Designing Multi-Service Cable Networks

Marty Glapa, Chia-Chang Li, Amit Mukhopadhyay, Cezary Purzynski, & Bill Stenger
Lucent Technologies

Abstract

How to manage access bandwidth to accommodate multiple services with diverse bandwidth requirements, while meeting performance objectives and maximizing revenues/profits is an area of great interest in the cable industry. We have developed a network engineering tool for efficient and effective multi-service cable network design. In addition, the tool can be used to effectively evaluate the impact of service definitions and growth scenarios on bandwidth utilization and performance. In this paper, we discuss the principles of multi-service network design, review some performance considerations, and present the capabilities of the network engineering tool with example results.

1. Introduction

Today's cable services are primarily entertainment based but there is a rapid expansion underway into high-speed data and telephony. The traditional, and still the primary business of the cable industry is one-way broadcast television service delivering up to about 80 channels with more on the way. Pay channel subscriptions that are broadcast over the network have been successfully selling monthly subscription services for about the past 25 years. Pay Per View works similar to the pay channel subscriptions except there is a billing reporting mechanism – often done by polling the subscriber's set top box by telephone or a store and forward RF system using two–way compatible plant. Modern addressable set top boxes have allowed cable operators to scramble analog channels and offer them as a tiered service. MPEG-2 (Moving Picture Experts Group) compression technology and QAM (Quadrature Amplitude Modulation) have enabled the cable operators to offer broadcast digital channel with tiered services, which too are gaining in popularity. A true departure from traditional entertainment video is high-speed data with well over 1 million subscribers in service. It allows a subscriber with a cable modem to receive and transmit high-speed data at rates far beyond those of traditional analog modems.

There are many competitive pressures on cable companies today that can effect the set of core services that are offered. To help combat the competitive pressures, cable operators are looking to new future services to retain and attract customers. These services include a mixture of video services, telephony, and multimedia services. Narrowcast ad insertion could be targeted to specific subscribers based on an individual customer's profile. Circuit based telephony, while being deployed today in some markets, will be quickly replaced by IP based telephony. IP telephony leverages a common logical layer infrastructure (the IP layer) for both voice and data services – resulting in lower capital and operations costs. Video on demand is surfacing as a narrowcast service that allows individual subscribers to order, and watch a movie or event. Work-at-home, and ideal service for an IP based infrastructure, extends LAN and PBX functionality to a workstation in an employee's home. Home security monitoring and energy monitoring/management services can be rather easily deployed using an HFC infrastructure. And, streaming video, video conferencing, video telephony and interactive gaming can open new opportunities for services delivered using an IP data infrastructure.

Each new service will have its own set of bandwidth requirements. These new services will compete for bandwidth in both up and downstream directions on cable plant. Managing this bandwidth, especially those services enabled by a common logical IP layer, to maximize efficiency and minimize cost is of great interest to the industry and is the topic of the remainder of this paper.

2. Service and Traffic Scenarios

As indicated in the previous section, cable networks today support a limited set of services, i.e., broadcast video, Pay Per View, Video On Demand, data services, and, in some cases, circuit voice services. These services typically occupy different part of the spectrum. They coexist on the same HFC plant but are engineered and provisioned separately.

The introduction of DOCSIS [1] brought about an Internet Protocol platform on cable networks that enables multiple service offerings. Internet access services have been widely available in the many areas in the United States and many other countries. Voice over IP (VoIP) services have been in trials while the technology and PacketCable specifications are being developed and implemented. The first service scenario we consider includes VoIP and basic Internet access services.

In addition to voice and basic Internet access services, the potential of new services for the cable operators is enormous. Cable operators can provide different grades of service to create tiers of services. Different tiers of Internet service could be associated with different bandwidth guarantees, maximum delays, and maximum packet loss probabilities. Massive bandwidth available on cable will open up the opportunities of many new services, e.g., streaming video, interactive gaming, and VPN connectivity for work at home. Table 1 shows the typical set of traffic parameters associated with some of the new services.

