Upstream * FN per CMTS***

Result: BW Data per CMTS in the Upstream = 368 * 8 = 2950 kbps

If the CMTS is to be a nonblocking switch, then it needs to support the sum of all the connections brought to it from the fiber nodes. The number of backhaul connections for voice traffic per CMTS follows:

Formula (98) ***Connections per CMTS = Connections per FN * FN per CMTS***

Result: Connections per CMTS = 47 * 8 = 376

The size of the VoIP packets is different on the IP backhaul since the protocol header is different from DOCSIS and there is no payload header suppression. The best approach is to convert the CMTS connection number into a bandwidth number. For that, a conversion factor is required that depends upon voice packet length, compression, and frame format. In the data world, an Ethernet frame is a common frame format that is commonly used for comparisons of bandwidth. The amount of voice bytes in that frame depends upon the frame length and the encoding rate. For G.711, 10 ms, the number of voice bytes in a packet follows:

Formula (99) ***Voice Bytes per Packet = Codec Bite Rate / 8 * VoIP Frame Time***

Result: Voice Bytes per Packet = 64,000 / 8 * 0.01 = 80

In the IP data world, the term packet refers to a Layer 3 protocol entity, while a frame refers to a layer protocol entity. Now that we know the size of the packet payload, we can map that into a Layer 2 Ethernet frame and calculate a bit rate. IP adds a 12-byte Real-Time Transport Protocol (RTP) header, an 8 byte User Datagram Protocol (UDP) header, and a 20-byte IP header. Ethernet adds a 14-byte header and a 4-byte cyclic redundancy check (CRC).

Formula (100) ***BW per Connection = (Voice Bytes per Packet + Packet Overhead + Frame Overhead) * 8 bits per byte / VoIP Frame Time***

Result: BW per Connection = (80 + 40 + 18) * 8 / 10 = 110 kbps

Note that the preamble for Ethernet is not included in these types of calculations. Instead, the preamble and issues due to collisions are included in how efficient a particular link is. For example, a 100-Mbps Ethernet connection might be considered only as 70 percent, meaning a maximum usage of 70 Mbps, because of preamble or other considerations.

Thus the bandwidth loading due to voice per CMTS follows:

Formula (101) ***BW Voice per CMTS = Connections per CMTS * BW per Connection * 2 directions***

Result: CMTS Total Voice BW = 376 * 110 = 83 Mbps in each direction

The CMTS will have some signaling going on with the network for setting up calls, making policy decisions, and just receiving routing updates. Let's assume a 2-percent overhead. Note that the previous maintenance and signaling budget was specific to DOCSIS only.

Formula (102) ***BW Total CMTS = (BW Data CMTS + BW Voice CMTS) * (1 + IP signaling Overhead)***

Result: BW Total CMTS in the Downstream = (8.8 + 83) * (1 + 0.02) = 94 Mbps

Result: BW Total CMTS in the Upstream = (2.95 + 83) * (1 + 0.02) = 88 Mbps

Assuming a symmetrical media for backhaul, the larger of these two bandwidth numbers is chosen.

Formula (103) ***BW CMTS Backhaul = Max (BW Total CMTS in the Downstream, BW Total CMTS in the Upstream)***

Result: BW CMTS Backhaul = Max (94, 88) = 94 Mbps

If the required backhaul bandwidth exceeds the usable bandwidth of the backhaul media, then the traffic loading of the CMTS needs to be reduced until the bandwidths match.

5.2 IP Backhaul Strategies

The IP backhaul consists of any equipment that connects one or more CMTSs at a hub together, onto the fiber link, and then into a router complex at the headend. Mapping this bandwidth to a backhaul media depends upon the media raw bandwidth and its efficiency. Several scenarios are possible.

5.2.1 Ethernet

Ethernet, usually 100 Mbps or 1 Gbps, tends to be the least expensive backhaul mechanism. For colocated CMTSs at a headend, many CMTSs may be aggregated together with Ethernets and Ethernet switches. For remotely located single-CMTS-per-hub installations, a common cost-effective approach is to take the data out the 100BaseT Ethernet port, convert it to optical, and run it over dark fiber to the headend.

The efficiency of Ethernet depends on how it is used. For example, if multiple CMTSs are connected together by 100-Mbps Ethernet into a hub before being passed to an edge router, there would be collisions and the efficiency of the Ethernet would be lower. If the edge router directly terminated each of the Ethernets, then there would be no collisions and the Ethernet efficiency would be much higher. If the Ethernet is not full duplex, then again there will be collisions, so its duplex bandwidth will be less than half of 100 Mbps.

5.2.2 SONET

SONET OC-3 has a base bit rate of 155 Mbps. It does not have collisions like Ethernet may have, but it does have different framing protocols. IP with SONET framing is good for about 134 Mbps, IP over ATM over SONET framing will be less because of the additional ATM overhead, while Packet over SONET (POS), which maps packets directly to unframed SONET, is good up to about 148 Mbps.

5.2.3 Dynamic Packet Transport

A more futuristic approach is to adopt the new DPT (Dynamic Packet Transport) OC-12 ring and embed it directly into the CMTS. DPT will optically bypass a CMTS that has functionally failed, or will close a ring if a CMTS has power failure. The equipment in each hub is the DPT OC-12 interface card in each CMTS. The DPT would terminate on one or two routers at the headend, which would then interface into one or more OC-12, OC-48, or OC-192 rings in the backbone. The second router at the headend would be for redundancy.

The OC-12 DPT consists of two counter rotating rings, referred to as the inside and outside rings that connect a series of nodes together. Let's assume that the headend is one node, and the hubs are the other nodes, and that all hub traffic comes to and from the headend. The downstream traffic will come on one of the rings from the headend node to a hub node, while the upstream traffic will come back on the other ring, following the same path as the downstream traffic. The effect of using two rings doubles the rated OC-12 bandwidth. Since the physical paths are symmetrical, the best approach for network design is to take the greater of the upstream or downstream bandwidth, and compare it to the OC-12 rated bandwidth.

DPT can send downstream data in one of two directions from the headend, and it allows traffic to exist simultaneously on one side of the ring or the other. This setup again doubles the bandwidth. However, if the design is intended to cover ring failure, worst-case ring failure will not allow that final doubling.

The bandwidth of the network connecting the hubs to a headend would be as follows: Note that the redundant CMTSs are not included in this calculation because they are not generating traffic.

Formula (104) ***BW HE = Hubs per HE * Working CMTS per Hub * BW CMTS Backhaul***

Result: BW HE = 5 * 5 * 94 = 2.4 Gbps

The number of rings required to support this number follows:

Formula (105) ***DPT Rings per System = Ceiling (BW HE / BW DPT Ring)***

Result: DPT Rings per HE = 2.4 Gbps / 2.4 Gbps = 1 (OC-48)

The number of DPT line cards required is equal to the number of the CMTSs, regardless of how many rings actually exist, since there is one backhaul connection per CMTS.

Formula (106) ***DPT Line Cards per HE System = CMTS per Hub * Hubs per HE***

Result: DPT Line Cards per HE System = 6 * 5 = 30

In addition, there will be a DPT card in each of the headend routers.

6 Related Topics

6.1 Support for Legacy Analog Modem and Fax

6.1.1 Analog Modem Support

With the introduction of DOCSIS and high-speed broadband services, many users who formally used an analog modem will now directly connect their computer to the data port of the cable modem and enjoy connectivity 1000 times faster at 27 Mbps instead of 28 kbps. Even if the downstream path is rate limited to 2 Mbps, the peak data rate will still be 100 times faster.

Still, there may be a minority of subscribers who still need to connect to their data provider with an analog modem. One application is for connecting into corporate intranets. Corporate firewalls usually block Internet access, so the only way to authenticate and allow in a remote user is often through a dialup connection. Eventually, virtual private network (VPN) technology will duplicate that same secure connectivity by generating an encrypted tunnel between the remote user, over whatever IP network that may exist, and the corporate intranet.

Another scenario is the data subscriber who wants to bypass the $40 per month high-speed data service offered by the local MSO. The subscriber does this by connection via an analog modem over the voice port to another lower priced ISP. The additional voice port might cost only $5 per month, with the budget ISP being $10 per month. Not only does this scenario represent lost revenue, but it might also represent lost bandwidth.

One of the things that every telephony provider will attest to is that modem traffic does not conform to classic voice models in traffic engineering and ties up valuable switch resources. Instead of 0.139 erlang of traffic, the modem call could be up for hours at a time. If it is a local call, that one erlang of traffic is not generating additional revenue.

This problem can be avoided in the VoIP environment through clever use of TCP/IP protocols. Consider two implementation approaches. In both scenarios, an application on the user's computer is generating and receiving IP packets that are converted to and from modulation tones by a V.28 analog modem.

In the first implementation, the analog modem tones are received by a media terminal adapter (MTA) in the cable modem and placed into a G.711, 64-kbps, bidirectional VoIP bearer path. This process is referred to as analog modem pass-through.

