Traffic Analysis and Simulation of a Prioritized Shared Medium *

Jeffrey J. Evans
Department of Electrical and
Computer Engineering Technology
Purdue University
401 N. Grant St.
West Lafayette, IN 47907
email: jjevans@tech.purdue.edu

Cynthia S. Hood, Phillip M. Dickens
Department of Computer Science
Illinois Institute of Technology
10 West 31^{st} St.
Chicago, IL 60616
email: {hood, dickens}@iit.edu

Abstract

This work studies the delay effects resulting from heterogeneous traffic types competing for communications resources on a prioritized shared medium using a fixed access and admission control scheme. This research focuses on a specific class of Hybrid Fiber Coaxial (HFC) residential Internet access system. Such systems have three separate traffic classes: voice, modem, and high-speed data. We use analysis, modeling and simulation, to understand how the combination of these three traffic classes, all competing for finite and prioritized bandwidth resources, impacts the "perceived" quality of service (QoS) by the end user. We develop analytic models for the various traffic types being considered, and validate the individual models using simulation. We then conduct simulation studies addressing the impact of the combination of these three data sources all vying for the limited bandwidth resources under varying source and offered load conditions. We show that traditional modeling techniques are not able to effectively capture the behavior of such real-world systems, and demonstrate that our techniques are quite useful for gaining insight into the behavior and performance of these systems.

1 INTRODUCTION

The vast majority of residential and small business users (on the order of 95%) obtain Internet access by way of an analog modem communicating over their telephone line [16]. Recently however new alternatives for telephone and Internet access have emerged including Digital Subscriber Line (DSL), Satellite, and Hybrid Fiber Coaxial cable (HFC). The objective of these alternatives is to provide a broader range of services to the consumer together over a single transport medium. Ideally, a service provider would like to offer video, data, and reliable (lifeline) telephone service over the same medium. However there is currently no single system that can provide all three services economically and reliably. For example, today's DSL systems can deliver reliable telephone service and high speed, always on, Internet access. Today's DSL however is not useful for delivering a wide range of video services. Conversely, satellite systems can now provide excellent quality video and bi-directional high speed Internet access, however reliable telephone service remains a challenge.

We focus our attention on the class of Network access systems (NAS) that use Hybrid Fiber Coax (HFC) to provide point to multi-point (shared medium) access from the public switched telephone network (PSTN) to the end user (consumer). Recently, the capabilities of these systems have been augmented to include direct access to the packet switched network (the Internet) in an effort to take advantage of otherwise unused bandwidth. This is done in such a way as to provide a prioritized set

*To appear in SIMULATION: Transactions of the Society for Modeling and Simulation International

of services over the shared medium where voice (and modem) traffic is given *strict priority* over data traffic. The problem is that the design and deployment of these Network access systems is based upon conservative telephone traffic engineering principles, and such principles do not necessarily apply to the new services being offered. For example, the service (call-hold) time for modems is substantially longer than that of the voice traffic for which these systems have been designed. The analysis and network design problem is further complicated by the addition of high-speed data users to this traffic mix of voice and modem users. Thus new tools are needed for the traffic analysis and modeling of these emerging Network access systems, where such tools take into account the statistical variability of the different traffic types (i.e. voice, modem, and high-speed data) simultaneously.

In this paper, we begin the task of understanding the individual statistical characteristics of voice, modem, and data traffic, and the traffic engineering ramifications of these source types superimposed on a single prioritized shared medium. We argue that using traditional techniques valid for voice traffic has resulted in a failure to sufficiently consider the ramifications of data traffic, be it via modem or direct high-speed (available bandwidth) access. Also, it is quite likely that the HFC systems under consideration will over time experience a gradual transition in the mix of voice, modem, and high-speed data users. In particular, we believe that modem users will eventually convert their Internet access to the high-speed data service as their requirements for bandwidth intensive applications increases. Thus our problem consists of trying to gain insight and intuition of the ramifications, if any, of subjecting a prioritized, shared medium to traffic types whose statistical and time scale characteristics vary greatly, and where the mix of such traffic types will change over time. Because the medium is not fair, with a bias toward voice and modem users, we wish to determine:

1. The degree to which modem users disrupt the traditional voice traffic model, affecting QoS for voice and modem users, and,

2. The degree to which high-speed data users can expect to suffer substantial network delays under typical deployment scenarios with generally accepted voice/modem traffic loads.

We begin by analytically modeling the various traffic source types that request bandwidth resources from the prioritized network access system. We then determine by way of related work and published reports the *"Busy-hour, Busy-day"* (BHBD) characteristics of each traffic type in order to establish any gross phase relationships between the arrival patterns of the different traffic types. Next, we develop simulation models to represent each of the traffic sources as well as the prioritized, shared medium using a tool such as OPNET Modeler. Such simulations are needed to verify the statistical properties of individual source types, as well as the properties of multiple instances of the same source type. We can then develop simulations that combine the different source types in such a way as to "mimic" actual BHBD traffic on typical deployments.

The modeling and simulation of these different classes of traffic sources (and the bandwidth allocation mechanism) will provide valuable insight into system-level behavior under realistic operating conditions. This intuition can be used in several ways. For example, it can be used as the basis for further work in designing reactive and proactive mechanisms for delivering the highest Quality of Service (QoS) to all users. Also, these results can be used to better plan the migration of modem users to high-speed data users. Ultimately, understanding the arrival, service time, and variation characteristics for these different sources of communications traffic will provide valuable assistance in better designing NAS topologies.

The contributions to the simulation community include:

- Demonstrating the importance of all the statistical properties of each of the individual traffic types in the performance analysis of shared medium access systems.

- Creating and verifying analytical and simulation models for each of the important elements (traffic source type and bandwidth server) of an important class of real-world shared medium access systems.

- Combining the individual simulation models to produce models of system-level behavior. These simulations encompass many scenarios with different numbers (and ratios) of users and offered load conditions.
- Providing valuable insight into the relationship between the behavior of the mixed traffic sources (superimposed on a shared medium access system) and the Quality of Service available to each class of network traffic.

The rest of this paper is organized as follows. In Section 2, we provide background and begin developing our model of a class of real-world Network Access Systems used to transport voice, modem, and high-speed data traffic simultaneously. We analyze, then simplify characteristics of the physical system that do not substantively contribute to the general behavior of the system, allowing us to focus attention on those aspects of the system's behavior that most affects the user's QoS. In Section 3, we discuss previous work in the area of modeling communications traffic at the source level. In Section 4, we determine the Busy-Hour, Busy-Day behavior of each traffic source type. By showing the "cyclical" nature of the users (voice, modem, and high-speed data) access to the medium, it is established that we can simulate all traffic types attempting to access the shared medium during the same time period. Also, we develop the analytical models necessary for investigating the bandwidth utilization behavior of the individual traffic source types on the shared medium, as well as the bandwidth server entity used to control medium access. However, we acknowledge that there is no means of combining these models into a single analytical expression. Thus in Section 5 we create simulation models of the entities, and then use simulation studies to verify their individual performance behavior against their corresponding analytical models. We then run simulations combining the individual traffic source models in such a way as to simulate Busy-Hour, Busy-Day mixed traffic in realistic access network deployment scenarios. We provide our conclusions and discuss areas of future work in Section 6.