3. Multi-Service Network Design Principles

The key to designing any efficient network is a good model of bandwidth requirements. Multi-service networks provide the additional challenge of estimating bandwidth requirements for several services. These services may have to satisfy multiple performance criteria and they can also have different busy hours – overlapping or non-overlapping.

Let us take the simple example of designing a residential voice and data network over the cable infrastructure. An important performance parameter for voice service is probability of blocking while for data service a key criterion is throughput. First, let us

Services	Bandwidth Requirement	On-Line Ratio	Active Ratio	Busy Hour Erlang
G.711 Voice with PHS - 2 way	121.6 kbps	N/A	N/A	0.346
Basic Data Service		60%	40%	N/A
Downstream	400 kbps			N/A
Upstream	160 kbps			N/A
Video Telephony - 2 way	400 kbps	30%	100%	N/A
Streaming Video				N/A
Downstream	1.5 mbps	25%	100%	N/A
Upstream	0			N/A
Interactive Gaming		25%	80%	N/A
Downstream	500 kbps			N/A
Upstream	100 kbps			N/A
SOHO - 2 way	1.5 mbps	100%	80%	N/A

Table 1: Example Application Bandwidth Requirements

assume that there is a common busy hour for both the services – 8:00-9:00 p.m. The network is loaded to its limits for both signaling as well as bearer traffic during this hour. Proper network engineering will require ensuring the desired blocking probability for voice traffic and providing the required throughput for data traffic. Secondly, let us now assume that the busy hour for voice is 7:30-8:30 p.m. and the same for data is 8:30-9:30 p.m. In this scenario, the design needs to meet the blocking criterion during the voice busy hour and the throughput criterion during the data busy hour. Lastly, let us assume that voice busy hour is from 7:30-8:30 p.m. and data busy hour from 8:00-9:00 p.m. In this case, the design needs to meet blocking criteria 7:30-8:00 p.m., blocking and throughput criteria 8:00-8:30 p.m. and throughput criteria 8:30-9:00 p.m.

The above is probably the simplest example of a multi-service network – only two services with one performance criterion each – and yet it poses a formidable challenge already to the network designer. In the very near future, cable operators will provide a combination of voice, data and video services to a mixture of residential, telecommuting, business, SOHO and other types of customers over a common platform. Each of the services may also be offered in different tiers, such as silver, gold and platinum classes. It is obvious that the solution space is multi-dimensional and the possible number of dimensions could grow out of control if it is not properly constrained by certain network planning and design framework.

In order to properly bound the problem, we first define as the building blocks several basic services, and, for every basic service, we assign certain traffic parameters based on a collection of expert opinions and market data. Note that these traffic parameters are part of the modeling tool inputs and can be varied to fit different service assumptions. Several basic services then can be grouped into a service package. Growth or service penetration rates then are associated with the service packages. Finally, we compute the weighted averages of the parameter values across all services for the network. This approach can be used in parts of a network (e.g., the region served by one Fiber Node) or to the entire network (e.g., a region served by one or more Head Ends). Figure 1 shows sample upstream and downstream bandwidth requirement at a Head End with respect to varying market sizes and service penetration rates.

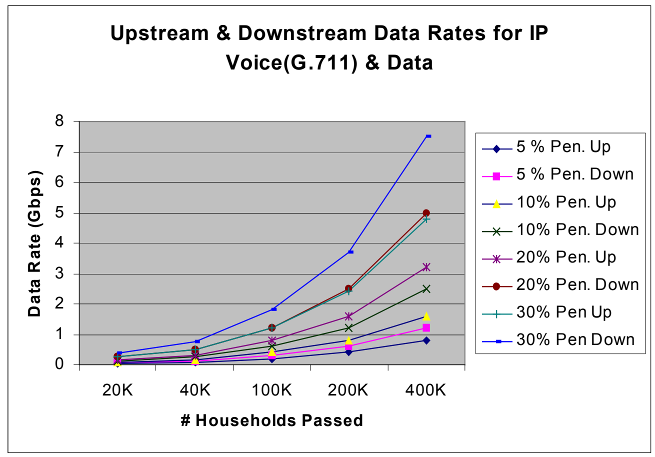

Figure 1: Voice and Data Bandwidth Demand

The basic concept of busy hour is still valid, but it has to be interpreted in the proper perspective. We no longer look into busy-hour traffic data for individual services; what we need, is to establish the busy hours for the aggregate bandwidth demand. Today, we have extensive data on voice bandwidth demand and have some measurements of Internet data traffic demand. As other types of data usage increases and as multimedia traffic grows, there is an increased need for keeping track of subscriber bandwidth requirements.