In the second implementation, the DSP processor in the MTA is used to receive and decode the modem tones, just like an answering modem would do. After decoding the modem tones, the bytes are extracted and packaged into new TCP/IP packets. These packets are sent to a centralized modem pool that takes the packets off the IP network, remodulates them back into analog modem tones, and delivers them to the PSTN. This technique is referred to as analog modem relay.

The analog modem pass-through technique is constantly sending packets across the IP network at a rate of about 110 kbps. The analog modem relay is sending data at the same rate that the originating computer is sending it, which is approximately 28.8 kbps. The ratio of these rates follows:

Formula (107) ***Analog Modem Bandwidth Ratio = BW Pass-Through / BW Modem Relay***

Result: Analog Modem Bandwidth Ratio = 110 / 28.8 ~ **4 X**

The analog modem pass-through technique is active 100 percent of the time. The modem relay technique sends packets across the IP network only when it receives bytes from the computer. On average, let's assume that a user is actually sending or receiving data only 20 percent of the time.

Formula (108) ***Analog Modem Usage Ratio = Usage of Pass-Through / Usage of Modem Relay***

Result: Analog Modem Usage Ratio = 100% / 20% = **5 X**

Modem traffic at 28.8 kbps will typically support low-bandwidth data applications such as Web traffic or e-mail. The bulk of the data transfer is downstream, with the upstream being TCP ACKs or key stokes. CableLabs tests have found the ratio between upstream and downstream traffic to be 19:1. To be conservative, half that would give a ratio of:

Formula (109) *Analog Modem Symmetry Ratio = 10 X*

Multiply these three ratios to give the ratio of bandwidth used on the IP network for analog modem pass-through versus analog modem relay.

Formula (110) *Analog Modem Relay Efficiency Ratio = Bandwidth Ratio * Usage Ratio * Symmetry Ratio*

Result: Analog Modem Relay Efficiency Ratio = 4 * 5 * 10 = **200 X**

This ratio is conservative. If the actual 20 X ratio was used for symmetry and a more realistic usage ratio of 10 percent usage was used, this ratio could be as high as:

Result: Analog Modem Relay Efficiency Ratio = 4 * 10 * 20 = **800 X**

With the conservative approach, it can be concluded that the same amount of upstream bandwidth can support 200 times more users with analog modem relay compared with analog modem pass-through.

Since the call holding times for voice traffic and analog modem pass-through traffic are so radically different, the two sources of traffic should be kept separate in traffic engineering calculations. An average call holding time or an average erlang value that includes both traffic types should be avoided because that average will depend upon the market penetration of the two applications. Numbers for market penetrations change often, whereas numbers for call holding times tend to be rules-of-thumb that do not change often. Thus averaging the two values introduces yet another assumption with an error factor that will increase over time.

The real problem with analog modem traffic is the connect time. A typical voice call is considered to be between 3 minutes for business and 8 minutes for residential. A short analog modem session could last 30 minutes. More typically, it may last a few hours. For users who are not being charged a connect time, and have e-mail programs that download new e-mail every 10 minutes, the session might last for days.

In reality, most users will connect directly to a cable modem and enjoy high-speed access at 10 times the amount of bandwidth they used to enjoy. Even then, direct IP connectivity is allowing 20 times as many users, each getting 10 times the level of service.

6.1.2 Fax Modem Support

The same options exist for transporting fax connections. The standard approach would be to send fax modem tones through a G.711 connection. The other approach would be to implement a fax relay protocol. In fax relay, the fax tones generated by the fax terminal are received by a DSP in the CM/MTA and converted back into bytes. The original data stream was on to an IP packet, so the bytes are packetized using the ITU-T T.38 protocol, and sent out over the IP network to a far-end T.38 endpoint. The transport can be TCP or UDP.

Fax calls start off as pass-through calls since it is unknown at call setup time that a fax terminal is attached. The call setup will have the same efficiency in its use of bandwidth, but for this analysis, it is assumed that the transfer time of the call dominates. Since fax is 9600 or 14,400 baud, the bandwidth ratio between the two scenarios is even higher than the analog modem relay scenario.

Formula (111) *Fax Modem Bandwidth Ratio = BW Pass-Through / BW Fax Relay*

Result: Fax Modem Bandwidth Ratio = 110 / 9.6 ~ **10 X**

Once the fax call is set up, it is constantly sending data, so both techniques will use 100 percent of their bandwidth in one direction.

Formula (112) *Fax Modem Usage Ratio = Usage of Pass-Through / Usage of Fax Relay*

Result: Fax Modem Usage Ratio = 100% / 100% = **1 X**

Faxes are definitely one way. The frame size is 64 bytes or 256 bytes. Multiple frames can be put into one packet. Assuming 256-byte frames and two ACKs per return packet with large size (if TCP is used as the transport, UDP would not have ACKs):

Formula (113) ***Fax Modem Symmetry Ratio = 8 X***

Putting these three ratios together yields

Formula (114) ***Fax Modem Relay Efficiency Ratio = Bandwidth Ratio * Usage Ratio * Symmetry Ratio***

Result: Modem Fax Efficiency Ratio = 10 * 1 * 8 = ***80 X***

This is conservative. If UDP is used as a transport, the symmetry ratio would more than double, bringing the combined ratio up to ***160 X*** or ***200 X***.

Fax, unlike analog modem traffic, tends to have a finite call holding time. Many fax calls may even be shorter than voice calls, meaning that pass-through fax traffic is more easily tolerated than pass-through analog modem traffic. However, the majority of third-line voice deployments may be targeted at home fax applications. So with fax relay, the same amount of bandwidth and the same amount of equipment will simply be able to support more users.

6.1.3 The Economic Trade-Off

So if analog modem and fax relay technology is so compelling from a technology viewpoint, why not always use it? There are some practical and business trade-offs. The first is that when protocols like fax or analog modems are interpreted, does that interface have to support every possible device that was ever made? If yes, that could be a lot of testing and potentially a lot of field problems if the technology is not designed correctly. This decision should be considered carefully. However, it is not uncommon for new technologies to face these problems as they are deployed into an environment where backward compatibility is a requirement. With good planning, this problem is solvable.

Another issue is licensing fees. Pass-through approaches do not have an incremental cost above the existing hardware and software platforms. Analog modem relay and fax relay software is usually purchased software that comes with a licensing fee. So the traffic engineer must do a cost analysis to decide which is more economical. Of course, if the analysis shows that it is more economical to increase the network equipment costs instead of paying licensing fees, some purchasing agent will be happy to use that analysis to get the fees down to where they are economically viable.

Perhaps the final issue is reliability. Analog and fax relay would have to interoperate with all or almost all analog or fax modems out in the world today in order to be successful. Is that possible? The design of the portion of the analog relay DSP code that terminates and generates modem tones is the same as the code that is in the current dialup servers. So the difficult problem has already been solved in a different part of the network. It is a matter of moving that code to a new operating point in the network and adding packetization. When it comes to fax relay, there are fewer fax standards and variants of fax technology in the market place than analog modem technology, so fax relay is also a bounded and solvable problem.

Analog modem and fax modem pass-through, although simple in concept, can be complicated in implementation. The pass-through mode, which is really an approach taken from circuit-switched technology, will work only if the network is virtually loss-less, as is the case with circuit-switching. In theory, a loss-less IP packet network can be built with multiple tiers of QoS and strict backbone reservation mechanisms, the ultimate goal of IP backbone networks. But it is a feature of IP networks, and all the protocols that run over IP networks, that packet loss can be tolerated. That feature is one of the important aspects of the IP network that gives it resilience and scalability.

By introducing one protocol that cannot tolerate packet loss, the entire network design is impacted. Existing, stable, IP networks may not be reusable as well as the networking technology used to build them. That scenario points to higher costs of deployment with longer time frames, higher costs, and a greater reliance on newer protocols, a scenario that can lead to less initial stability and reliability.

The irony of such a decision would be that after taking longer to deploy, spending more money to do it, and losing bandwidth efficiencies along the way, the real benefactors of analog modem pass-through may not be the service provider who is building the network, but the very companies who are trying to bypass the service provider's network in the first place.

Ultimately, the solution must match the technology. And if the technology is an IP-based network, then the best solution is an IP-based transport relay protocol for analog modem and fax modem.

6.2 CMTS Admission Control

6.2.1 Goals

This white paper has proposed traffic engineering methodologies for dealing with a variety of data, voice, and video sources. The goal has been to specify and build a multimedia environment that has its transport layer managed by the CMTS. The picture would not be complete without some comments on CMTS admission control.

CMTS admission control refers to the rules that the CMTS will follow when allocating bandwidth to different resources. These rules apply not only to the different types of traffic, but also to the parameters associated with the traffic. They are also affected by the link-layer capability of DOCSIS. This section specifically focuses on admission control for voice in the presence of data. In particular, the follow traffic is considered:

- Voice with or without compression
- Voice with multiple flow specs
- Voice with an emergency services call such as 911
- Voice with VAD
- Voice with Payload Header Suppression (PHS)
- Data

The assumption will be made that voice has higher priority than data. The focus is on determining different admission policies for various voice traffic scenarios. There are further policies for differentiating data flows that are not discussed.