2 BACKGROUND

In the coaxial cable area, two classes of systems are being developed and deployed. Both classes are capable of delivering excellent quality broadcast video, due to the enormous bandwidth capacity of the cable itself. In both classes, there is a "one-to-many" relationship between the "Local Exchange" (otherwise known as the central switching office or headend) and the end user.

2.1 Packet-Switched (PS) Cable Systems

A more recent trend has been the introduction of standardized data delivery using *"cable modems"* and their associated central office components (termination systems, routers, etc.). We refer to this class of system as *"packet-switched"* (PS) cable systems, which resemble other data-centric, shared medium access systems. These systems use aggressive transport mechanisms (modulation techniques over radio frequencies) resulting in more efficient use of the transport spectrum. This comes at the price however of requiring a well-controlled (reduced noise) physical layer. Since these systems are designed to transport packet (Internet) data traffic (not telephone traffic), the bandwidth allocation tends to be asymmetric. The downstream direction (toward the end user) tends to be large (on the order of 30Mbps) while the upstream direction (toward the Internet) is much smaller (on the order of 2Mbps) [8]. This is primarily due to frequency spectrum allocation, or more specifically, *where* the gaps in TV channels are and *how large* these gaps are. This bandwidth allocation is then shared among many (several hundred) end users.

To communicate, each end user requests bandwidth from their *server* at the headend, and is then assigned a set of highly granular "timeslots" or *"Bandwidth Units"* (BU's) to transmit their payload. To receive, each end user listens to all packets (which use a different sized BU) in the downstream, processing those packets destined for that end user. This packet-switched model is adequate for today's Internet traffic, and is well suited for the delivery and control of future video services. However, significant challenges remain in providing reliable, scalable, and affordable telephone service.

One reason for this is the continued evolution of a consistent, scalable network architecture, and its resulting standards and specifications. Another is the challenge of distributing traditional PSTN functionality (such as call admittance and control, security, billing, etc.) across data networks on a large scale, while maintaining the reliability and Quality of Service (QoS) presently seen in the PSTN. The interested reader is directed to [5] for a thorough discussion of these issues.

2.2 Circuit-Switched (CS) Cable Systems

A different class of coaxial cable system, which is the focus of this research, is designed specifically to provide reliable telephone access (in addition to video transport). We will call this class the *"circuit-switched"* (CS) HFC class. We do not imply or suggest that the CS class is superior or inferior to the PS class. It is just different. A typical CS system is generally designed using less aggressive, but more robust transport mechanisms (modulation schemes) and optimized (although often proprietary) transport, access, and control protocols. While resulting in less efficient use of the transport spectrum when compared to cable modems, it is symmetric in nature, which is necessary to transport bi-directional voice traffic. Figure 1 provides a simple illustration of such a system.

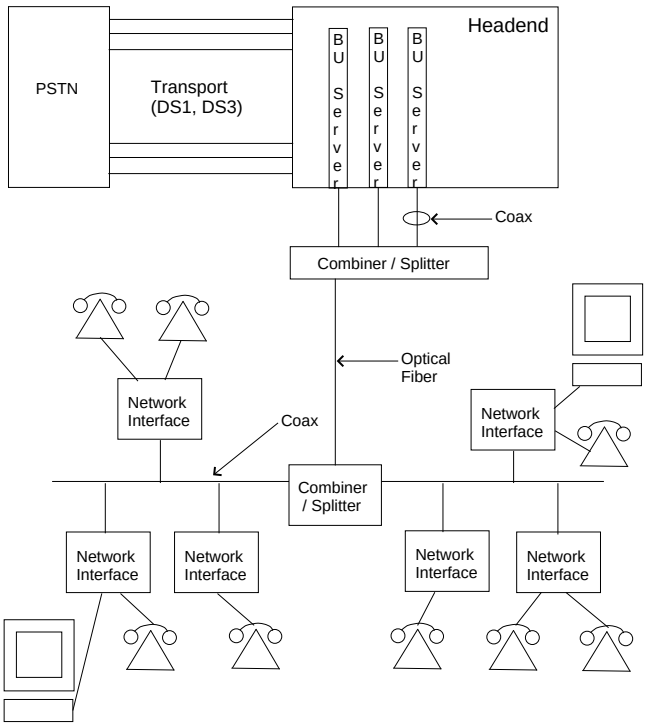

Figure 1: HFC Telephone Network Access System

In this system, which is connected directly to the PSTN, many end users (Network Interfaces) compete for a relatively small number of fixed sized, symmetrical BU's controlled by a headend device (BU Server). This is different from the PS class where the BU's differ in both size and symmetry. The fixed size BU normally takes the form of a 64Kbps timeslot, identical to a DS0 in telephony jargon. When one of these bi-directional BU's is allocated to the end user, information flowing toward the network is re-mapped and aggregated at the headend onto other forms of *"Time Division Multiplexed"* (TDM) transport. These transport mechanisms (such as telephony T1, E1, OC-X) are used for transport to the PSTN, by way of a Local (telephone) Exchange (or Local Digital Switch). Note in Figure 1 that some of the endpoints are represented with both a telephone and a

computer host. The host gains access to the packet switched network (WWW) by way of an analog modem, but must go through (or at least into) the PSTN first.

2.3 Emergence of Data Services

More recently, high-speed direct Internet access capability has been added to the CS class of HFC access systems. The motivation for this additional service is the fact that at almost any time there is a significant amount of unused bandwidth. This stems primarily from the fact that these systems were designed using very conservative telephone traffic principals, leading to an over-provisioning of network resources. Thus while there may be less total bandwidth available in these systems in comparison to cable modems, the set of end users is approximately an order of magnitude smaller [8].

Figure 2 illustrates the CS class of HFC access systems. The system has a similar topology to Figure 1, with the exception that direct access to the packet switched network is available when BU's are available. A data user can access the shared medium's resources (BU's) at a reasonably granular level in time and is not limited to using only one BU at a time. The admission control policy however is that access to the medium is always granted to a telephone user first. Thus when a telephone user requests service (i.e. goes off-hook or is called), the resources required for the call are dynamically re-allocated to that telephone user reducing (or eliminating) the resources available to the data user (as well as other telephone users). The effects are twofold.

1. First, extended burst times result when bandwidth resources are reduced (but not eliminated) due to higher priority voice (or modem) traffic.

2. Secondly, added delay due to queueing results in the case where bandwidth resources are completely consumed by the higher priority telephone traffic. Moreover, once resources are consumed by telephone calls, they tend to be consumed for minutes at a time.

Such resources may be consumed even longer if the user is gaining access to the Internet via a modem rather than via the higher speed data service. As an example, consider that America Online reports that the average user spends 55 minutes per day on the Internet [7]. Since we are studying a system which is currently deployed world-wide our focus is to study potential delay effects of heterogeneous traffic types competing for resources. We therefore do not consider alternative admission control policies in this work.