Once the basic traffic requirements are established, the next step is to set the proper performance objectives. These performance criteria will typically be different for various types and classes of service. The challenge is to engineer a network that provides the required bandwidth while satisfying the required performance criteria.

4. Performance Considerations

In order to capitalize on the promise of converged networks, it is necessary to develop models and network design techniques that allow the network designer to engineer a single infrastructure capable of supporting multiple services. These models and techniques have to be flexible enough to give services a degree of isolation from one another in terms of network characteristics to meet service-specific requirements. The designer takes into account all of these requirements in the form of delay, delay jitter, bandwidth requirements, BER, packet loss rates etc. By evaluating the performance impact of various engineering the parameters of the MAC and IP layers as well as choice of QoS schemes and proper network sizing attempts to satisfy the requirements of all services. The first step is to understand the acceptable levels of performance required by each service and identifying the major factors impacting performance. Until very recently, the Internet access over cable was the predominant, non-traditional service offered by the MSOs. The service requirements were very loosely defined. For both delay and packet loss, the general understanding was that less is better, but no specific targets have been established. This very loose definition of service levels has led to under-engineered networks and unacceptable performance levels as the subscriber population grew. Planned introduction of voice services changes that picture entirely. Voice services, and primary line services in particular, are characterized by very strict delay, delay jitter and packet loss requirements. Legacy services, such as fax and voice band data, place even stricter requirements on those performance measures. Below we discuss the major factors affecting voice and data performance in networks with HFC access. Figure 2 shows the major architectural components of such a network.

4.1. Delay

As mentioned above, voice services place very stringent delay requirements. These follow from the desire to maintain PSTN levels of voice quality in the VoIP services in the cable environment. The general requirement is that the bearer channel delay should be less then 300 ms round trip. It is widely supported by perception studies and standards. Primary line service is also characterized by strict signaling delay requirements, driven by two factors. The first is driven by customer perception of what constitutes good service and second, maybe even more important, is the desire to maintain all of the optional features (call waiting, caller ID etc.), some of which have strict timing requirements. In the following we identify the end-to-end delay components, with particular emphasis on bearer traffic delays.

4.1.1. Delay in Multimedia Terminal Adapter/Cable Modem (MTA/CM)

MTA performs A/D conversion and packetization of the speech samples. The delay introduced by both is driven by the choice of codec. All codec's have an associated block size i.e. the number of speech samples that need to be accumulated before the codec can process them. The block size ranges from the low of 1 sample (i.e. a delay of 125 microsecond for 8000 sample/s) to a high of 240 samples (or 30 ms) for G.723 codec. Some codec's also require a look-ahead time, which can be as high as 7.5 ms for G.723. The packetization time has to be an integer multiple of block size and therefore its theoretical lower limit is a single block size for the codec of choice. However, one has to consider the issue of bandwidth efficiency and the longer the packet in terms of bytes, the smaller is the overhead associated with packet headers and physical overheads. In some sense we have to strike a balance between the delay and bandwidth efficiency.

