CAPACITY PLANNING FOR CABLE HIGH-SPEED DATA SERVICES[†]

K. R. Krishnan, Martin Eiger, Arnie Neidhardt, and Tamra Carpenter
Telcordia Technologies, New Jersey

Abstract

The engineering of cable networks for IP-based voice and data services presents new planning challenges to cable operators. Unlike the broadcast video services for which cable networks have traditionally been designed, the traffic of these new services is directed to individual subscribers. We describe algorithms for estimating capacity requirements to support IP-based services at acceptable levels of QoS as well as the traffic models on which they are based. The algorithms account for the efficiencies realized from the statistical multiplexing of independent traffic streams of different subscribers. We provide examples of their use in investigating various "what-if" scenarios. By combining the capacity estimation algorithms with methods for deriving the parameters of the traffic models from network measurements, one could create a monitoring and planning system for provisioning IP services on cable networks.

INTRODUCTION

The engineering of cable networks for IP-based voice and data services, as defined in CableLabs PacketCable™ and DOCSIS® specifications, presents new planning challenges to cable operators, since the traffic of these new services is directed to individual subscribers, unlike the broadcast video services in traditional cable networks. To be successful in offering these IP services to subscribers, cable operators need a new set of algorithms to determine the capacity requirements for providing acceptable levels of QoS for the services [1,2]. In this paper, we describe capacity-estimation algorithms for IP services and the traffic models on which they are based. The traffic models offer a mathematical description of the traffic of the IP services, and the capacity-estimation algorithms determine resource requirements, at various points in the network, to support traffic loads at specified QoS levels. The algorithms account for the efficiencies that are realized from the statistical multiplexing of independent traffic streams of different subscribers for each service (multiplexing the streams of *heterogeneous* services is fraught with problems, as pointed out later, and is not attempted). We show by means of examples the use of the algorithms for investigating various "what-if" scenarios, including the trade-off between QoS guarantees and network resource requirements, and the projection of network capacity requirements for various scenarios of demand growth.

The services considered in this paper are Voice-over-IP and High-Speed Data. We present mathematical models for the traffic streams of these two services and determine the bandwidth requirements (in the upstream and downstream directions) at various points in the network to meet specified levels of QoS. The mathematical models characterize the random

[†] Copyright 2004 Telcordia Technologies, Inc. All rights reserved.

fluctuations of traffic rates in typical sessions of each service, and enable us to determine the capacity requirements as a function of traffic loads and QoS specifications. This capability lends itself for use in a "what-if" tool to answer various questions, e.g., what QoS levels that can be supported on the existing network, where would capacity augmentation be required, either for the current demand or forecast demand.

We consider the application of our algorithms to two examples. The first example shows the trade-off between capacity requirements and QoS constraints. In the second, we consider the evolution of demands over a 5-year horizon and show the capacity savings achieved by the proposed algorithms, which account for multiplexing efficiency, in comparison with linear extrapolations that fail to account for the multiplexing gain.

In principle, the parameters of the mathematical models of traffic can be estimated from traffic measurements that are collected at a fine enough time-resolution. However, such detailed measurements may not always be practical or economical in all networks. If the model parameters can be estimated from the *routine* operational traffic measurements in a network, then the capacity-estimation algorithms could become part of an integrated monitoring and planning system. Such a system will enable a network operator to plan and install new capacity before existing capacity runs out. The integration of parameter-estimation methods with capacity-estimation algorithms is a subject for future studies.

NETWORK ARCHITECTURE

The architecture of a cable network is designed for the efficient distribution of broadcast television services. Figure 1 shows the typical two-level hierarchical structure of cable networks. For broadcast television, the signal feeds enter the network at the head-end, from which the traffic is transported over a high-speed backbone ring to various distribution hubs (each of which is the site of one or more Cable Modem Termination Systems (CMTSs)). Each hub sends the traffic to each of its subtending fiber nodes, which then distribute the signals to individual homes over coaxial distribution networks.