2.4 Shared Medium Access Model

Consider a system such as that illustrated in Figure 2. The shared medium between any BU Server and its set of Network Interfaces consists of an approximately symmetric, TDM transport. Specifically, the downstream (toward the end user) transport is loosely based on a frame comprised of an integer number of usable payload timeslots, with each timeslot consisting of eight bits, again based on voice traffic transport. The transmission scheme is such that the aggregate bit rate is on the order of 2Mbps. A subset of timeslots is utilized for non-payload carrying functions such as synchronization (framing), signaling (on-hook/off-hook), and a lower speed communications channel. This channel is used by the single point "Bandwidth Unit (BU) Server" to communicate to the customer premise equipment (multi-point "Network Interfaces"). This communication consists of provisioning, admission request and control, and diagnostic requests.

The upstream transport is similar in capacity and functionality. For robustness reasons however, its timing is modified such that many payload "bytes" are buffered in each Network Interface, then each Network Interface is given a "burst opportunity" to transmit the contents of its payload buffer. In the voice (and analog modem) case the upstream payload is then "de-multiplexed" and placed on an appropriate DS0 in real-time for transport to the PSTN. Medium access (for admission control) for voice traffic bandwidth is realized over a low-speed data channel. This channel is again part of a subset of timeslots for non-payload carrying functions. Communication is realized using a common

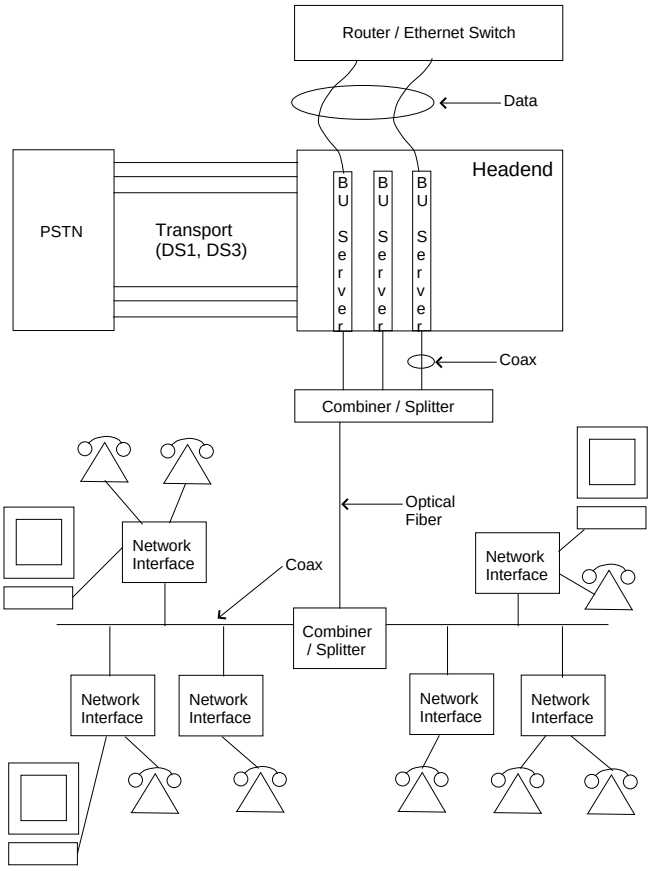

Figure 2: HFC Telephone / Data Network Access System

medium access technique, where each Network Interface may attempt to communicate (request bandwidth) at a single point in time (timeslot). The voice traffic arrival rate on this system is of a time-scale several orders of magnitude greater than that of the granularity of the medium access mechanism. Therefore, it appears reasonable to disregard consideration of the details of the medium access protocol itself, allowing us to abstract away this portion of the Network Interface entities of Figure 2.

A different medium access system is employed by which access to bandwidth is possible for data traffic. The method employs control bits in the downstream manipulated by the BU Server. These bits inform the listening Network Interfaces as to when they may place data traffic onto the transport as well as how much data they may place on the transport. Additionally, messages decoded in hardware are used to dynamically "re-map" the available payload timeslots as traffic (of any type) comes and goes. This dynamic mapping of available payload bandwidth is tightly coupled with the medium's priority mechanism. The volume of data access may be as high as the full capacity of the transport bandwidth (on the order of 2Mbps). The upstream timing granularity of access is on the order of a millisecond. In the downstream, a single BU Server controls data access; therefore, the granularity of access is much finer, on the order of microseconds. Lastly, the medium acts asymmetrically for data access. Thus the BU Server can transmit data in the downstream direction to Network Interface A while Network Interface B (or a different channel on Network Interface A) transmits data in the upstream direction toward the network.

Control of the shared medium's resources (i.e. voice, modem, and high-speed data traffic) is

handled solely by the BU Server. We therefore believe it is reasonable to abstract away the Network Interface entity in our model development, leaving the BU Server, the shared medium, and a set of highly variable traffic sources. The resulting shared medium access model is illustrated in Figure 3. Note that there is a subtle, yet very important distinction illustrated in Figure 3. Some of the computers are 'connected' to a telephone rather than being directly to the shared medium. This distinguishes computer users accessing the Internet by way of a conventional modem from those users taking advantage of the high-speed data access service (which are connected directly to the shared medium).

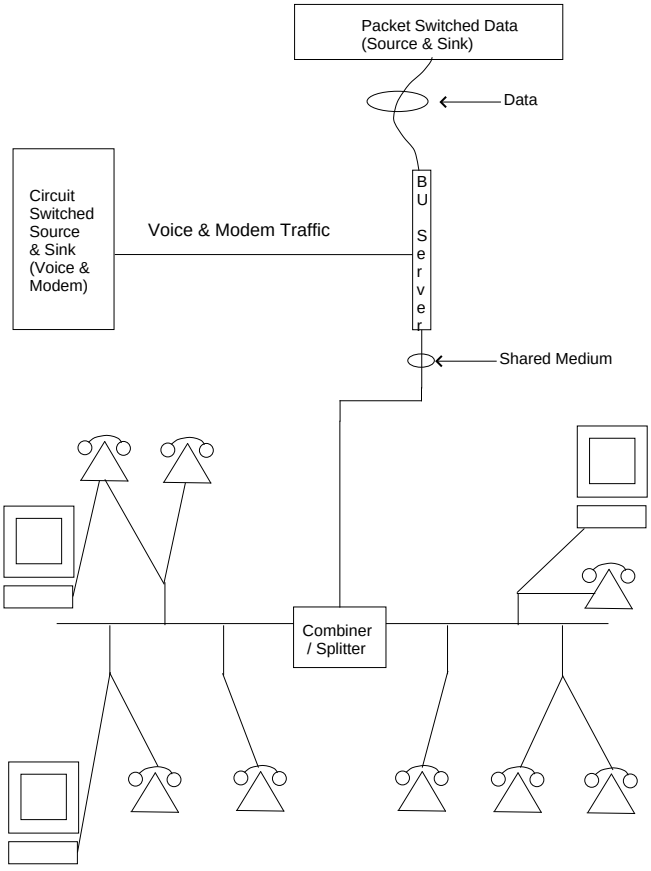

Figure 3: Shared Medium Access Model

3 RELATED WORK

The notion of analyzing and modeling different traffic types presenting loads to a prioritized shared medium is not new. As far back as the late-1980's work has been done in this area, beginning with the Integrated Services Digital Network (ISDN). ISDN has seen limited use in the United States but has gained wider acceptance in other parts of the world. ISDN is a TDM paradigm, where a wide range of services can be supported using multiple TDM channels. These channels can be used individually or combined to form higher bandwidth "connections" to support data centric services such as video conferencing along with traditional voice. As will be seen, work in the ISDN area serves as a foundation for modeling the behavior of the BU Server.