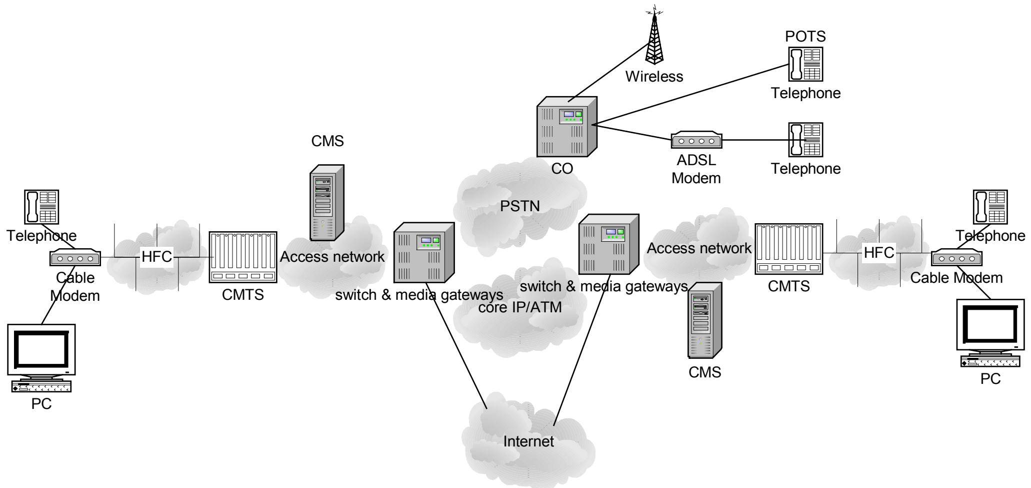

Figure 2: A converged network with ingress cable access and various possible egress networks

The additional factor to consider is that the more efficient the codec i.e. it generates smaller number of bytes per unit time of speech, the higher (in percentage terms) is the overhead, for a fixed packetization time. We will come back to that subject later. It suffices to say here that the "optimal" packetization time ranges someplace between 5 and 20 ms with 10 ms being commonly used. Additional delays are incurred at the MTA for packet processing (i.e. header generation etc.).

For data traffic the estimate is much simpler. As to data traffic, the CM is just a bridge/router with its well understood delay sources.

4.1.2. Delay in HFC Plant

The HFC plant delay includes not only the propagation delay (which is negligible) but also the Medium Access scheme. The switch from data oriented DOCSIS 1.0 to integrated services DOCSIS 1.1 has become the enabler of real time services. For real time services, Unsolicited Grant Service (UGS) is the preferred MAC access method. UGS service grants periodic, fixed size transmit opportunities to established CBR flows. In order to minimize the access delay an MTA can synchronize its packet generation time with the expected grant arrival time and hence not incur any delay (other then a small safety margin). Should the MTA choose not to implement grant synchronization, the MAC access delay becomes a random variable (uniformly distributed) in the ranging from 0 to 1 packetization time, as the packet might be generated at any time between successive transmit opportunities. That delay is random only from call to call and should remain constant for the duration of a particular call. The picture becomes much more complex, when multi-line MTA is considered, where the processor has to perform a juggling act between encoding and decoding operations on multiple connections and still try to keep these operations synchronized with the transmit opportunities.

4.1.3. Delay in CMTS and IP Interconnecting Network

Once the voice packet reaches the CMTS successfully, it will be treated preferentially, as voice packets carry very high ToS or DS marking. The delay therefore is negligible, as the only traffic a voice packet competes against is that of other voice packets. A simple estimate, based on the relative sizes of the ingress and egress interfaces and the packet size, shows that delay in the worst case is no more than 0.5 millisecond. Same conclusion remains true (and actually improves) throughout the IP interconnecting network, as the speed mismatch between ingress and egress is the worst at the CMTS. One should however emphasize that this delay is highly dependent on the interconnecting network architecture.

4.1.4. Delay in Gateways and Core Networks.

Once the traffic emerges in the primary headend, it can take two distinct paths, depending on the called number location and more importantly, transport network that a service provider has at its disposition. As of today, there are few QoS enabled (a.k.a. managed) IP backbones in existence as the technology to build them is not yet mature enough. Therefore, for the sake of this analysis, one should primarily consider two choices. A direct ATM backbone connection and PSTN hop-off. The delay analysis of the ATM backbones is relatively easy, as they are connection oriented (i.e. allow for traffic engineering), QoS enabled with call admission control. Under such a set of conditions, one can easily predict (and guarantee) the delays, which for a backbone have two major components: switch delays and propagation delays. The picture is much more complex, when a hop-off to PSTN is needed. Hopping off to the PSTN entails restoring original packet spacing, commonly referred to as dejittering, and playing the voice packets out. Playing out involves restoring packets into a TDM structure, an operation that ranges in complexity from simple recovery of individual bytes from a packet for G.711 to a full decoding operation for compression codec's. The dejitter buffer will delay packets by a fixed amount that is programmed based upon the amount of delay variation experienced by packets form the same flow on their way from the MTA to the PSTN gateway. The less jitter, the smaller the dejittering delay. This is where proper network design in terms of sizing and QoS will have a major effect.