Earlier the concept of a maximum threshold for voice calls that was less than the bandwidth of the link was introduced. That concept will be expanded upon in this section to deal with the different types of voice traffic.

The protocols mentioned such as Network Control Signaling (NCS), Distributed Control Signaling (DCS), and Distributed Quality of Service (DQoS) are PacketCable protocols[1]. RSVP is an Internet Engineering Task Force (IETF) protocol that has been adopted within the PacketCable DQoS architecture.

6.2.2 CMTS Call Model

The CMTS owns admission control. In a DCS environment, the CMTS will receive reservations from RSVP messages that originate at the CM or MTA. RSVP-based reservations that are DCS/DQoS related will be preauthorized at the CMTS from the DCS call server by means of the Common Open Police Service (COPS) protocol.

In a NCS environment, the CMTS will rely initially on native DOCSIS reservations, and eventually the same RSVP/DQoS reservation.

In both the NCS and DCS environments, the CMTS will always accept native DOCSIS reservations. The CMTS is the only entity that has observability on DOCSIS bandwidth requests. This is why the CMTS becomes the platform that has to aggregate all bandwidth requests for the HFC upstream.

6.2.3 Reservation Technique

Normally, RSVP would specify the actual bandwidth to be used. In cases like VAD, there is no definitive answer. Instead, the answer lies in a realm of statistics and policy decisions. For that reason, RSVP shall deliver to the CMTS the raw bit rate of the VoIP stream along with any call attributes such as VAD or header suppression or call priority. The CMTS admission control, based upon some algorithm or statistic, shall be responsible for calculating the bandwidth savings from these attributes.

For multiple flow specs, RSVP will report both flow specs to the CMTS. The CMTS will decide how to deal with the problem.

For emergency services such as 911, RSVP will include in the flow spec an admission object that the CMTS will know how to interpret.

1. More information on PacketCable can be obtained from its Web site at *http://www.packetcable.com*

For VAD, RSVP will include the raw bit rate of the voice and an attribute indicating that VAD is available on this connection. The CMTS will decide how to map VAD traffic onto DOCSIS.

For header suppression, RSVP will deliver the raw voice bit rate along with an object indicating which fields in the packet can be suppressed. The CMTS will decide whether to actually perform the suppression, and what bandwidth savings will be realized as a result.

6.2.4 Admission Decisions

6.2.4.1 General Approach

Consider the CMTS upstream first. The CMTS upstream has a fixed amount of bandwidth. Within this bandwidth, the CMTS must constantly schedule all traffic on an ongoing basis. This scheduling includes not only all voice and data traffic, but all DOCSIS overhead such as signaling, request slots, initial maintenance slots, station maintenance slots, and so on. Further, there are variances in the mapping of packets to DOCSIS minislots that can result from concatenation and fragmentation.

The approach to admission control is to establish a series of thresholds, each with a set of rules. These thresholds are a percentage of the maximum amount of voice traffic that the upstream will handle. The rules will relate to the priority levels and various types of voice traffic.

Choosing the value of the threshold is the task of traffic engineering. Using customer penetration numbers, call densities, and erlang tables, the traffic engineer will generate the number of calls that must be supported on an upstream configuration. They will also generate the maximum usage of an upstream that can be used for voice in order to provide the specified data performance and a maximum usage of an upstream if there is no data. These usage numbers can be mapped to the different admission thresholds.

In theory, there can be an infinite number of thresholds and priorities. In practice, it helps to pick a small number of thresholds that all applications can map into. This white paper suggests three fundamental thresholds based upon expected voice usage on DOCSIS.

6.2.4.2 Normal Traffic

As an example, consider a standard DOCSIS upstream running QPSK modulation with a symbol rate of 1280 ksps that results in 1.6 MHz of RF bandwidth and 2.56 Mbps of raw bandwidth. With 100-percent voice utilization, and depending upon specific DOCSIS physical level characteristics, that upstream will handle 23 G.711 (64 kbps) calls or 50 G.728 (16 kbps) calls. However, at 100-percent utilization, there would be no room for data or DOCSIS overhead, so a lower operating threshold must be chosen.

To continue the example, a fiber node might require 25 G.728 voice connections based upon customer penetration and call densities. Data performance requirements and DOCSIS signaling in this example limit the upstream to 65 percent voice with data present, or 80 percent with no data present. Taking 65 percent of the 50 call limit of the upstream provides an operating limit of 32 calls. However, the actual call load is 25 / 50 = 50 percent.

The offered load is 50 percent (25 calls). The available bandwidth is 65 percent (32 calls). The total bandwidth is 100 percent (50 calls).

The traffic engineer has two choices available for an admission threshold for normal voice traffic, 50 percent and 65 percent. The first threshold for voice calls is thus set at 50 percent, allowing the remaining 15 percent to be used for other voice services.

This first threshold shall be referred to as the voice operating threshold. All new calls will be accepted until the aggregate bandwidth of all calls meets the voice operating threshold. Then all new calls with normal admission characteristics will be denied service.

6.2.4.3 Multiple Flow Specs and Exception Conditions

There are a variety of exception conditions that may merit higher-priority treatment. One of these exception conditions is the DOCSIS multiple flow specs. The multiple flow specs allow a cable modem to make two reservations, each at a different rate, but use only one at a time. The significance of this is that on a fully loaded system, reservations that increased their bandwidth should not be denied since the user has already been admitted. For example, if someone initiated a G.728 voice call, he or she could prebook G.711 in case the call flipped into a fax call. The ridiculous case to deal with is where 25 G.728 calls get placed, and all of them might flip up to G.711. How much bandwidth is required?

Another example of the use of multiple flow specs is when a G.711 call is established, and a G.728 call comes in on call waiting. The called party flips over to the new call and receives a G.728 reservation. It may possible at this time for new calls on the system to take the released network bandwidth and fully book the available normal voice bandwidth. The goal of the multiple flow spec feature is to allow the called party to flip back to the higher-bandwidth G.711 call even though the network is not allowing any new call setups.

These cases represent exception conditions. Even though new calls are being denied service, providing existing calls with extra bandwidth is reasonable, especially since there was a flow spec requesting service. The amount of extra bandwidth required depends upon the nature of the exception. In the case of G.728 connections flipping into G.711 fax calls, that would be related to how many subscribers actually have a fax machine and how often that fax machine is in use matched to the desire to block calls only 1 percent of the time. This setup is handled by introducing a second threshold. Traffic engineering chooses to set that threshold in this example at 65 percent.

This second threshold shall be referred to as the voice exception threshold. Existing calls may increase their bandwidth based upon multiple flow specs until the aggregate bandwidth of all calls meets the Voice Exception Threshold. In addition, new calls that meet admission characteristics considered as exception conditions may also be admitted until the Voice Exception Threshold is met.

6.2.4.4 Emergency Services

Emergency services calls such as 911 calls present a special case where the call should be placed through the network, even it means first preempting data and then preempting other normal voice traffic. This setup is handled with a third threshold. In this example, traffic engineering is going to set the threshold equal to the maximum voice traffic with no data that is 80 percent. Thus 911 calls would preempt data.

This third threshold shall be referred to as the voice preemption threshold. New calls that meet the admission characteristics considered for data preemption, such as emergency services calls, will be admitted up to the voice preemption threshold. After that point, the CMTS may choose to begin dropping lower-priority voice calls in order to make room for new 911 calls. This setup would be another policy decision at the CMTS.

6.2.4.5 VAD

VAD (Voice Activity Detection) is a voice algorithm that allows silence intervals to be suppressed and not transmitted. For DOCSIS to take advantage of VAD, a specific VAD reservation must be made in DOCSIS that runs a scheduling algorithm that alternates between polling and unsolicited grants. For this reason, the CMTS specifically has to know whether VAD is active or not.

The goal is to provision VAD without introducing significant blocking. Two criteria are important:

- That a reasonable maximum number of VAD calls fit within the maximum allowable bandwidth; this setup prevents call blocking under extreme circumstances, although data blocking may occur
- That the nominal number of VAD calls fits within the target voice region; this setup statistically allows enough bandwidth for data services

In this example, in the upstream, the total bandwidth is 50 calls and the available bandwidth for voice and data is 40 calls. The new example offered load is 45 VAD calls. Will this system work, and what is the performance?

In this example, traffic engineering has two choices to accommodate the peak 45 calls. Spend money and add another upstream to increase the call capacity to 80 calls, or accept potential blockage of 5 VAD calls.

It would be rare for all 45 VAD calls to be active at once. That model predicts that 45 total VAD calls will produce 28 active VAD calls. This VAD activity factor would be 62 percent (28/45).

A single upstream channel would block 5 VAD calls, or 11 percent (5/45) of the offered VAD traffic. The VAD nonblocking number of 89 percent (40/45) is well above the nominal VAD call activity level of 62 percent. This level satisfies the first criterion of a reasonable maximum.

The nominal VAD offered load of 28 calls provides a nominal upstream loading of 56 percent (28/50), a level below the 65 percent preferred limit. This level matches the second criterion.