The work presented in [18, 21, 20], is applicable to our data traffic model, tying together the

notions of packet trains [11] for source-destination pairs and self-similarity [14] for aggregates. As discussed below, this work does provide significant insight into the high-speed data traffic generator employed in our study.

3.1 Voice and Modem Traffic

An abundance of studies in the literature have theorized, measured, and analyzed the nature of telephone voice traffic. It is well understood that voice traffic arrivals to the PSTN follow cyclical patterns. These patterns are different depending on the time scale (hours, days, weeks, years), and can be modeled successfully using homogeneous Poisson processes. Designers of telephone networks typically design in order to address what is known as the "Busy-Hour, Busy-Day" (BHBD) parameter, with the goal of handling BHBD traffic with a low probability of blocking (typically less than one percent).

Recent reductions in telephone service cost have resulted in changes in arrival rate and service time characteristics. Studies such as [3] explore the modeling of call hold time distributions to better understand the ramifications to call admission control (signaling) traffic in the presence of short (less than 10 seconds) calls. In this study, two normal distributions are combined to show a bi-modal call hold time characteristic. It also identifies that an exponential curve fit for call hold time deviates considerably from an empirical distribution, even when ignoring a large portion (17%) of the calls over 10 minutes in length. In [4], measured data is presented in the form of average daily telephone use. Also presented is busy-hour use as a percentage of the daily use, as well as growth rates for average daily use over the past several years. This data is used in our study to establish baselines for the voice traffic arrival and call hold time behavior presented in section 4.

Studies such as [6] concur with [3], and offer further insight into the effects of modem traffic on the overall call hold time distribution. The authors state that the use of the exponential distribution for call holding time (service time) "seriously underestimates the actual numbers of very long calls (e.g. analog modem 'data' calls that last for many hours)". Then the *"heavy-tailed"* nature of call holding time distributions using lognormal, Weibull, and Pareto distributions are explored. Our interest in modem traffic behavior lies in the more recent trends suggesting that residential users typically access the Internet for 55 minutes per day on average [7], where the vast majority (95%) gain access using modems [16]. To date, no research (of which we are aware) specifically explores the arrival and service time characteristics of this very important modem *"session"* traffic. This may be attributable to the difficulty of determining (by some measurement) if a call is a voice or modem call. Monitoring admission control traffic provides call hold time (call arrival and release) information, but cannot distinguish the *"class"* of a call. We assert however, that the trends cited in [3] and [6] are continuing and have accelerated since the time of those studies [7, 16]. Thus rather than attempting to develop a single model encompassing voice and modem traffic, we separate the two into independent traffic classes and examine their individual class behavior. Each class (voice and modem) will have its own arrival and service time characteristics, and can be more easily monitored separately for simulation and data collection purposes.

3.2 Data Traffic

As recently as the early 1990's researchers have been misled by the statistical nature of data traffic. For example, the advent of ISDN in the 1980's saw research in the area of modeling both narrow band and wide band services being accessed on the same medium. The models had their foundations in multiserver queueing theory [17, 13, 19], where the source models, both narrow (voice) and wide (data) band, took the form of Poisson processes with exponentially distributed service times. While in some cases these techniques are useful, it has been recognized that in general these notions of modeling data access have become outdated [12, 15]. It is also now generally recognized that modeling data traffic using Poisson or compound Poisson processes is inappropriate (or at least outdated) [11, 12, 15, 18].

In [11], the authors demonstrated a clear deviation from Poisson packet arrivals from their analysis of measured LAN data at M.I.T. Their log plot of the histogram of interarrival times

revealed neither a Poisson (straight line) nor compound Poisson (a straight-line with a spike near the origin) process. The author's observation of "source locality" is also important. In their creation of the "packet train" model (also known as the ON/OFF model) they observed that measured traffic exhibited a phenomenon such that "successive packets tend to belong to the same train". In other words, there was a high probability that a packet going from point A to point B would be followed by either another packet from point A to point B or a packet from point B to point A.

In [21], Ethernet LAN traffic at the source level is statistically analyzed. The mathematical results indicate that the superposition of many ON/OFF sources with strictly alternating ON-periods or OFF-periods exhibit what is termed the *"Noah Effect"* (high variability or infinite variance). This in turn produces aggregate network traffic that exhibits what is termed the *"Joseph Effect"* (self-similarity or long-range dependence). Theorems are developed and traffic measurements performed to describe and verify the statistical behavior for large M (i.i.d. traffic sources) and T (time).

An approach is then provided through which self-similar traffic can be generated. For our purposes however there are two problems with the proposed approach. First, a massively parallel computing environment is required to produce synthetic traces of self-similar traffic in reasonable time. Secondly, the assumptions of a large number of sources M and time T do not apply to our particular modeling problem. We thus consider the simpler techniques outlined in [6, 8], that can generate near self-similar traffic behavior based on Pareto and Weibull distributions producing a *"heavy-tailed"* packet arrival effect. We do however take from [21] the idea of accounting for the Noah Effect with individual traffic sources by using infinite variance distributions.

Other studies consider different application types that are common in the Internet environment. For example, in [8] it is shown that HTTP sessions tend to produce a "burstier" traffic pattern than a "streaming" session such as "RealAudio" or "RealVideo". The addition of substantially more home businesses and telecommuters make it increasingly difficult to speculate who will use these streaming applications, as opposed to browsing, and at what time of the day they may be used. For our purposes however, we restrict the scope of the application domain to more bursty applications such as browsing and light file transfer.

In [20], textured plots are used to offer a visualization of the packet arrival process at the source/destination level. Limit results for aggregate WAN traffic are presented where it is assumed that "sessions" (i.e. FTP, HTTP, Telnet) arrive according to a Poisson process, then transmit packets deterministically at a constant rate, then cease transmitting packets. The main stochastic element left undefined is the session length, which is presented as possessing a long-range dependence or heavy-tailed property. We are interested in these WAN traffic results because our system exhibits virtually no intra-LAN communication (i.e. data users sharing the same medium do not communicate directly with each other), and we consider and use these concepts in the development of the high-speed data traffic generator model discussed in section 5.

3.3 The Shared Medium (BU Server)

The authors of [13, 19] explore matrix-analytic queueing analysis and state transition tables respectively to model access control strategies in an ISDN. In both cases, the medium is modeled as a multiserver $M/M/N$ queue, with infinite capacity servicing two classes of traffic, called NB (Narrow-Band) and WB (Wide-Band), again modeled as Poisson arrival processes. The medium's channel capacity consists of basic bandwidth units (BBUs), where each BBU is synonymous with a server. In both cases, various policies for the handling of blocking and queueing were considered. In [13], both non-preemptive and pre-emptive session-level admission policies were analyzed and simulated to compare the performance of the various policies. In [19], case equations and performance curve illustrations are given. In both cases, the emphasis was limited to session level blocking probabilities and queueing delays, and no appreciable investigation into packet level queueing delays was performed. In this case, the resultant simulation data would be expected to provide an indication of delay due to congestion, which could provide for the development of a QoS metric. In our case, we will need to examine different (generic) arrival and service time characteristics and to consider a finite number of available resources. These considerations are made in section 5.