The propagation delay through the PSTN (or any network, for that matter) is impacted by the facilities length, a factor related but distinct from the distance between the end points. It takes into account the fact that facilities do not necessarily follow the shortest path, as well as the fact that for reasons ranging from traffic engineering to protection switching the even the shortest facilities path may not be used. The correction factor of 1.3 to 1.8 is commonly used. Once the packets arrive at the egress network attached to a cable plant, all the processes the packet went through on ingress, i.e., packetization at the gateway, dejitter, and play-out at the terminating MTA will be repeated. The only difference is that now we send bits downstream on the cable plant. The transit delay is negligible, with the biggest component being the interleaving delay.

4.2. Delay Jitter

The delay jitters have several components. The first occurs right at the MAC access where the transmission opportunities can be scheduled within a certain window. This picture becomes much more complex if VAD is used, which employs UGS with poling. This access method works similarly to the UGS service, except that when silence is detected the flow of grants is stopped and resumed again when speech is detected. There is a delay associated with restarting the flow of grants and that additional delay will impact the first packet of the talk spurt. Additionally, the CMTS upstream MAC scheduler might not be able to schedule the transmission opportunities for the flow at the same position as previously. This will add to jitter. Furthermore, each router, including the CMTS, can experience varying levels of congestion and hence packets of the same flow waiting for an access to an outgoing link will experience variable delay through each router. Similarly, the egress (CMTS downstream) link buffers and scheduling algorithm at the CMTS should be engineered carefully to avoid potentially excessive jitter.

Since the packets have to be played out evenly spaced, dejitter buffers are used at end points to properly re-space the packets. The amount of time packets is placed in that buffer before they are played out should be slightly larger than the maximum expected jitter, otherwise packet loss will occur. A buffer that is too large leads to an unnecessary additional delay. The upper bounds on jitter can be estimated by adding all jitter elements mentioned above. Jitter in routers can be estimated by calculating the maximum number of packets from competing voice connections. Finally, it should be pointed out that jitter can become a component of constant delay if the first packet (its time of arrival establishes the play-out reference point) incurs maximum jitter.

Generally, the requirements for data are not as strict as they are for voice. It, of course, depends on type of data. Interactive gaming has actually stricter delay requirements than voice. Other types of data, such as e-mail, are practically insensitive to delays. But most of data, as represented by Internet traffic or NCS signaling traffic, do not have to be delivered to the destination with accuracy of several (tens) of milliseconds and virtually no jitter. However, one should take into account that many of these types of traffic are riding on the TCP protocol, which can be profoundly effected by long delays, excessive jitter and packet loss. All of these can throw the TCP behavior out of balance and limit its throughput disproportionately to the incremental impairment.

4.3. Bandwidth Efficiency

It is well known that IP is extremely bandwidth inefficient for short packets, such as voice packets. As discussed previously there exists a trade-off between delay and bandwidth efficiency. The headers associated with bearer traffic are RTP/UDP/IP headers, Ethernet MAC and DOCSIS MAC. Additional overhead is incurred by FEC and mapping of packets into minislots. Finally, there is physical layer overhead associated with guard time and preamble. As some of these overheads are unavoidable, in particular given the harsh conditions of the HFC plant, others are constant form packet to packet and can therefore be suppressed. The mechanism of the PHS can be used to suppress the packet header constant fields in particular the entire UDP, IP headers and nearly entire Ethernet header in the upstream direction can be suppressed. It is accomplished through a mechanism PHS mask (byte wise mask of suppressed fields) and telling the CMTS (or the CM) the values of the constant fields so it can restore them. The header suppression technique is not in general available in the IP backbone network. The overhead becomes more pronounced as the number of bytes in the packet gets smaller, as is the case when one shortens the packetization time and/or deploys a more efficient coder/decoder (codec). PHS is not used for data packets.