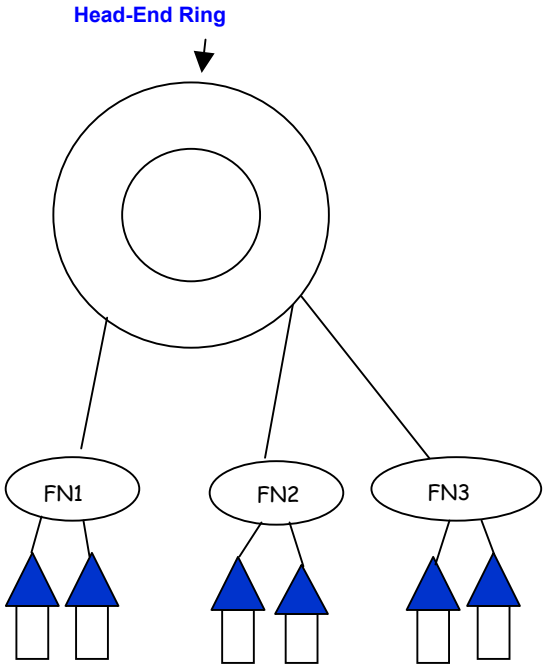

Figure 1: Architecture of Hybrid Fiber Coax (HFC) Cable Network

The introduction of IP-based voice, data, and video services over this architecture raises new issues for consideration:

a) Since the traffic for these services is dedicated to *individual subscribers*, this portion of the system's requirements for bandwidth is determined by the number of subscribers served and their traffic demands.

b) The routing of downstream traffic to the right subscribers is more complicated than in traditional broadcast television networks that offer tiered packages of video channels.

c) The routing of upstream traffic from subscribers on a shared medium is a new problem that is absent in traditional broadcast television networks (although some cable systems carry upstream traffic in support of pay-per-view video services).

IP-based services are accommodated within the distribution network by dedicating some bandwidth spectrum for digital traffic. Typically, the upstream traffic occupies the spectrum from about 5 MHz to 42 MHz [3-4], while one, or possibly two, 6 MHz channels at higher frequencies are set aside for downstream digital traffic. At the distribution hub, analogue and digital traffic are combined in the downstream direction and separated in the upstream direction. Within the backbone network, analogue and digital traffic are carried on separate facilities.

For the digital traffic of the IP-based services, each CMTS is a point of aggregation for the traffic of all the subscribers served by it, while the backbone ring aggregates the traffic of all the CMTSs at the various hubs. Designing a cable network to support IP-based services requires that the traditional method of designing for signal integrity within a certain bandwidth range must be combined with *a method for ensuring that the bandwidth at distribution hubs and on the backbone ring is sufficient to handle traffic loads and deliver a given set of services at desired levels of performance.* We now turn to the methods for estimating these capacity requirements.

CAPACITY DESIGN

Load Estimation

The load of a service at a CMTS at a hub can be determined in two steps:
In the first step, we estimate the average number of *simultaneously active users* for the service in the cluster of subscribers served by the fiber node or nodes associated with the CMTS. Typically, we consider the load during the network *busy hour* for the service. Given the number of homes served by the CMTS, the penetration of the various IP services for that cluster of homes, and the activity level for a typical subscriber for each service, we can determine the average number of simultaneous users.

In the second step, we combine this estimate of the number of simultaneous traffic streams with information about the *shape* or intrinsic characteristics of the typical traffic stream generated by a user of that service, to determine the aggregated traffic of the service passing through the CMTS, in each direction, in the busy hour. Similarly, the aggregation of the traffic streams of all hubs gives us the aggregate traffic of the service on the backbone ring.

Capacity Calculation

For each service, the bandwidth requirements at the hubs and on the ring are determined on the basis of the required levels of QoS and the *aggregated* traffic at the corresponding points in the network, thereby taking advantage of the efficiency of capacity utilization arising from the statistical multiplexing (superposition) of the individual traffic streams. The algorithms for these calculations of bandwidth requirements are presented below for the two services considered in this paper: Voice-over-IP (VoIP) and Data.