The BU Server work described in this section provides a starting point from which we can expand. The notion of two independent Poisson arrival processes using different arrival rates and service times can be used to address voice and modem traffic. However, our medium requires the addition of another traffic class, that being high-speed data traffic, with packet-level granularity for admission control and preemptive controls for priority handling of voice or modem arrivals. This complicates the problem, and the approach to handling this greater complexity is provided in Section 4. However, since our system has fixed bandwidth allocation and admission control policies, we need not be concerned about considering multiple policies or optimizing the control scheme in terms of blocking probability or loss, which is the objective of studies such as [2]. Rather, we are concerned with the *"effects"* of various traffic types and loads on the medium and QoS *"ramifications"* to the end users.

3.4 Simulation of HFC Systems

In [8], modeling and simulation of performance on a shared 10Mbps HFC medium is addressed. The medium is not prioritized however, and the analysis addresses only data traffic (i.e. cable data modem on a PS HFC access system). The technique is developed to generate data traffic is an ON/OFF approach, using a Weibull distribution for ON periods and a Pareto distribution for the much longer OFF periods. These "heavy-tailed" distributions exhibit a significant number of large values, such that *"the probability density function decays with increasing values at a rate slower than that of an exponential distribution, such as Gaussian or Poisson."* The resultant model and simulation make the simplifying (yet reasonable) assumption of a "homogeneous, non-dispersive Web", in an effort to isolate the access portion of the data network from the Internet itself. Simulation results indicate that significant added latency (greater than one second) does not occur until 600 to 1000 users are simultaneously active, depending on the Internet data rate, which is assumed to be constant.

Similar to the work in [8], we also generate data traffic using an ON/OFF approach. However, the notion of utilizing unused bandwidth on a circuit-switched HFC access system for packet-switched high-speed data is new. Since current CS HFC access systems are designed based on conservative voice-only traffic models, our work breaks new ground by investigating the combined effects when modem and high-speed data users are added to the traffic mix.

4 Traffic and Medium Characteristics

In order to run meaningful simulations we must identify and model the statistical behavior characteristics of the traffic sources and the shared medium (BU Server). The traffic source classes are voice, modem, and high-speed data, and each traffic source class behaves according to its own stochastic process. Furthermore, in order to decide how long (in simulation time) the simulations should run, it is necessary to identify on a larger (session level) time scale whether a traffic source class is attempting access to the medium or not. In other words, when voice users are accessing the medium (during BHBD) are modem users also trying to access the medium? What time-of-day do high-speed data users try to access the medium? Does it coincide with voice and/or modem use?

For the shared medium (BU Server), we must identify and define just what the BU Server is. Is it single or multiple servers? Does it have a single queue or series of queues? How deep are the queues? We must also address and handle gracefully those scenarios identified earlier in section 2. These scenarios include admission policy (priority management), preemption policy, and arrival collision detection and management. In the most detailed case, we seek to understand the behavior of the BU Server when a voice or modem arrival occurs and a data traffic burst (which is allowed to consume all remaining resources on the medium) also is occurring.

4.1 Voice Traffic

The probability distribution most commonly used for modeling voice traffic is the Poisson distribution. The combination of Poisson call arrivals, exponentially distributed service times, and the

notion of a predictable "busy hour" has given rise to traffic intensity measurements of absolute traffic volume in terms of Erlangs and Century-Call-Seconds (CCS). These traffic benchmarks are commonly used around the world to quantify voice traffic, and in most cases traffic engineers can rely solely on traffic tables to effectively design a network topology to achieve a desired QoS (typically 0.01 calls blocked per busy hour of traffic) [1].

If we consider each traffic source a stochastic Poisson process with arrival rate λ, the probability of n arrivals at time t can be expressed as;

$$Pr[n,t] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \tag{1}$$

If two Poisson processes N_1 and N_2 with arrival rates λ_1 and λ_2 are impinging on a single medium, the resulting process is a Poisson process with arrival rate $\lambda = \lambda_1 + \lambda_2$. Expanding this result to n sources results in a Poisson process whose arrival rate is the sum of the individual arrival rates;

$$\lambda = \sum_{i=1}^{n} \lambda_i. \tag{2}$$

We now determine the traffic intensity load offered to the medium (for an individual voice source) to be used is our analysis. From [4], we establish the following call traffic benchmarks from 1997:

- Average Telephone line is in use 57 minutes per day.
- 12% of the per day usage is during the busy hour = 6.84 min. = 410.4 sec. during the busy hour.
- 410.4 sec. = 4.1 CCS

Using an exponentially distributed service time statistic with a mean value u of 4.0 minutes (240 seconds) we can calculate the per source arrival rate λ_i as;

λ_i = offered load per line / offered load per call
offered load per line = 4.1 CCS
offered load per call = 240 sec. per call / 100 sec./CCS = 2.4 CCS per call
λ_i = 4.1 / 2.4 = 1.7 call per hour (BHBD)

We increase the call rate slightly to 2.0 call per hour in order to have round numbers to work with. This increases the per line intensity from 4.1 CCS to 4.8 CCS (480 sec. during busy hour). Stated another way, the average line daily usage increases from 57 minutes per day as cited above (1997 data) to 67 minutes per day. In [4], traffic intensity growth rate increases from 1994 to 1997 that are on the order of a few percent per year. Extrapolating conservatively from 1997 data gives us 68 minutes per day, which is within two percent of our 67 minute per day figure derived above. Lastly, our mean service time value of four minutes per call results in eight minutes per hour, per line, or 11.9 percent of the total minutes per day of 67. This matches well with the 12 percent of daily usage occurring during the busy hour as cited above. Therefore, when a simulation is run using 120 voice sources, the arrival rate will be 240 (120 * 2) calls per hour, with calls having a mean service (or call hold) time of 4 minutes, with an exponential distribution.

4.2 Modem Traffic

The creation and subsequent rapid growth of the Internet has resulted in modifications to how one thinks of telephone traffic. Home use of computers has risen to the point where users will use their modems to get 'on-line' and stay on-line. For modem traffic, we must first answer the question of whether the busy hour for voice and modem users overlap. That is, when voice traffic is in the busy hour are modem users also present in a similar busy hour fashion? The research presented in [15]

(Fig. 1, page 228) helps to provide an answer to this question. In particular, the authors show significant variations in the daily connection arrival rate for various data connection types (TCP applications such as FTP, Telnet, SMTP, etc.). While the purpose of this data was to demonstrate that a simple homogeneous Poisson process cannot be used to model the arrival process, one can also take from this data another significant point. That is, the time of day and period of time that the data arrivals are highest overlaps that of BHBD telephone traffic. Therefore, for the purposes of this study, we will use these results and assume that data, and therefore modem arrival processes, coincide with each other over the course of a BHBD simulation interval.