Traffic engineering management chooses to save money. The CMTS is given a voice operating threshold of 80 percent. The voice exception threshold and voice preemption thresholds are also set at 80 percent.

6.2.4.6 Payload Header Suppression

Payload header suppression (PHS) is a DOCSIS protocol invented by the author of this paper that allows the repetitive bytes in the Ethernet and TCP/IP headers to be suppressed at the transmit location and restored at the receive location. Payload header suppression is intended primarily for the HFC upstream, but can be used in the downstream as well.

Typically, for an RTP packet in the upstream, the Ethernet header and the entire TCP/IP header (with the possible exception of the 2-byte checksum) will be suppressible. In the downstream, only the TCP/IP header can be suppressed since the cable modem filters on the MAC address of the data frame. In the upstream, PHS leaves a DOCSIS header, and RTP header, the payload, and a CRC. Non-RTP packets typically will not be suppressed as they will have a large variance in header values.

Payload header suppression is an option at the CMTS, and external reservation protocols are unaware if the CMTS will actually use it. The use of payload header suppression will impact the call densities in the upstream and downstream. Thus it will impact the number of calls supported for a given admission level and must be allowed for in the call density calculations by traffic engineering.

6.2.5 Putting It All Together

Bandwidth requests arrive at the CMTS from protocols such as RSVP and DOCSIS. Voice bandwidth requests contain the voice raw bit rate and traffic attributes but do not make any attempt to interpret the impact these attributes have on the final bit rate. The CMTS combines the information in the bandwidth requests with its own internal state information to arrive at a decision of how much bandwidth to use for the call and whether to admit it or not.

The CMTS manages its upstream admission control using a series of thresholds, each with its own rule set. The admission objects in the RSVP requests are mapped into these thresholds. These thresholds correspond to bandwidth limits that depend upon the physical parameters of the channel and link protocols such as header suppression.

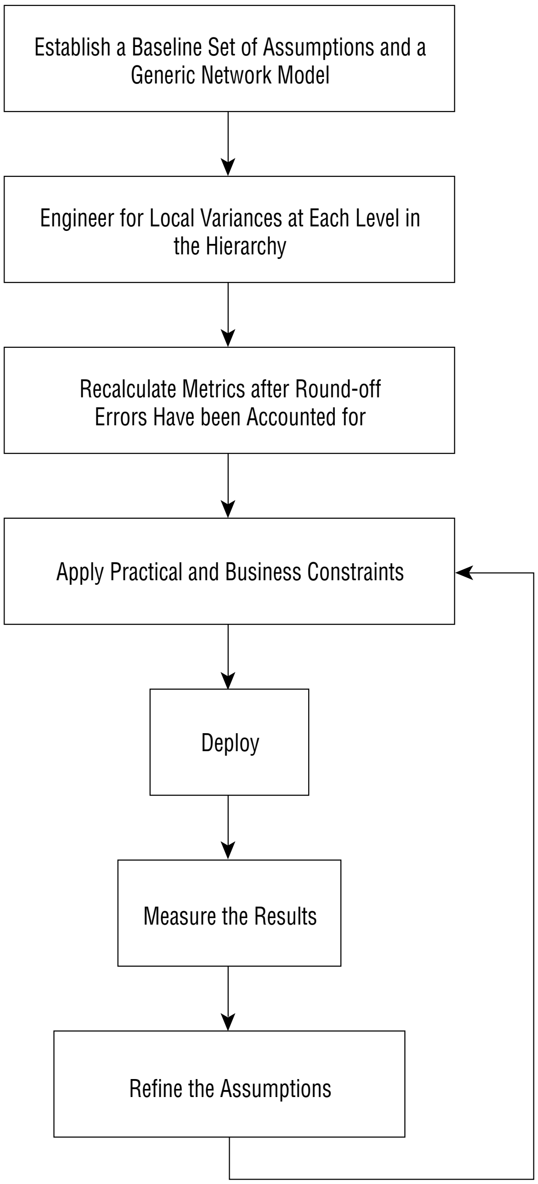
Three proposed thresholds along with their rule set follow:

- Voice operating threshold
- Voice exception threshold
- Voice preemption threshold

New voice calls with normal priority are accepted until the voice operating threshold is reached. Higher-priority calls and known exception cases such as multiple call flows are allowed up to the higher voice exception threshold, and emergency service calls such as 911 are admitted up to the voice preemption threshold. If more emergency service calls are received, the CMTS may decide to actually preempt voice calls with lower priority.

7 Summary

There is a saying in the statistical industry that there are lies, damn lies, and then statistics. Traffic engineering has its roots in statistics and assumptions, all the way from what the market penetration will be to the Poisson distribution used in the erlang B formula. No two fiber nodes are really the same size. They vary, based upon geography and age of the plant. Instead of always being 500 HP, they could range from 20 to 2000 HHP or more. The combination of fiber nodes to hubs to headends is always unique for the same reasons. Even market penetrations will vary, based upon demographics, pricing, or available competition.



The objectives of the traffic engineering on the HFC plant were to map fiber nodes to ports on a CMTS. That mapping introduced huge inconsistencies. The choice between using one or two upstream receivers could introduce a 100-percent error factor. One of the typical traffic engineering mistakes is to assume a completely generic plant, engineer for that plant, and then apply the results blindly to all plants from that point forward.

The goal of this paper was to present a balanced approach between theoretical and engineering requirements on one side, and practical and business requirements on the other side. With such round-off errors existing, and such a large amount of assumptions, doing a generic traffic engineering design once is only the first step.

Traffic engineering should be an ongoing process. First a baseline set of assumptions should be made as well as a first-pass at a generic network model. Then the engineering should be redone at a more local level to accommodate local differences in fiber nodes, hubs, and headends.

This scenario will generate round-off errors, so the next step is to take the actual design, generate metrics, and compare those to the original assumption. Once these calculations are made, practical constrains such as cost and maintenance should be considered.

The next step is to deploy and measure the results. Only by measuring the results will the network provider know which areas of the network will need upgrading first, and which areas may never need upgrading. This step also provides better statistics for generating design and marketing assumptions for the next installation.

Appendix A: Definitions

ABSBH

ABSBH is an acronym for Average Busy Season Busy Hour. The Average Busy Season Busy Hour is defined as the average traffic load in the busy hour of the Average Busy Season. Busy season data excludes Mother's Day, Christmas, or extremely high traffic days that can be attributed to unusually severe weather or catastrophic events, and are not reasonably expected to recur from year to year. Generally (but not always), the busy season data excludes weekends. The ABSBH represents the day's greatest number of calls per unit time that a network element will have to process with unimpaired performance.

Busy Hour

Busy Hour is the hour during the day that contains the peak amount of traffic. For residential voice, this tends to be about 7:00 p.m. Unfortunately, this is also the busiest time for data as well. An interesting rule found in reference [1] is that the traffic during busy hour represents 15 to 17 percent of the traffic during a one-day period.

CAT

CAT is an acronym for Call Attempts, usually for a one hour period of time.

CMS

CMS is an acronym for call management server. A CMS manages call control on a VoIP network.

CMTS

CMTS is an acronym for Cable Modem Termination System.

CMTS Domain

In DOCSIS, there is an association between a downstream frequency and multiple upstream frequencies, referred to in this white paper as a CMTS domain. Typically these domains are “1 down, n up,” where n is from 1 to 8. In theory, there can be any number of downstreams grouped with any number of upstreams.

In practice, a cable modem (CM) will be connected to a CMTS domain. The CMTS equipment will contain some number of these domains. A CMTS domain may be part of a line card, an entire line card, or spread across several line cards.

CCS

CCS is an acronym for Calls per Centum Seconds. A centum second represents 100 seconds. There are 3600 seconds in an hour, so there are 36 centum-seconds in one hour. An erlang is dimensionless but is usually referenced to one hour, so one erlang is equivalent to 36 CCS.

Ceiling

A ceiling is a mathematical function that rounds up a whole number to the nearest integer.

Coded

Codec is an acronym for coder/decoder. A codec converts voice from analog to digital, usually with an 8-kbps sampling clock.

Data Rates

A peak data rate represents the maximum allowed data rate on the media when no other subscribers are present, and is implemented with rate shaping at the CMTS. The data rate for a 5-minute interval represents the committed average data rate that a subscriber will receive during data busy hour. The data rate during a 1-second interval represents the committed peak data rate that a subscriber will receive during data busy hour.

DOCSIS

DOCSIS is an acronym for Data over Cable System Interface Specification.

DPT

DPT is an acronym for dynamic packet transport.

DWDM

DWDM is an acronym for dense wave division multiplexing.

Erlang

An erlang describes call densities. One erlang is the maximum physical capacity of a line or trunk. Erlangs are dimensionless, but are usually referenced to one hour. One erlang for one hour is equivalent to 36 CCS, which is the equivalent of 60 minutes of connect time; it could be one 60-minute call in an hour or ten 6-minute calls over a period of 6 minutes.