4.4. Packet Loss and FEC

In order to prevent excessive packet losses (voice packets are particularly sensitive, as they cannot be retransmitted), one can protect the packets with an error correcting code. The more extra bytes are added, the more errors can be corrected. However, it also results in lower bandwidth efficiency. The issue of FEC depth is of particular importance in the upstream direction where interleaving cannot be used. The FEC depth required is related to the noise characteristics in the HFC plant. In the high ingress levels will contribute to lower bandwidth efficiency, higher bit error rates and packet loss rates, which for voice services will lead to deteriorating voice quality. In general, losses in the range of 0.1% - 1% are service effecting, with the exact number highly dependent on the codec type used. In order to mask the impact of packet losses on voice quality, an error concealment method should be implemented in the voice decoder. For voice-band data (VBD) and fax, the requirements are even more stringent (0.01%

packet loss), and failure to achieve them might lead to unusable service. It should also be pointed out that fax and VBD services usually require G.711 coding, which is less bandwidth efficient, as compression codec's do not reliably transfer the analog signaling. Reliability of DTFM tone transmissions in compression codec's is of concern as well.

5. Lucent Cable Network Engineering Tool (LuCNET)

Lucent's Cable Network Engineering Tool (LuCNET) has been developed to minimize equipment cost while still satisfying necessary performance criteria. The input to the tool comprises traffic demand, network topology and service definitions (service area characteristics) and performance objectives.

The user can specify Head End (HE) and Distribution Hub (DH) location on a map and input traffic parameters on an interactive input screen. The tool generates Customer Premises Equipment (CPE) to DH bandwidth estimate. It then uses various in-built engineering rules and performance objectives to determine various equipment configuration. Once the tool is run, the user can view graphically a schematic representation of the network, sizing of inter-connecting links, and equipment sizing at each node. Numerous sensitivity analyses charts can also be generated. Network costs for various configurations can be determined interactively, when the cost data is available. The cost elements can be fed into a business modeling tool in conjunction with revenue and other relevant data to generate various financial reports. A representative process flow chart for the tool is presented in Figure 3.

6. Example LuCNET Results

LuCNET is highly flexible to generate different types of results that may be of interest to a network designer. The basic output comprises a schematic network topography connecting the Distribution Hubs and Headends in a single or hierarchical dual ring structure. The tree structure connecting the Distribution Hubs to the Fiber Nodes and to the homes is also captured in network diagrams. Links are shown towards the PSTN, SS7 and Data networks. By clicking on one of these links, one can see the type of facility (e.g., DS3, OC3…), the number of facilities (e.g., 1, 2…) and the loads of these links (e.g., 65%, 80%…).

Clicking on the nodes can show details of the nodes on the diagram. They will show what equipment and how many of them are there (e.g., Cable Modem Termination System, Call Management System, etc.) By further clicking on the equipment, one can also see the detailed equipment configuration (e.g., number of chassis, plug-in cards, etc.)

Some of the most interesting outputs from the tool are the various sensitivity analyses charts. Figure 4 shows the sensitivity of cost per subscriber with respect to market size and take rate. The x-axis shows different market sizes (with respect to the number of households passed) and the y-axis represents normalized cost per subscriber. Each curve corresponds to a particular take rate. As expected, cost per subscriber goes down as the market size and penetration increases.

Figure 5 shows the sensitivity of the total network cost (normalized) with respect to different market sizes and penetration. On the x-axis, each bar corresponds to a market size (HHP – Households Passed) and a penetration value. The y-axis represents normalized cost for the entire network. The top part of each bar represents the cost contribution from the CMTS while the bottom part represents the cost contribution from the CMS.

Presented in this paper are only a few examples of sensitivity analysis charts of common interest. The tool can create numerous such figures based on individual needs of the users.