Though packet level characteristics cannot be modeled using Poisson processes there must be an initial macro session that can be thought of as being similar to logging on and logging off a LAN, or getting on-line then off-line. These are the modem arrival and service time characteristics. Therefore, we assume that a Poisson arrival process model is acceptable for these 'macro session' modem arrivals. Two significant differences in modem traffic behavior are present however. One is the service time, which is addressed later. The other is the probability sample space or population. For voice traffic, it is reasonable to assume that the population is infinite, due to the combination of originating and terminating calls. We know that for our HFC access system the number of originating call sources (voice plus modem) is finite, to the point of being considered small (less than 150). However, when we consider the notion that any outside source can call any of the finite sources, we conclude that the number of sources is large enough to be considered infinite. This is not true however for analog modem traffic. We assume that all modem calls are originating calls, and, since the number of modem users relative to the number of voice users is also small (less than 50), the population should not be considered infinite.

For this case we can turn to birth-death theory [10] to develop the birth and death rates of the system:

Births (arrivals)

$$\lambda_n = \begin{cases} (M-n)\lambda & : \quad 0 \leq n \leq M \\ 0 & : \quad n \geq M \end{cases} \tag{3}$$

and deaths (service times)

$$\mu_n = \begin{cases} n\mu & : \quad 0 \leq n < c \\ c\mu & : \quad n \geq c. \end{cases} \tag{4}$$

In this state-dependent process, M is the number of modem sources and c is the number of servers (resources) in the system. The state then is defined as the number of modem sources in a call. The state probabilities then reduce to:

$$p_n = \begin{cases} \binom{M}{n} \left(\frac{\lambda}{\mu}\right)^n p_0 & : \quad 1 \leq n < c \\ \binom{M}{n} \frac{n!}{c^{n-c}c!} \left(\frac{\lambda}{\mu}\right)^n p_0 & : \quad c \leq n \leq M, \end{cases} \tag{5}$$

which in our case further reduces to:

$$p_n = \binom{M}{n} \left(\frac{\lambda}{\mu}\right)^n p_0, \tag{6}$$

by virtue of the fact that we are dealing with both a finite number of sources and servers and, arrivals occurring when the system is full are "lost" (i.e. not queued). Since this form of p_n does not allow for a closed form calculation of p_0, we must calculate each of the coefficients multiplying p_0 in equation 6 (which we call $\{a_n, n = 0, 1, 2, 3, \ldots, M\}$ and $a_0 = 1$) and complete the computation as:

$$p_0 = \frac{1}{1 + a_1 + a_2 + \ldots + a_{M-1} + a_M}. \tag{7}$$

We can now use the definition of expected value to develop the average number of modem sources in a call (L) when the system is in a steady state:

$$L = \sum_{n=1}^{M} np_n = p_0 \sum_{n=1}^{M} na_n. \tag{8}$$

We also need to consider the service time distribution and mean value for analog modem traffic. It would be advantageous to run simulations using either an exponential service time characteristic or one that could introduce a "heavier" tail as suggested in [6], which can be achieved by using a Weibull or Pareto distribution for the service time. By adjusting the "shape" parameter α value to be $0 < \alpha < 1$, we can produce service time distributions ranging from exponential to one whose service time distribution possesses a "heavy" tail, with a greater number of service times being longer than the mean.

We will adjust the source mean service time μ_i, and the source arrival rate λ_i, to compensate for the potentially long service times. For example, from [7] we use a mean service time μ_i of 55 minutes per source. With this service time mean and considering its exponential distribution characteristic, we must then reduce the per source arrival rate λ_i, to a more reasonable value, such as between 0.5 and 1.0 call per hour.

4.3 Data Traffic

As stated earlier, voice arrivals follow cyclical daily, weekly, and even yearly patterns. In [9, 15] the authors concur on one crucial point: that there is an arrival "pattern" for data packet intensity, or number of data packets per unit time, that is strikingly similar to that of the voice arrival daily cycle. In [9], it was observed that daily peak-hour utilization was on the order of 30%. From these studies, we conclude that from the macro, or longer time (minutes or hours) perspective, the traffic intensity for voice, modem and high-speed data traffic will all overlap. This means that when simulating several hours of system behavior we need not concern ourselves with "phasing-in or phasing out" entire classes of traffic.

In our system we will be examining the behavior of a small number (less than 50) of ON/OFF sources. These sources could be thought of as a number of individual computer users accessing the Internet, or an even smaller number of users who are each engaged in several data "channels", or applications, each being considered as a source. We are interested in producing data traffic sources whose resultant offered load has high variability, both individually (Noah Effect) and when aggregated (Joseph Effect). We are less concerned with the exact statistical nature of each source (i.e. the individual application(s) that each user might be running at any instant). This behavior can be realized by mimicking packet arrival processes at the source level as seen in [20]. We can do this by creating sessions of packets to mimic application sessions such as FTP, HTTP, and Telnet while ignoring the details of the underlying transport, network, and link layer protocols. These sessions would arrive according to a Poisson process, they then transmit packets deterministically at a constant rate, then cease transmitting packets. Each of these sessions would have a probabilistic session length, or duration. We ensure that the session length would have the statistical characteristic of a long-range dependence, or heavy-tailed property.

4.4 Shared Medium Access (BU Server)

There are two primary objectives in modeling the shared medium (BU Server). The first is to avoid modeling some of the low-level access methods described earlier in section 2, where we abstracted away medium access details for voice and modem traffic. This is a reasonable abstraction given that the time-scale between the arrival rate of voice and modem traffic and medium access are different by several orders of magnitude. Conversely, the BU Server was made solely responsible for medium access of high-speed data and overall resource allocation, including the prioritization, scheduling and queueing mechanisms. The second objective in the modeling of the BU Server is to design and implement the model with sufficient flexibility to provide ease of use and visualization of system behavior.

The BU Server simulation model was designed and implemented in phases. However, to accommodate our overall objectives the BU Server model takes the form of a hybrid queueing system. By hybrid we mean that the queueing model is different depending on traffic type. We are less interested in the subtle aspects of telephone voice (and modem) traffic. For example, consider source behavior when a call is blocked. When a voice user is blocked, do they try again soon, do they wait, or do they give up? Rather than complicate the voice and modem traffic generator models with 'per-source' finite state machines to account for these effects, we simply drop blocked call attempts for voice and modem traffic. In telephony terms this is known as "blocked calls lost". This behavior is that of a $M/G/N/N$ queueing system, where N is the number of servers and the capacity of the system.

We are most interested in understanding the behavior of high-speed data traffic when inter-mixed with voice and modem traffic. As such, a queueing model to express the behavior of the high-speed data traffic is that of a $G/G/N/\infty$ system, where N still represents the number of servers (BUs) but we allow for infinite system capacity in order to queue high-speed data traffic. Blocked data bursts are queued in FIFO fashion, as would generally be the case with Ethernet data traffic, relying on the upper software layers to handle excessive delay scenarios.