The traffic aggregation, however, is considered only for traffic streams of the *same service* and *same levels of QoS*, i.e., we multiplex only *within* each class of traffic, and *not across* traffic classes, because the multiplexing of streams of different characteristics and QoS requirements may offer no benefits, and, in fact, might require additional control mechanisms to ensure that each class of traffic receives its proper QoS. Therefore, we adopt the conservative rule of merely adding the separate bandwidth requirements of each class of traffic to arrive at the requirements of the combined traffic of heterogeneous streams.

IP SERVICES

We now present the traffic models, QoS parameters, and capacity calculations for Voice-over-IP and Data. The model and QoS parameters for these two services, together with some representative default values, are summarized in Table 1.

Voice-over-IP

a) Traffic Model

For the traffic produced by a single Voice-over-IP call, we use an ON-OFF model for the traffic rate, with the rate alternating between a constant peak value during ON-intervals (which correspond to talk spurts), and zero during OFF-intervals (which correspond to intervals of silence).

The peak rate resulting from the standard digitizing of voice-samples at the Nyquist rate is 64 kb/s, and is then subject to modification by the coding rate, and by the overhead involved in forming IP packets from segments of talk spurts. The ON and OFF intervals are treated as random variables of exponential distribution. Thus, the parameters characterizing this model for Voice-over-IP traffic are:

τ_{on} = mean duration of a talk spurt

τ_{off} = mean duration of a silent period

R = coding rate

h = length of header of an IP packet

c = "packetizing policy"
= duration of talk-spurt segment
collected as IP packet-payload

P = peak rate
$= R + \frac{h}{c}$

The first two parameters, τ_{on} and τ_{off}, pertain to the characteristics of voice traffic, for which extensive experimental studies have produced typical default values that can be used in the absence of user input. The coding rate R and the coding policy c (which affect customer-

perceived quality of voice-connections) are parameters of the coder, and thus assumed known. The header length *h* pertains to the IP protocol that is implemented.

b) QoS parameters

For VoIP, the QoS parameters are packet-loss rate and delay-jitter. In addition, since VoIP is a real-time connection-based service, subject to blocking if the network cannot meet the packet-level QoS constraints, the probability of such blocking becomes an additional, *connection-level* QoS constraint.

c) Load Calculations

Consider the calculations for homes (customers) subtending a given CMTS. We want to know the total offered load A in Erlangs due to these customers in the busy hour. Let C be the number of customers, and let θ be the penetration factor for the Voice-over-IP service. Then, $C\theta$ is the number of subscribers to this service. If u is the utilization factor for a typical subscriber, i.e., the fraction of time during the busy-hour that a typical subscriber would spend on voice-calls *if the subscriber suffered no blocking*, then the offered load per subscriber is u Erlangs. Then, the total offered load equals $A = C\theta u$ Erlangs, which can also be viewed as the *average number of simultaneous calls that would be in progress during the busy-hour, if there were no blocking.*

We assume that there is admission control for voice calls (i.e., a new call attempt would be blocked if sufficient bandwidth cannot be provided to it), and that the probability of a call attempt being blocked should not exceed a specified value b. To meet the blocking criterion for the offered load A, the *minimum* number of simultaneous calls that must be supported at the CMTS is the smallest integer N for which $B(N, A) \le b$, where $B(N, A)$ is the Erlang-B blocking function [5]. This is the number of calls for which the VoIP traffic model above would be used to determine the required bandwidth at a CMTS, for specified QoS constraints on packet loss, delay, and jitter.

d) Bandwidth Calculations

As an example of bandwidth calculations using the VoIP model, we present below the formula [6] for the bandwidth L required for supporting N simultaneous calls at a loss rate of r, given a buffer of size B.