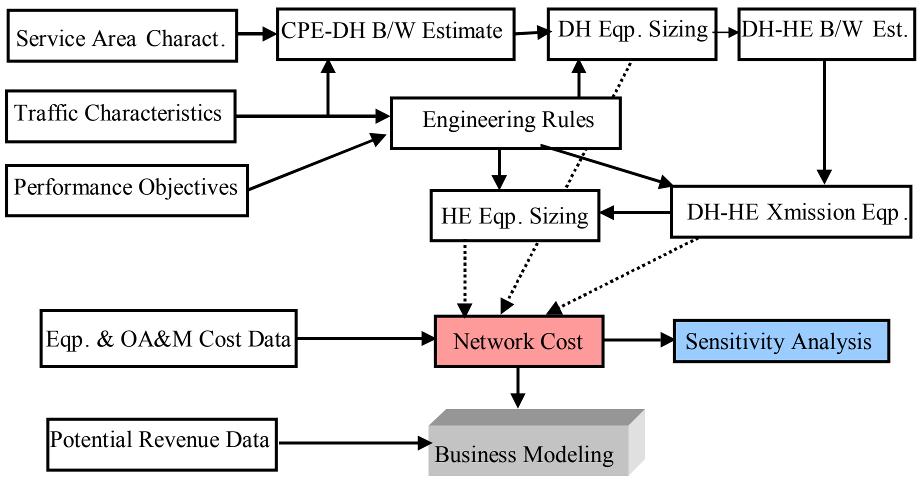

Figure 3: Network Modeling Process Flow

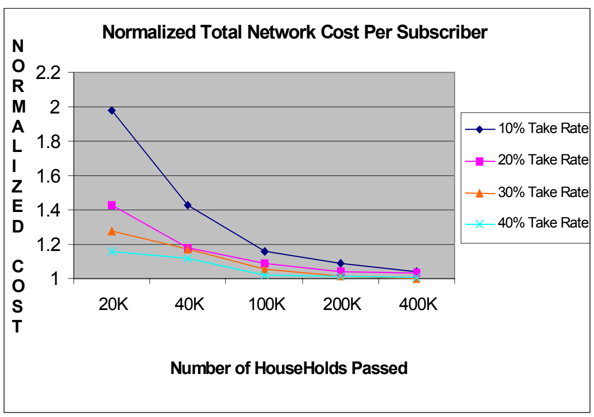

Figure 4: Cost Per Subscriber Sensitivity vs. Network Size and Service Penetration Rate

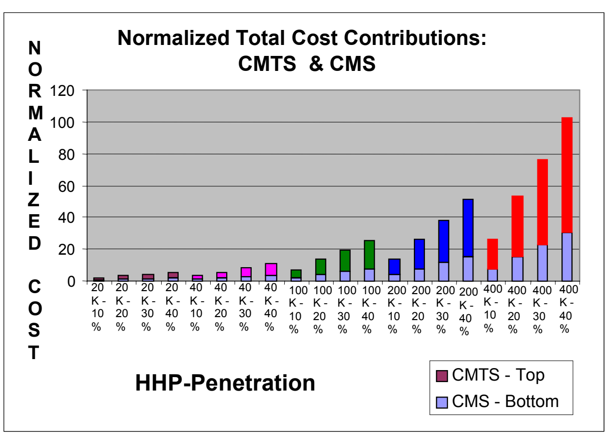

Figure 5: Normalized Cost Contribution by CMS/CMTS

7. Concluding Remarks

Cable access networks present a special challenge to network planning and design. Sharing of the network resources extend from the core to the access, which traditionally is dedicated to a single user. With the enormous amount of bandwidth on cable access, many new services will become available. These new services will further complicate the tasks of designing an efficient access network while meeting QoS objectives of all the services. In this paper, it is shown that, with appropriate design tools and good understanding of service characteristics, a cable access network can be built efficiently and effectively to offer multiple services on an IP platform.

8. References

[1] Cable Television Laboratories, Inc., Data-Over-Cable Service Interface Specifications.

Contact Author Name: Chia-Chang Li
Mailing Address: 4K430, 101 Crawfords Corner Rd., Holmdel, NJ 07733
Telephone Number: (732)949-3101
Facsimile Number: (732)949-7425
Email Address: chiali@lucent.com