The combination of these queueing system models present challenges in terms of resource (BU) management such as prioritization and scheduling. Take a prioritization case for example, of when a voice or modem call arrives and no BUs are available but there happens to be a data burst in progress. Recall from section 2 that the BU Server operates with a time scale granularity significantly different (more granular) for high-speed data traffic medium access when compared to voice (and modem) traffic. When a voice or modem call arrives with a data burst in progress, the data burst BU's must be reduced by exactly one BU. The remaining burst length must be calculated, and in terms of the simulation model, the data burst's original termination event is canceled and a new termination event is scheduled based on the remaining burst time.

A scheduling challenge arises if for example, a data burst is in progress using a single BU, and that burst is "interrupted" because of a voice or modem arrival. In this case, the burst's remaining data is placed at the head of the queue, contrary to other blocked data bursts, which are placed at the tail of the queue. In the overall BU Server model we develop and keep statistics involving blocking and delay characteristics as well as data indicating both the desired and actual bandwidth unit requirements for voice and modem traffic.

5 Simulation Models, Simulations and Results

Our simulation software package (OPNET Modeler) uses a hierarchical approach for developing a *network* to be simulated. At the highest level there are *"networks"*, consisting of one or more *"sub-networks"*, comprised of one or more *"nodes"*, which are composed of one or more *"processes"*. For our study, we employ a simple network consisting of a single node. This node is made up of several processes. The processes consist of a single *"bandwidth server"*, and various *"bandwidth request generators"*. Each generator requests bandwidth from the server according to its generating function (Poisson, ON/OFF, etc.).

5.1 Models

The BU Server (medium) was initially modeled as a $M/G/\infty/\infty$ queue. This was done in order to verify the voice and modem traffic source model statistical characteristics. Behaviorally, any source asking for any amount of bandwidth (or bandwidth units) will get it. By allowing infinite system capacity we can easily visualize and deduce whether the voice and modem traffic sources are behaving as expected according to the analysis in section 4. Later, the simulation model was modified to represent the high priority part of the Server, which handles voice and modem traffic prioritization, acting like a $M/G/N/N$, where N represents both the number of BU's on the medium and the system capacity. This is done by introducing resource (BU) limiting and prioritization behavior and additional run time parameters such as the "Number of Bandwidth Units" (N) (adjustable at simulation time), and statistics such as "blocked arrivals", indicating arrivals (by

type, voice/modem/data) that were rejected by the medium. At this point the BU Server model was used to verify the behavior of the high-speed data traffic source model.

Lastly, the low priority section of the Server, where an infinite capacity queueing mechanism for blocked data traffic arrivals was developed, was modeled as behaving like a $G/G/N/\infty$ queueing system. Control mechanisms were developed and implemented to properly handle additional prioritization and scheduling cases discussed in section 4. This final model offers the flexibility of easily being able to modify simulation parameters at run time. Additional statistics were added involving blocking and delay while maintaining data associated with resource (BU) requests and allocation. Some of the statistics kept in the BU Server process model included bandwidth usage (both requested and actual), service (hold or burst) time, and parameters associated with blocked and queued traffic. We take advantage of OPNET's subqueue statistics to acquire data specific to data traffic, such as the size of the queue and the queueing delay.

The resultant OPNET process model for the bandwidth server consists of six states. An initialization state sets up simulation variables and statistics. The idle state waits for events to occur, providing the correct state transitions and a returning point once the event has been handled. Two states are designed for traffic arrivals, one for voice or modem traffic and the other for data traffic. The remaining two states handle traffic terminations, performing statistics updates and data burst queueing checks when a termination event occurs. Arrivals are handled when a source produces an "arrival" event by sending a service setup packet over a logical message-passing stream to the server process. This happens in zero simulation time, and the server process is "interrupted" by this message. It then sets up (or blocks) the service, updates statistics and schedules a self-interrupt to "terminate" the service based on a "length" (hold time) parameter received in the service setup packet.

The voice traffic process model consists of an initialization state and a generator state. Run time adjustable simulation parameters include arrival distribution (Poisson), arrival rate, hold time distribution (exponential), mean hold time, and number of sources. The number of sources parameter is used to calculate the aggregate arrival rate as a sum of arrival processes (i.e. $\lambda = \lambda_1 + \lambda_2 \ldots \lambda_n$ where n = number of sources). Individual statistics for the voice source model consist of the number of arrivals generated, call length, offered load, and total offered load.

The modem source process model ends up being a near copy of the voice source process model. This model also consists of initialization and generator states. Run time simulation parameters include arrival distribution (Poisson), arrival rate, hold time distribution (exponential or Weibull), mean hold time, number of sources, and a "birth-death" indicator. A birth-death indicator is used with the number of sources and number of bandwidth units parameters to calculate the average number of sources in the system as developed in section 4 (eq. 8). The resulting simulation model behavior is one that effectively "slows down" the arrival process such that when the simulation reaches steady state the number of modems in the system equals the average number calculated. The modem call generator has the same individual statistics as that of the voice call generator.

The data traffic source model, called a variable bit rate (VBR) traffic generator, consists of a parent process that creates "child" processes. Each child process is created according to a probability distribution determined at compile time. The 'session' length of each child process, packet arrival, and packet length pdf's are all individually configurable, also at compile time. Thus the resulting arrival processes (children) can be viewed as either individual users, multiple applications running on one or more machines, or even sub-processes of an individual "session". The VBR parent process initializes itself, then begins to generate child processes according to simulation parameters determined by the user at compile time. The behavioral parameters are used by each child process created by the parent process. Each child process then acts independently as a function of the various probability distributions.

The resulting bandwidth allocation node model is shown in Figure 4. As can be seen, it is comprised of two Poisson process call generators (one for voice traffic and another for modem traffic), the VBR data burst generator, and the bandwidth server process model, and is required as a fundamental building block in order to perform OPNET simulations.

5.2 Model Verification Simulations

The desired end-result is to model and simulate a system containing multiple sources of different traffic types trying to access finite bandwidth. As such, we established "deployment rules" to model the number of sources and their loads presented to the medium (BU Server) in such a way so the simulations would mimic real-world deployment and traffic load scenarios. For example, using the bandwidth capacity described in section 2 (30 BU's) the medium can adequately support roughly 135 telephone lines offering voice traffic (i.e. no modem or high-speed data traffic). In this case, each line offers a load of 2 calls per hour with a mean call hold time of 4 minutes, with a probability of blocking of 0.5%. In terms of average BU resource consumption, this equates to roughly half (15 of 30 BU's) as discussed in section 4. We can then add-in factors such as a percentage of the phone lines being used for modem service, and still another proportion of users attempting higher speed Internet access via the data service. We can then let the numbers (sources and loads) grow to the point where obvious congestion and blocking is taking place.

We considered three *"deployment scenarios"*, based on concentrating (over-subscribing) the sources on a 30 BU medium by factors of two, three, and four, concentration factors commonly found in actual system deployments. In terms of telephone lines (also used for modem data access), this represents 60, 90, and 120 lines (sources) respectively. For simulations of voice and modem source types, we have chosen a worst case scenario of 67 percent of the sources being voice and the remaining 33 percent modem. For the added cases of high-speed data access, the mix of traffic source types becomes a bit more complicated, resulting in the following deployment rules:

- The total number of data users (analog modem + high-speed data access) is limited to 33 percent of the total number of sources simulated. This is actually lower than the estimated number of home computer users (in the United States) who are "on-line" (roughly 38%).
- As high speed data users are added to simulation scenarios, an equal number of analog modem users are removed from the scenario. However each analog modem source will now become a voice source, with a worst case high speed data scenario being 120 voice sources plus 40 (33 percent of 120) high speed data sources (160 total traffic sources).
- Simulation run time was three hours (simulation time) in order to allow the traffic generator processes to stabilize. This figure was arrived at empirically based on trial simulations running in excess of 10 hours simulation time.