Define

$$\phi = \frac{P\tau_{on}\tau_{off}}{B(\tau_{on} + \tau_{off})} \ln\left(\frac{1}{r}\right);$$

$$f = \frac{\tau_{on}}{\tau_{on} + \tau_{off}}$$

$$m = Pf$$

Then, $L = Min[L_1, L_2]$, where

$$L_1 = NP\left[\frac{\phi - 1 + \sqrt{(\phi - 1)^2 + 4\phi f}}{2\phi}\right]$$

$$L_2 = Nm + \sqrt{[-2\ln r - \ln(2\pi)]Nm(P - m)}$$

Corresponding expressions can be derived for the bandwidth required to meet jitter constraints (often, the jitter constraint is treated as a bound on the maximum delay, experienced when the buffer allocated to the service is full). The maximum of the bandwidths determined by the loss and jitter constraints is then the bandwidth required to meet *all* the QoS constraints.

Data

a) Traffic Model

In pioneering studies at Telcordia Technologies, high-speed data traffic was shown to be characterized by burstiness over many time scales, a phenomenon known as "long-range dependence" [7-9]. A fluid model known as Fractional Brownian Motion (FBM) [6] was shown to be capable of representing the aggregated traffic of a large number of independent streams, such as those generated by users downloading files from the World Wide Web. The FBM model is a Gaussian model with stationary increments, and is specified by the parameters (m, a, H), where

m = mean traffic arrival rate

a = peakedness parameter

H = Hurst parameter, with $0.5 \le H \le 1$

The peakedness parameter a describes the variance of the fluctuations in traffic rate at the time scale used in the model description, and H characterizes the persistence of correlation in traffic rates with time lag, i.e., is a measure of long-range dependence, with $H = 0.5$ corresponding to short-range dependence and $H > 0.5$ corresponding to long-range dependence. It can be shown that the multiplexing of n independent FBM sources, each described by (m, a, H), gives rise to the FBM process (nm, a, H).

b) QoS parameters

For data sessions, we assume that there is no admission control, and hence no blocking constraint to be considered. The QoS parameters are, therefore, packet loss rate and mean delay.

c) Load Calculations

The demand will be specified in terms of the requirements of the aggregate data stream that has to be supported. Then, just as in the case of voice calls, we arrive at $A = C\theta u$ as the average number of simultaneous data sessions in progress, where the penetration and utilization factors now pertain to data service. If the traffic of a single session is described by the FBM process (m, a, H), the aggregate traffic of A simultaneous and independent sessions is given by (Am, a, H).

d) Bandwidth Calculations

As an example of bandwidth calculations with the FBM model, we present below the formula that determines the mean delay d when FBM traffic (m, a, H) is offered to a link of bandwidth L [6]:

$$d = \frac{K^{-\frac{1}{\theta}}}{L}\Gamma\left(1 + \frac{1}{\theta}\right),$$

where

$$\theta = 2(1 - H),\ K = \frac{(L - m)^{2H}}{2am(1 - H)^{2(1-H)}H^{2H}},$$

and $\Gamma(z)$ is the Gamma function

We can invert the formula to determine the bandwidth L required to achieve a *given* mean delay d by doing a binary search for L.

EXAMPLES

a) QoS and Capacity Requirements

This example illustrates how the capacity-estimation algorithms can be used as a “what-if” tool to determine the effect

of the QoS levels specified for the Voice-over-IP and Data services on the capacity requirements.

Network

We consider a symmetric network of 4 hubs, each with one CMTS. At each CMTS, the upstream traffic has a channel of bandwidth 2.2 Mb/s, with a buffer of 300 kbits, and the downstream traffic has a channel of bandwidth 27 Mb/s with a buffer of 3 Mbits. The bidirectional ring has a bandwidth of 24 Gb/s, with a buffer of 3 Mbits.

Load

At each CMTS:

VoIP = 30 erlangs

Upstream data rate = 0.5 Mb/s

Downstream data rate = 0.5 Mb/s

The parameters for the traffic models for VoIP and data are taken to be those given in Table 1.

QoS

We first consider the following choice of QoS parameters (QoS-1):

VoIP: Connection blocking = 0.1%
Maximum delay = 10 msec
Bit-loss rate = 5%

Data: Average delay = 50 msec
Bit-loss rate = 1%

For the loads assumed, the bandwidth required at each CMTS to support VoIP is 1.4 Mb/s, while the bandwidth required for the upstream data (which turns out to be the bottleneck here) is 0.856 Mb/s. Thus, the total bandwidth needed for the upstream channel is 2.256 Mb/s, which *exceeds* the available upstream channel bandwidth of 2.2 Mb/s. Thus the network *cannot* support the services at the performance levels in QoS-1.