Fiber Node (FN)

A fiber node interfaces between a point-to-point analog fiber link and a local analog coax plant. Upstream from a fiber node is a hub.

Floor

A floor is a mathematical function that rounds down a whole number to the nearest integer.

Grade of Service (GoS)

GoS describes the blocking probability of a collection of trunk circuits. A GoS of 1 percent means that one call will be blocked out of 100 call attempts. Since voice networks are designed to handle busy hour, HDBH, or ABSBH, this level of blockage would occur only during those times.

Headend (HE)

The headend is the location in the cable plant where the video feeds are brought in, modulated, and transmitted into the local receiving area. Downstream from a headend is a hub.

High Day Busy Hour (HDBH)

HDBH is the one day that has the highest traffic during the busy hour. The traffic level in the busy hour of the high day is termed the "HDBH load." Practically speaking, the HDBH occurs in the first working day after the Thanksgiving holiday. Historically, the traffic in the HDBH is approximately 20 percent greater than during the ABSBH.

Households Passed (HHP)

HHP is used to measure the size of an HFC plant serving area. HHP refers to the number of homes that the coax passes and thus the number of homes that the MSO could potentially provide service to.

Hub

The hub is a concentration point between fiber nodes and the headend. The connection to the fiber nodes is usually point-to-point analog fiber. The connection to the headends is usually a digital fiber ring.

Hybrid Fiber-Coaxial (HFC)

HFC is the description of the modern day cable plant, which consists of a digital and analog fiber backbone with coax at the edges.

Ineffective Call Attempts

An ineffective call attempt within a local switching system is any valid bid for service that does not complete or outpulse, as appropriate, in the normal manner, including false starts and permanent signals.

- False starts occur when the calling party picks up the receiver and hangs up before dialing digits.
- Permanent signals occur when the calling party picks up the receiver and leaves it off hook.

Ineffective call attempts are not counted as completed calls.

Internet Protocol (IP)

IP is the standard Layer 3 transport protocol for the Internet.

ksps

ksps is an acronym for kilosymbols per second. It is the unit of measure for the symbol rate of QPSK or QAM modulation.

Lines

Also known as subscriber lines, lines allow telephone sets to be connected to telephone switches, like private automated branch exchanges (PABXs) and central office (CO) switches.

MAC

MAC is an acronym for Media Access Control. MAC is often used interchangeably to refer to the data link layer (Layer 2) of the OSI stack.

MPEG

MPEG is an acronym for Motion Picture Experts Group and is usually associated with the video compression work that the group has done.

MSO

MSO is an acronym for multisystem operator. The term MSO is used to refer to the companies that provide cable access service.

PABX

PABX is an acronym for private automated branch exchange.

PHY

PHY is an acronym for physical layer. PHY is often used interchangeably to refer to the physical layer (Layer 1) of the OSI stack.

POTS

POTS is an acronym for plain old telephone service and refers to standard analog phone service delivered over a twisted pair.

QAM

QAM is an acronym for quadrature amplitude modulation.

QPSK

QPSK is an acronym for quadrature phase shift keying.

TGW

TGW is an acronym for trunk gateway. A trunk gateway connects between a circuit-switched environment such as T1 and a packet or ATM switched environment such as Ethernet or ATM over OC-3. A trunk gateway converts between circuit-switched voice and packet-switched voice, as well as performing voice compression, VAD, de-jittering, and echo cancellation.

Trunks

Trunks interconnect switching equipment. Typically a piece of voice switching equipment such as a PABX will have a line side that connects to the subscriber's telephones, and a trunk side that connects to the next level of switching hierarchy. One significance of trunks is that they carry concentrated traffic. This concentration is usually expressed as the line-to-trunk ratio.

uBR

uBR is an acronym for the universal broadband router. The uBR is the Cisco Systems product name for a CMTS and CM.

VAD

VAD is an acronym for Voice Activity Detection. It is also known as silence suppression. When VAD is enabled, silent intervals are not packetized and sent.

VoIP

VoIP is an acronym for voice over IP. It refers to packetizing voice and transporting it over an IP network.

Appendix B: Formula Summary

Background

Telephony Traffic Engineering Basics

Formula (1) $$P_r = \frac{\frac{A^N}{N!}}{\sum_{J=0}^{N} \frac{A^j}{j!}}$$

Formula (2) *N = Erlang-B[A, Pr]*

Formula (3) *N = Erlang-BL[N, Pr] = Max[N / 90%, Erlang-B[N, Pr]]*

Formula (4) *Call Attempts per Hour = Erlangs * 3600 / Call Holding Time*

Formula (5) *Erlang = CCS / 36*

HFC Plant Layout

Headends, Hubs, and Fiber Nodes

Formula (6) *HHP per HE = HHP per FN * FN per Hub * Hubs per HE*

Formula (7) *HHP per Hub = HHP per FN * FN per Hub*

Equipment Description

DOCSIS CMTS

Formula (8) *BW Raw = Bits per Symbol * Symbol Rate*

Formula (9) *BW Usable = BW Raw * (1 – PHY and MAC Overhead)*

HFC Traffic Sizing

Data

Formula (10) *Data Asymmetry Worst Case = (2 * Average Downstream Packet Size) / (Average Upstream Packet Size)*

Formula (11) *Active Subscribers = HHP * Data Penetration * Data Average Activity Factor*

Formula (12) *Peak Subscribers = HHP * Data Penetration * Data Peak Activity Factor*

Formula (13) *BW Data Average = (Active Subscribers * Bandwidth per Subscriber) * (1 + MAC and PHY Overhead)*

Formula (14) *BW Data Peak = (Peak Subscribers * Bandwidth per Subscriber) * (1 + MAC and PHY Overhead)*

Formula (15) *BW Data = Max[BW Data Average, BW Data Peak]*

Voice

Derivation of Subscriber Density Numbers

Formula (16) *Total Market Penetration = S Separate Market Penetrations - Overlap*

Formula (17) *Nth Line Relative Penetration = Line Penetration / First Line Penetration*

Formula (18) *Lines per Home = S Line Market Penetration / First Line Penetration*

Formula (19) *Lines per Home = S Line Relative Penetration / 100%*

Formula (20) *Traffic per Line = S (Traffic per Line * Market Penetration) / (First Line Penetration * Lines per Home)*

Formula (21) *Traffic per Line = S (Traffic per Line * Relative Penetration) / Lines per Home*

Formula (22) *Codec Share of Traffic = Â (Traffic per Line * Relative Penetration per Codec) / (Lines per HHP * Traffic per Line)*

Connectivity

Formula (23) *Lines per Node = HHP per Node * Market Penetration of Voice * Lines per HHP*

Formula (24) *Traffic per Node = Lines per Node * Traffic per Line*

Formula (25) *Connections per FN = Erlang-BL (Erlangs per Node, Grade of Service)*

Formula (26) *Line-to-Trunk Ratio = Lines per FN / Connections per FN*

Multiple Codecs

Formula (27) *Connections per Upstream Max = Floor [Â (Connections per Upstream per Codec * Codec Share of Traffic)]*

Formula (28) *BW Voice = Bandwidth per Upstream * Connections per FN / Connections per Upstream*

Video

Formula (29) *IPVS Users = HHP per FN * FN per IPVS Downstream Group* IPVS Market Penetration * IPVS Activity Factor * IPVS Sessions per Home*

Formula (30) *IPVS Unicast Sessions = IPVS Users * IPVS Unicast Usage * 1 User per Unicast Session*

Formula (31) *IPVS Multicast Users = IPVS Users * IPVS Multicast Usage*

Formula (32) *IPVS Multicast Sessions Max = IPVS Multicast Available * IPVS Multicast Used Max*

Formula (33)

$$\text{IPVS Multicast De-rating Factor} = \frac{Min\left[U, Ceiling\left\lceil M \times \left\{1 - exp\left(-\frac{U}{M} \times F\right)\right\}\right\rceil\right]}{M}$$

Formula (34) *IPVS Multicast Sessions = IPVS Multicast Sessions Max * IPVS Multicast De-rating Factor*

Formula (35) *IPVS Sessions = Ceiling [IPVS Unicast Sessions + IPVS Multicast Sessions]*

Formula (36) *IPVS Network BW = IPVS Sessions * BW per IPVS Stream*

Formula (37) *IPVS per RF Channel = Floor [BW per RF Channel / BW per IPVS Session]*

Formula (38) *IPVS RF Channels = Ceiling [IPVS Sessions / IPVS per RF Channel]*

Maintenance and Signaling

Formula (39) *BW M&S = HHP per FN * Total Penetration * Overhead per Modem*

HFC Equipment Sizing

Equipment Sizing Based upon Plant Topology

Formula (40) *CMTS per HE Min = Hubs per HE * 1 CMTS per Hub*

Formula (41) *HHP per CMTS Max = HHP per HE / CMTS per HE Min*

Equipment Sizing Based upon Upstream Combining

Formula (42) ***FN per Receiver Max 1 = Min [Floor (HHP per Upstream Limit /HHP per FN), FN per Upstream Limit]***