We next consider the following set of less stringent QoS parameters (QoS-2):

VoIP: Connection blocking = 1.0%
Maximum delay = 10 msec
Bit-loss rate = 10%

Data: Average delay = 100 msec
Bit-loss rate = 5%

The bandwidth for VoIP is now 1.2 Mb/s, while that for the upstream data is 0.777 Mb/s, for a total bandwidth requirement on the upstream channel of 1.977 Mb/s, which is smaller than the given channel bandwidth of 2.2 Mb/s. Thus, the existing network can support the two services at the performance levels in QoS-2.

b) Multi-Year Capacity Planning

This scenario deals with capacity planning over a 5-year horizon, under a given forecast of load evolution, and shows the benefit of taking account of the statistical multiplexing that occurs in aggregating the traffic streams of different subscribers for the same service. The network is the same as in the previous example, and the forecast 5-year load evolution is given below in Table 2, along with the results of calculation for the total

upstream bandwidth at each CMTS, using the performance levels specified in QoS-1 above.

Once again, the bottleneck is the upstream channel bandwidth, which remains adequate to support the loads for the chosen QoS parameters in Years 1-4, but becomes inadequate in Year 5, according to the results of the capacity-estimation algorithms appearing in Row 3 of Table 2.

Suppose, on the other hand, that one merely looked at the loads and bandwidth requirements in Year 1, and estimated the bandwidth requirements in the future years by linear extrapolation of the bandwidth requirements in Year 1. The results of such extrapolation (Row 5 of Table 2) would lead one to the false conclusion that the network runs out of capacity even for Year 3. The comparison of bandwidth requirements determined by the capacity-estimation algorithms with those calculated by linear extrapolation shows that the penalty in failing to exploit multiplexing gain increases with increasing load. So, *this scenario demonstrates the potential benefit of the algorithms in deferred capital investments, by the explicit accounting for multiplexing gain in calculating capacity requirements.*

CONCLUSIONS AND FUTURE WORK

The engineering of cable networks for IP-based voice and data services presents new planning challenges to cable operators. The traffic of these new services is directed to individual subscribers, unlike the broadcast video services for which cable networks have traditionally been designed. To be successful in offering these IP services to subscribers, the network must have sufficient capacity to provide acceptable levels of QoS. There is a need for a new set of algorithms for planning and provisioning new IP services on cable networks.

In this paper, we have described traffic models and algorithms for estimating capacity requirements to support IP services at specified levels of QoS. We model the digital portion of the cable network in terms of resources and their capacities, along with mathematical models that capture the characteristics of the traffic of the IP-based services that these resources must accommodate. The capacity-estimation algorithms determine resource requirements at various points in the network to support traffic loads at specified QoS levels. Where sufficient resources are present, they are partitioned among the different types of traffic. Locations with insufficient resources are identified and the shortfall is determined. The capacity-estimation algorithms can be used for investigating a wide variety of "what-if" scenarios, including the trade-off between QoS guarantees and network resource requirements, as shown in the examples that we consider.

The capacity estimation algorithms account for the efficiencies that are realized from the statistical multiplexing of traffic streams of different subscribers for the same service. By considering a scenario of multi-year evolution of subscriber demands, we demonstrate the capacity savings achieved by the multiplexing efficiency built into our algorithms, in comparison with the approach of linear extrapolation of capacity requirement in the number of subscribers, which could lead to gross overestimation of capacity requirements. More accurate estimates will lead to better strategic decisions on when, where, and how to offer new services.

For mathematical models of traffic to be useful, one must be able to determine proper values for their parameters, to obtain a reasonable fit to the traffic being described. In principle, the model parameters can be estimated from traffic measurements collected at a fine enough time-resolution. However, such detailed measurements may not always be practical or economical in all networks. If the model parameters can be estimated from the *routine* operational traffic measurements in a network, then one could derive the inputs to the capacity-estimation algorithms from traffic measurements and load projections, creating an *integrated* monitoring and planning system. Such a system will enable a network operator to plan and install new capacity before existing capacity runs out. The integration of parameter-estimation methods with capacity-estimation algorithms is a subject for future studies.

It would also be desirable to expand the system to propose remediation when it finds that demand exceeds network capacity. We envision optimization algorithms that will propose ways that cable operators might add network capacity at minimal cost, determining appropriate QoS parameters for viable Service Level Agreements, and implementing controls to meet them.

We also have to investigate whether other services such as Streaming Video, Video-on-Demand, and Video Games can be represented in terms of existing traffic models or will require the construction of new models and corresponding capacity-estimation algorithms.

REFERENCES

[1] P. Bates, T. Carpenter, Y. Chandramouli, J. DesMarais, M. Eiger, K. R. Krishnan, and A. L. Neidhardt, "Planning and Provisioning for Cable Internet Services", National Fiber Optic Engineers Conference, September 2002, Dallas, Texas.

[2] J. T. Chapman, "Multimedia Traffic Engineering for HFC Networks," White Paper, CISCO Systems, 1999.

[3] Cable Datacom News, Overview of Cable Modem Technology and Services,http://www.cabledatacomnews.com/cmic/cmic1.html.

[4] W. Ciciora, J. Farmer, and D. Large, Modern Cable Television Technology, Morgan Kaufman Publishers, Inc., San Francisco CA, 1999.

[5] R. B. Cooper, Introduction to Queuing Theory, 2nd ed., New York: North Holland, 1981.

[6] K. R. Krishnan, A. L. Neidhardt, and Y. Chandramouli, "Traffic Models for Cable Network Services", Telcordia Technologies, 2001.

[7] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long Range Dependence in Variable Bit-Rate Video Traffic," IEEE/ACM Transactions on Networking, 43, pp. 1566-1579, 1995.

[8] M. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-similar Video Traffic," Proc. ACM SIGCOMM, London, September 1994, pp. 269-280.

[9] W. Willinger, M. Taqqu, W. Leland, and D. V. Wilson, "Self-similarity in High-speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements," Statistical Science, 10, pp. 67-85, 1995.

Table 1: Traffic Model and QoS Parameters
(with reasonable default values)

	Voice-over-IP	Data
QoS	• Jitter tolerance (0.01 sec) • Loss tolerance (0.5%) • Blocking tolerance (0.1%)	• Loss tolerance (1%) • Average delay tolerance (0.05 sec)
Traffic Shape[1]	• Talk spurt (1.004 sec) • Silence (1.587 sec) • Packet header (416 bits) • Analogue-to-digital coding rate (6.4 kb/s – 64 kb/s) • Voice frame size (0.010 – 0.030 sec)	• Peakedness[2] (61000 bits/sec^{2H-1}) • Hurst parameter (0.85)
Intensity[1]	• Erlangs	• Bits/second

Table 2: Five-Year Evolution of Loads and Capacity Requirements

	Y1	Y2	Y3	Y4	Y5
VoIP Load (erlangs)	15	20	25	30	35
Upstream Data Rate (Mb/s)	0.25	0.30	0.35	0.40	0.45
Upstream Bandwidth Requirement (Mb/s)	1.415	1.664	1.948	2.113	2.385 **(X)**
Linear extrapolation from Y1 (Mb/s)	1.415	1.821	2.228 **(X)**	2.635 **(X)**	3.041 **(X)**
Overestimate in extrapolation	**0%**	**9.5%**	**14.3%**	**24.7%**	**27.5%**

Entries marked with an "**X**" exceed the upstream channel bandwidth of 2.2 Mb/s

[1] Traffic shape and intensity are given independently for upstream and downstream traffic. The parameters should assume the same values in the two directions for symmetric services, such as voice.
[2] Peakedness is measured in bits/sec^{2H-1}, where H is the value of the Hurst parameter.