Equipment Sizing Based upon Bandwidth

Upstream Connectivity—A Theoretical Approach

Formula (43) ***BW Total = BW Data + BW Voice + BW Video + BW M&S***

Formula (44) ***Receivers per FN Min = Ceiling(BW Total Upstream / BW per Receiver)***

Formula (45) ***FN per Receiver Max 2 = Floor (BW per Receiver / BW Total Upstream)***

Formula (46) ***FN per Receiver = Min [FN per Receiver Max 1, FN per Receiver Max 2]***

Upstream Connectivity—A Practical Approach

Maximum Voice Density

Formula (47) ***Data Region Min = Max Data Rate / Usable Channel BW***

Formula (48) ***Maintenance Region = Maintenance Time / Voice Frame Time***

Formula (49) ***Voice Region Target = 100% – Max (Data Region Min, Maintenance Region)***

Matching Fiber Nodes to the CMTS Receivers

Formula (50) ***Receivers per FN = Ceiling [Connections per FN / (Connections per Upstream * Voice Region Target * (1 + Voice Region Allowance))]***

Downstream Connectivity

Formula (51) ***FN per Transmitter Max = Floor(BW per Downstream Raw / BW Total Downstream)***

Adjustment for Voice Activity Detection

Formula (52) ***VAD Region Max Target = (100% - Maintenance Region) * (1 + VAD Region Allowance)***

Formula (53) ***VAD Connections Max = Connections per FN***

Formula (54) ***VAD Receivers per FN = Ceiling (Connections per FN / (Connections per Upstream * VAD Region Max Target))***

Formula (55) ***VAD Connections Nominal = Min [VAD Connections Max, Erlang-BL [VAD GoS, VAD Connections Max * VAD Activity Level Two Way]]***

Formula (56) ***VAD Activity Factor One Way = VAD Connections Nominal / VAD Connections Max***

Formula (57) ***VAD Region Nominal Target = VAD Region Max Target * VAD Activity Factor One Way***

Formula (58) ***VAD Region Nominal Target £ Voice Region Target***

Making an Optimized Decision

Formula (59) ***FN per CMTS Domain = Floor (Receivers per CMTS Domain / Receivers per FN Min)***

Formula (60) ***FN per CMTS = FN per CMTS Domain * CMTS Domains per Line Card * Line Cards per CMTS***

Formula (61) ***HHP per CMTS = FN per CMTS * HHP per FN***

Formula (62) ***Line Card Utilization = (Receivers per FN * FN per Domain / Receivers per Domain)***

The Final Metrics

Formula (63) ***Connections per FN Max = Connections per Upstream Max * Upstreams per FN***

Formula (64) ***Voice Region Actual = Connections per FN / Connections per FN Max***

Formula (65) ***VAD Region Max Actual = Connections per FN / Connections per FN Max***

Formula (66) ***VAD Region Nominal Actual = VAD Region Max Actual * VAD Activity Factor One Way***

Equipment Sizing Based upon Performance

Data

Formula (67) ***PPS Data per Direction = Data Rate / 8 bits per byte / Packet Size***

Formula (68) ***PPS Data per CM = 1.5 (Bandwidth per Subscriber / 8 bits per byte / Packet Size)***

Formula (69) ***PPS per CMTS = PPS per CM * Active Subscribers per FN * FN per CMTS***

Voice

Formula (70) ***PPS Voice per Connection = 2 Directions * 1 / VoIP Frame Interval***

Formula (71) ***PPS Voice per FN = PPS Voice per Connection * Connections per FN * VAD Activity Level Two Way***

Formula (72) ***PPS Voice per CMTS = Voice PPS per FN * FN per CMTS***

Signaling

Formula (73) ***CAT per CMTS = CAT per Line * Lines per FN * FN per CMTS***

Formula (74) ***PPS Signaling per CMTS = Call Attempts per CMTS * CMTS Packets per CAT / 3600 seconds per hour***

Formula (75) ***Ratio Bearer:Signaling = (Call Holding Time * Voice Traffic per Connection) / CMTS Signaling Packets per Call***

Total CMTS Traffic Loading

Formula (76) ***PPS CMTS = PPS Data per CMTS + PPS Voice per CMTS + PPS Signaling per CMTS***

Cable Modem Termination System Sizing

Without Redundancy

Formula (77) ***Working Line Cards per Hub = Ceiling(FN per Hub / (FN per CMTS Domain * CMTS Domains per Line Card)***

Formula (78) ***Working CMTS per Hub = Ceiling(Working Line Cards per Hub / Line Cards per CMTS)***

Adjustment for Redundancy

Formula (79) ***Protect CMTS per Hub = Ceiling(Working CMTS per Hub / N Max)***

Formula (80) ***Protect Line Cards per Hub = Protect CMTS per Hub * Line Cards per CMTS***

Formula (81) ***N Actual = Working CMTS per Hub / Protect CMTS per Hub***

The Result

Formula (82) ***Line Cards per Hub = Working Line Cards per Hub + Protect Line Cards per Hub***

Formula (83) ***Line Cards per HE = Line Cards per Hub * Hubs per HE***

Formula (84) ***CMTS per Hub = Working CMTS per Hub + Protect CMTS per Hub***

Formula (85) ***CMTS per HE = CMTS per Hub * Hubs per HE***

Cable Modem Sizing

Formula (86) ***CM per HE = HHP per HE * Market Penetration Total***

Call Management Server Sizing

Formula (87) ***Call Attempts per HE = Call Attempts per Line * Lines per HE***

Formula (88) ***CMS per HE = Call Attempts per HE / Call Attempts per CMS Max***

Trunk Gateway Sizing

Bandwidth

Formula (89) *TGW Ports per HE = Ceiling(Erlang-BL(Erlangs per HE * off net, Grade of Service))*

Formula (90) *TGW Line Cards per HE = Ceiling (TGW Ports per HE / TGW Ports per Line Card)*

Formula (91) *TGW Chassis per HE = Ceiling (TGW Line Cards per HE / TGW Line Cards per Chassis Max)*

Formula (92) *TGW Line Cards per Chassis = Ceiling (TGW Line Cards per HE / TGW Chassis per HE)*

Performance

Voice

Formula (93) *TGW Bearer PPS = PPS per Connection * TGW Ports per Line Card * TGW Line Cards per Chassis * VAD Activity Level Two Way*

Signaling

Formula (94) *TGW Call Attempts = Call Attempts per Line * Line-to-Trunk Ratio * TGW Ports per Line Card * TGW Line Cards per Chassis*

Formula (95) *TGW Signaling PPS = TGW Call Attempts / 3600 seconds per hour * Packets per CAT*

IP Network Sizing

CMTS Backhaul Bandwidth

Formula (96) *BW Data per CMTS in the Downstream = BW Data Downstream * FN per CMTS*

Formula (97) *BW Data per CMTS in the Upstream = BW Data Upstream * FN per CMTS*

Formula (98) *Connections per CMTS = Connections per FN * FN per CMTS*

Formula (99) *Voice Bytes per Packet = Codec Bite Rate / 8 * VoIP Frame Time*

Formula (100) *BW per Connection = (Voice Bytes per Packet + Packet Overhead + Frame Overhead) * 8 bits per byte / VoIP Frame Time*

Formula (101) *BW Voice per CMTS = Connections per CMTS * BW per Connection * 2 directions*

Formula (102) *BW Total CMTS = (BW Data CMTS + BW Voice CMTS) * (1 + IP signaling Overhead)*

Formula (103) *BW CMTS Backhaul = Max (BW Total CMTS in the Downstream, BW Total CMTS in the Upstream)*

IP Backhaul Strategies

Dynamic Packet Transport

Formula (104) *BW HE = Hubs per HE * Working CMTS per Hub * BW CMTS Backhaul*

Formula (105) *DPT Rings per System = Ceiling (BW HE / BW DPT Ring)*

Formula (106) *DPT Line Cards per HE System = CMTS per Hub * Hubs per HE*

Related Topics

Support for Legacy Analog Modem and Fax

Analog Modem Support

Formula (107) *Analog Modem Bandwidth Ratio = BW Pass-Through / BW Modem Relay*

Formula (108) *Analog Modem Usage Ratio = Usage of Pass-Through / Usage of Modem Relay*

Formula (109) *Analog Modem Symmetry Ratio = 10 X*

Formula (110) *Analog Modem Relay Efficiency Ratio = Bandwidth Ratio * Usage Ratio * Symmetry Ratio*

Formula (111)   Fax Modem Bandwidth Ratio = BW Pass-Through / BW Fax Relay

Formula (112)   Fax Modem Usage Ratio = Usage of Pass-Through / Usage of Fax Relay

Formula (113)   Fax Modem Symmetry Ratio = 8 X

Formula (114)   Fax Modem Relay Efficiency Ratio = Bandwidth Ratio * Usage Ratio * Symmetry Ratio

Appendix C: DOCSIS Bandwidths

The bandwidth that a DOCSIS system provides depends primarily upon channel bandwidth, symbol rate, and modulation techniques. Table C-1 through Table C-3 below show all the DOCSIS possibilities. Secondary considerations such as MAC-layer overhead, signaling messages, and maintenance activities are not included in these tables, but are allowed for in the traffic engineering calculations.

Table C-1 Downstream Bit Rates

RF Bandwidth	Modulation	Symbol Rate	Bit Rate	
			Raw	Payload
6 MHz	64 QAM	5.056941 Msps	30.34 Mbps	~ 27 Mbps
	256 QAM	5.360537 Msps	42.88 Mbps	~ 38 Mbps
8 MHz	64 QAM	6.74 Msps	40.44 Mbps	~ 36 Mbps
	256 QAM	7.15 Msps	57.20 Mbps	~ 51 Mbps

Table C-2 Upstream Bit Rates for DOCSIS 1.0 and 1.1

RF Bandwidth	Symbol Rate	Raw Bit Rate	
		QPSK	16 QAM
200 kHz	160 ksps	320 kbps	640 kbps
400 kHz	320 ksps	640 kbps	1.28 Mbps
800 kHz	640 ksps	1.28 Mbps	2.56 Mbps
1.6 MHz	1280 ksps	2.56 Mbps	5.12 Mbps
3.2 MHz	2560 ksps	5.12 Mbps	10.24 Mbps

Table C-3 Upstream Bit Rates for DOCSIS 1.2

RF Bandwidth	Symbol Rate	Raw Bit Rate				
		QPSK	8 QAM	16 QAM	32 QAM	64 QAM
200 kHz	160 ksps	320 kbps	480 kbps	640 kbps	960 kbps	1.28 Mbps
400 kHz	320 ksps	640 kbps	960 kbps	1.28 Mbps	1.92 Mbps	2.56 Mbps
800 kHz	640 ksps	1.28 Mbps	1.92 Mbps	2.56 Mbps	3.84 Mbps	5.12 Mbps
1.6 MHz	1280 ksps	2.56 Mbps	3.84 Mbps	5.12 Mbps	7.68 Mbps	10.24 Mbps
3.2 MHz	2560 ksps	5.12 Mbps	7.68 Mbps	10.24 Mbps	15.36 Mbps	20.40 Mbps
6.4 MHz	5120 ksps	10.24 Mbps	15.36 Mbps	20.40 Mbps	30.72 Mbps	40.80 Mbps

Appendix D: Voice-over-IP Call Density Tables

Table D-1 and Table D-2 contain the optimized configuration values that are required to achieve a high efficiency of voice traffic in the upstream. Table D-2 indicates the number of VoIP calls that can be established on one upstream. The values in these tables are dependent upon the upstream configuration parameters, and may be different for each system. The tables make the following assumptions:

- Voice scheduling is done with DOCSIS unsolicited grant service.
- All 100 percent of the bandwidth is used for grants. In reality, several voice calls should be subtracted to allow for other interval usage codes (IUC) such as maintenance slots, request slots, and MAC-layer traffic. De-rate accordingly.
- All 100 percent of the bandwidth is used for voice. In reality, the upstream will be provisioned for voice and data, with some upper limit to the voice calls, such as 60 percent. De-rate accordingly.
- Voice Activity Detection (silence suppression) is not active. If VAD is active, the call density potentially doubles. The recommended approach is to limit the number of calls to the maximum specified in Table D-2, and fill in the silence intervals with data traffic.

The equivalent voice bit rate is with respect to the raw bit rate of the channel; it was calculated by dividing the channel raw bit rate by the number of connections slots. The chart includes the equivalent bit rate for the 2.56-Mbps channel. The equivalent bandwidth gets slightly better with higher modulation and symbol rates because the channel efficiency is a bit higher with larger channels.

Some interesting observations may be drawn from these tables:

- The call density for a conservative, QPSK, 1280-ksps, 1.6-MHz, 2.5-Mbps upstream channel with G.711 (64 kbps) encoding, 20-ms packetization period, no VAD, and no payload header suppression is 24 calls. This number is comparable to a T1 link that can also handle 24 calls and has a data rate of 1.544 Mbps. The DOCSIS upstream, however, is many times more flexible and reconfigurable than a standard T1 link.
- Header compression increases call density from 20 percent for large packets (G.711, 20 ms) to 90 percent for small packets (G.729, 10 ms).
- Eight-to-one voice compression (G.711:G.729) results in a call-density increase between 1.7 (10 ms, no header suppression) to 3.4 (20 ms, header suppression).
- Doubling the symbol rate doubles the call density. Going from QPSK to 16QAM doubles the call density.

The most popular upstream packet for data, a 64-byte TCP ACK, is included with the most popular forward error correction code word (FEC CW) setting to indicate the efficiency of the upstream encoding when running both voice and data.

Table D-1 Optimized Upstream DOCSIS Configuration Parameters for VoIP

DOCSIS Parameter	Value
Minislot Size	8 bytes
Shortened Last Code Word	Enabled
Preamble	72 bits
FEC Code Word Size	(See Table D-2)
FEC Error Correction	T = 2, (4 bytes, ~ 8 percent)
Guard Time	8 symbols for 16 QAM, 4 symbols for QPSK
Interleaver Delay	This parameter does not affect VoIP density calculation

Table D-2 Optimized Upstream Voice Profiles

Input			Output			Configuration		Per-Connection Statistics		
Codec	**PHS**	**Length**	**Connections @ BW in Mbps**			**FEC CW**	**Fill**	**VoIP Frame**	**DOCSIS Frame**	**Equiv Bit Rate**
kbps	***on/off***	***ms***	***@ 10.24***	***@ 5.12***	***@ 2.56***	***Bytes***	***Bytes***	***Bytes***	***Minislot***	***kbps***
G.711 (64 kbps)	**on**	**10**	94	47	23	52	4	92	17	109
		20	118	59	29	52	0	172	27	86
	off	**10**	72	36	18	48	0	120	22	141
		20	96	48	24	52	4	200	33	106
G.728 (16 kbps)	**on**	**10**	200	100	50	54	0	32	8	51
		20	290	145	72	52	0	52	11	35
	off	**10**	114	57	28	52	4	60	14	90
		20	188	94	47	52	4	80	17	54
G.729 (8 kbps)	**on**	**10**	228	114	57	52	2	22	7	45
		20	400	200	100	54	0	32	8	26
	off	**10**	123	61	30	52	6	50	13	83
		20	228	114	57	52	4	60	14	45
TCP ACK	**on**	**n/a**	n/a	n/a	n/a	52	2	12	5	n/a
	off	**n/a**	n/a	n/a	n/a	52	2	64	12	n/a

Where:

Codec Coder/decoder: Specified the compression algorithm used for voice and at what bit rate.

PHS Payload Header Suppression.
- The Ethernet header is suppressed from 18 bytes to 4 bytes.
- The IP header is suppressed from 20 bytes to 0 bytes.
- The UDP header is suppressed from 8 bytes to 0 bytes.
- The RTP header is left at 12 bytes.

PHS allows for different configurations, a scenario that would then lead to different voice densities.

Length Recording time of the voice sample.

Connections The number of voice connections that a 100-percent utilized upstream would support. Conversion from modulation and symbol rates to raw bit rates are described in “Upstream Bit Rates for DOCSIS 1.0 and 1.1” on page 67.

FEC CW Forward Error Correction Code Word. This is the Reed-Solomon block used for error correction. A code word contains both the payload and the error correction bytes. For example, a 52-byte code word with T=2 would have 48 bytes of payload and 4 bytes of error correction.

Fill These are the number of bytes used as fill to round up the DOCSIS frame to the nearest last code word and nearest minislot boundary. Unoptimized profiles can cause this number to get as large as 31 bytes or more.

VoIP Frame This includes the voice payload, RTP, UDP, IP, and Ethernet overhead if present.

DOCSIS Frame This is the finished result of a VoIP frame with all the DOCSIS PHY and MAC overhead.

Equiv Bit Rate The total number of bytes in a DOCSIS VoIP frame, including all protocol, FEC, preamble, and fill bytes, divided by the length interval.

Appendix E: Erlang B Tables

Table E-1 Erlang B with 1.0 percent Grade of Service

Erlang	Trunk	Erlang	Trunk	Erlang	Trunk	Erlang	Trunk
1	5	46	59	91	108	136	155
2	7	47	61	92	109	137	156
3	8	48	62	93	110	138	157
4	10	49	63	94	111	139	158
5	11	50	64	95	112	140	159
6	13	51	65	96	113	141	160
7	14	52	66	97	114	142	161
8	15	53	67	98	115	143	162
9	17	54	68	99	116	144	163
10	18	55	69	100	117	145	164
11	19	56	70	101	118	146	166
12	20	57	71	102	119	147	167
13	22	58	73	103	121	148	168
14	23	59	74	104	122	149	169
15	24	60	75	105	123	150	170
16	25	61	76	106	124	151	171
17	27	62	77	107	125	152	172
18	28	63	78	108	126	153	173
19	29	64	79	109	127	154	174
20	30	65	80	110	128	155	175
21	31	66	81	111	129	156	176
22	32	67	82	112	130	157	177
23	34	68	83	113	131	158	178
24	35	69	84	114	132	159	179
25	36	70	85	115	133	160	180
26	37	71	87	116	134	161	181
27	38	72	88	117	135	162	182
28	39	73	89	118	136	163	183
29	40	74	90	119	137	164	184
30	42	75	91	120	138	165	185
31	43	76	92	121	139	166	186
32	44	77	93	122	140	167	187
33	45	78	94	123	142	168	188
34	46	79	95	124	143	169	189
35	47	80	96	125	144	170	190
36	48	81	97	126	145	171	191
37	49	82	98	127	146	172	192
38	51	83	99	128	147	173	194
39	52	84	100	129	148	174	195
40	53	85	101	130	149	175	196
41	54	86	103	131	150	176	197
42	55	87	104	132	151	177	198
43	56	88	105	133	152	178	199
44	57	89	106	134	153	179	200
45	58	90	107	135	154	180	201

(After 180, Trunks = Erlangs / 0.9 percent when using 90-percent limiting)

Table E-2 Erlang B with 0.5 percent Grade of Service

Erlang	Trunk	Erlang	Trunk	Erlang	Trunk
1	5	46	62	91	111
2	7	47	63	92	113
3	9	48	64	93	114
4	11	49	65	94	115
5	12	50	66	95	116
6	14	51	68	96	117
7	15	52	69	97	118
8	16	53	70	98	119
9	18	54	71	99	120
10	19	55	72	100	121
11	20	56	73	101	122
12	22	57	74	102	123
13	23	58	75	103	124
14	24	59	76	104	125
15	26	60	78	105	127
16	27	61	79	106	128
17	28	62	80	107	129
18	29	63	81	108	130
19	30	64	82	109	131
20	32	65	83	110	132
21	33	66	84	111	133
22	34	67	85	112	134
23	35	68	86	113	135
24	36	69	87	114	136
25	38	70	89	115	137
26	39	71	90	116	138
27	40	72	91	117	139
28	41	73	92	118	140
29	42	74	93	119	142
30	44	75	94	120	143
31	45	76	95	121	144
32	46	77	96	122	145
33	47	78	97	123	146
34	48	79	98	124	147
35	49	80	100	125	148
36	51	81	101	126	149
37	52	82	102	127	150
38	53	83	103	128	151
39	54	84	104	129	152
40	55	85	105	130	153
41	56	86	106	131	154
42	57	87	107	132	155
43	59	88	108	133	156
44	60	89	109	134	158
45	61	90	110	135	159

Table E-3 Erlang B with 0.5 percent Grade of Service

Erlang	Trunk
136	160
137	161
138	162
139	163
140	164
141	165
142	166
143	167
144	168
145	169
146	170
147	171
148	172
149	173
150	174
151	176
152	177
153	178
154	179
155	180
156	181
157	182
158	183
159	184
160	185
161	186
162	187
163	188
164	189
165	190
166	191
167	192
168	194
169	195
170	196
171	197
172	198
173	199
174	200
175	201
176	202
177	203
178	204
179	205
180	206
181	207
182	208
183	209
184	210
185	211
186	212
187	213
188	215
189	216
190	217
191	218
192	219
193	220
194	221
195	222
196	223
197	224
198	225
199	226
200	227
201	228
202	229
203	230
204	231
205	232
206	233
207	234
208	235
209	237
210	238
211	239
212	240
213	241
214	242
215	243
216	244
217	245
218	246
219	247
220	248
221	249
222	250
223	251
224	252
225	253
226	254
227	255
228	256
229	257
230	258
231	260
232	261
233	262
234	263
235	264
236	265
237	266
238	267
239	268
240	269
241	270
242	271
243	272
244	273
245	274
246	275
247	276
248	277
249	278
250	279
251	280
252	281
253	282
254	283
255	284
256	286
257	287
258	288
259	289
260	290
261	291
262	292
263	293
264	294
265	295
266	296
267	297
268	298
269	299
270	300

(After 270, Trunks = Erlangs / 0.9 when using 90-percent limiting)

Appendix F: Excel Spreadsheet Erlang Macros

```
Function Trunks(Erlangs, blockage)
'*********************************************************************
'*********************************************************************
'******
'****** This function calculates the number of trunks necessary
'****** in a network to achieve a
'****** and User-Defined maximum acceptable blockage rate
'****** when given the number of Erlangs on a network
'****** INPUTS: It accepts "Erlangs" and "Blockage"
'****** OUTPUT: It delivers "Trunks"
'******
'*********************************************************************
'*********************************************************************
    calcBlockage = 1
    nmbr_trunks = 0
    If (blockage < 0.0001) Then
        nmbr_trunks = -9999
    ElseIf (blockage > 0.99999) Then
        nmbr_trunks = -9999
    ElseIf (Erlangs < 0.00001) Then
        nmbr_trunks = 0
    Else
        While (calcBlockage > blockage)
            nmbr_trunks = nmbr_trunks + 1
            calcBlockage = Erlangs * calcBlockage
            calcBlockage = (calcBlockage /(calcBlockage + nmbr_trunks))
        Wend
    End If
    Trunks = nmbr_trunks
End Function


Function blockage(Erlangs, Trunks)
'*********************************************************************
'*********************************************************************
'******
'****** This function calculates the blockage on a
'****** network with a given number of trunks and
'****** traffic load (In Erlangs)
'****** INPUTS: It accepts "Erlangs" and "Trunks"
'****** OUTPUT: It delivers "Blockage"
'******
'*********************************************************************
'*********************************************************************
    countTrunks = 1
    blockage = 1
    While (countTrunks <= Trunks)
        junk = Erlangs * blockage
        blockage = junk / (junk + countTrunks)
        countTrunks = countTrunks + 1
    Wend
End Function
```

Appendix G: Biography and Acknowledgements

Biography

John T. Chapman graduated from the University of Alberta in Canada and began his career in 1984 designing analog line and trunk cards for the European market at ROLM / IBM / Siemens.

In 1989, John became an employee of Cisco Systems. John's first project included inventing the 52-Mbps HSSI (High-Speed Serial Interface), which is now an ANSI and ITU standard. Later projects included designing the 60- and 26-pin multiprotocol serial interfaces used on all Cisco routers, the Cisco 3000 product line including the industry's first ISDN router, coarchitecting the VIP card in the Cisco 7500 product line, as well as designing and managing the Mueslix Serial ASIC, which is widely being used in over 40 Cisco Systems products.

Since helping start the HFC Cable group at Cisco Systems in 1996, John remains the system architect. After helping define the data-over-cable effort, John has focused on evolving VoIP technology for the HFC environment. John has been a significant contributor to industry specifications including DOCSIS (PHS, VAD, DCC) and PacketCable. John currently holds the title of Cisco Distinguished Engineer. John holds 4 U.S. patents, 2 international patents, and has 14 patents pending.

John is blessed with a beautiful daughter, Mikayla Shae, and a very loving and supportive wife, Karean. One of his favorite pasttimes is white water canoeing; John currently holds a master level 5th degree black belt in the martial art of Tae Kwon Do as well as a 4th degree black belt in the martial art of Hapkido.

Acknowledgments

I would like to thank the many individuals who reviewed this paper and offered constructive comments.

Many thanks to Bill Lester, Gagan Choudhury, Rich Farel, and Dan Hedyman from AT&T. A special thank you to Rich, who contributed the 90-percent limit to the Erlang formula.

Many thanks to Jamie A. Chaikin of Telcordia for adding clarity to the definitions.

Many thanks to Victor Hou from General Instruments.

Last but not least, many thanks to Vojislav Vucetic from Cisco Systems.

Appendix H: References

[1] Cisco Systems, "Voice Design and Implementation Guide for the MC3810," 1998, http://www.cisco.com/warp/public/788/7.html

[2] CableLabs, *"DOCSIS 1.1 RF Interface Specification,"* SP-RFIv1_1-I03-991105, Interim version 3.0, November 5, 1999.

[3] CableLabs, *"PacketCable Specifications"*, Initial Release 1.0, November 1, 1999.

[4] Telcordia, TR-TSY-000517 (LSSGR Section 17) Issue 3, March 1989, Revision 1, May 1995.

[5] Telcordia, *"LSSGR: Common Section 11, Service Standards,"* TR-TSY-000511, Issue 2, July 1, 1987.

[6] AT&T, *"The AT&T Worldwide Intelligent Network – Facts and Figures,"* 1998, http://www.att.com/network/netfacts98.html

Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems Europe s.a.r.l.
Parc Evolic, Batiment L1/L2
16 Avenue du Quebec
Villebon, BP 706
91961 Courtaboeuf Cedex
France
http://www-europe.cisco.com
Tel: 33 1 69 18 61 00
Fax: 33 1 69 28 83 26

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Headquarters
Nihon Cisco Systems K.K.
Fuji Building, 9th Floor
3-2-3 Marunouchi
Chiyoda-ku, Tokyo 100
Japan
http://www.cisco.com
Tel: 81 3 5219 6250
Fax: 81 3 5219 6001

Lit# 953934 Release 1.0 12/99 LW