Voice traffic only simulations were performed using the deployment rules described above. Ten simulations were run, adjusting the seed value for OPNET's random number generator prior to each run. In the 120 voice source case, the results indeed showed the average number of bandwidth units required at about fifteen, as one would expect when referencing a standard call table.

Modem only simulations were also performed using deployment rules described above, again to validate the analytical models developed in section 4. Two configurations were considered, one using a strictly Poisson arrival process, and the other using the "birth-death" Poisson arrival process. Ten simulations were run on each configuration, adjusting the seed value for OPNET's random number generator prior to each run. In the 40-modem source case, the results indeed exhibited a marked difference between the simple Poisson process and the birth-death process. We observed in the Poisson process simulations with an infinite population (Figure 5) that the number of modem users in the system steadily increased to a value larger than the original number of modem sources, which is clearly impossible. In simulations using the birth-death model for modem traffic as shown in Figure 6, we observed that the arrival behavior is such that the BU resources required is limited to roughly one half the total number of sources over the course of the simulation. This turns out to be consistent with the analytical representation of equation 8. We therefore conclude that it is inappropriate to model analog modem traffic in the same fashion as we model voice call traffic.

Lastly, simulations of 40 modem sources using the birth-death model for arrivals were run. This time however, the Weibull distribution was used for the service time instead of an exponential distribution, using a shape parameter of 0.4, which produced a heavier tail, giving us more modem

calls in excess of 55 minutes. As the shape parameter is reduced, we see more arrivals whose service time (length) exceeds the mean, thus producing an even heavier tail. In simulations that mix all traffic types, we used a single value (0.4) for the shape parameter.

Our verification of the VBR traffic generator consisted of disabling both voice and modem arrivals. We then observed the behavior of the VBR generator using several scenarios. One scenario was to produce arrivals of child processes at a constant rate, with constant child duration shorter than the child process' inter-arrival time. By stipulating a constant packet size and inter-arrival rate, we then produced a constant stream of packets with no expected blocking or overlap. This was important in order to verify and debug the basic functionality of both the traffic generator and the BU Server. Subsequent scenarios consisted of manipulating the various process model attributes per the results and subsequent theorems presented in [20]. An example is a packet train with an arrival characteristic that is Poisson, with a fairly short duration and exponentially distributed or constant time between packets. Additionally, these trains could have inter-arrivals on the order of tens of seconds, mimicking the behavior of a single computer user who is downloading a web page, then pondering that page before moving on.

Consider such a data only simulation (i.e. no voice or modem traffic) whose parameters are such that the desired result might mimic a single computer user "browsing" the web. In this case, each child process acts like a "mini-session", with arrivals being Poisson distributed with $\lambda = 15$ seconds. This is intended to represent an entire ON/OFF sequence composed of an ON period of several closely timed bursts followed by many seconds of OFF (idle) time. We assume that each upstream request (mouse click to request a web page or file) is both short (a small packet) and always successful. Thus, the ON time can be assumed to be composed of primarily downstream traffic, ignoring the occasional upstream acknowledgements. During the ON time, packets arrive at a constant rate of every 0.051 seconds [10], whose size is normally distributed about a mean value of 992 bytes. The child duration pdf in this case has a Weibull distribution with a shape parameter of 0.6 and a mean time of 1.68 seconds (representing the ON portion of the ON/OFF sequence). Thus, an average ON period would offer about 32K bytes of data to the medium. The value of 32K bytes was derived empirically by counting the contents of a random sampling of several actual web pages.

In Figure 7, the individual graphs show that the traffic is indeed bursty when observed on several time scales. However, these traces in no way attempt to show or prove that the traffic contains any significant statistical trait such as self-similarity. The use of a Weibull distribution for the child duration (ON period) parameter does produce a heavy tail such that there will be more arrivals with longer ON periods. This enables us to represent up to approximately 1M byte of data offered to the medium for some ON periods.

5.3 Mixed Traffic Simulation Results

The central purpose of this work is to explore and gain an intuition into potential QoS ramifications of mixed traffic classes super-imposed on a prioritized shared medium. Recall from sections 1 and 2, that our model represents a real world system, designed and deployed initially to accommodate telephone voice traffic and using traditional telephony (Poisson process) principles. The emergence of modem access to the Internet introduces a class of traffic not considered in the initial design of such systems. The addition of high-speed data access capability to the system further complicates the traffic mix. Assuming reasonable and common deployment ratios (users to BUs), we are looking primarily for two things:

1. What is the impact on the perceived QoS when modem users are introduced into the traffic mix?

2. What is the impact on the perceived QoS as the "mix" of data users shifts from modem use to high-speed data use?

In mixed traffic source combination simulations, we examine three commonly deployed concentration ratios namely, 2:1, 3:1, and 4:1. Within each concentration ratio, four different traffic class

'mixes' were simulated, for a total of twelve different simulation "scenarios". The basic "voice-to-data user" mix was based on estimates of home computer penetration and the percentage of these users who access the Internet by way of modem. Each scenario was simulated at least three times, using different seed values for the OPNET random number generator. No significant statistical variation was observed across simulations where only the seed value was altered.

For the 2:1 concentration ratio, simulations were run on a traffic mix of 40 voice, 20 modem, and zero data users. The traffic source mix then was altered to introduce a single data user, becoming 41 voice, 19 modem, and 1 data user in accordance to our deployment rules. The mix was then altered again producing 50 voice, 10 modem, and 10 data users, then finally 60 voice, zero modem, and 20 data users. We are interested in examining any queueing delay experienced by high-speed data users resulting from prioritization of the mixed traffic which could be perceived as degraded quality of service. The 2:1 concentration ratio results show negligible delay for the data user, as clearly shown in Figure 8. The only difference of note is that as the number of data users increases (bottom two traces) the small queueing delay of only a few milliseconds is constantly present instead of only occasionally as seen in the top trace.

The 3:1 concentration ratio scenarios begin to exhibit more interesting behavior. In the zero data user case, the simulations suggest that the medium may be in danger of becoming congested with voice and modem traffic. Using different seed values in this scenario produced only a minor variation (1 or 2 bandwidth units) in the average number of bandwidth units consumed. Figure 9 shows that a single data user (top trace) could be at risk for a perceived degradation in QoS. This is due to the observed queueing delay momentarily exceeding one second in duration. However as the number of modem users is reduced in favor of high-speed data users, we note that the queueing delay quickly dissipates to a level at least an order of magnitude less than the single high-speed data user case. This implies that if modem use could be drastically reduced (or eliminated) high-speed data users should not be concerned with QoS degradation.

Finally, the 4:1 concentration ratio behavior begins to show not only QoS degradation for data users, but also for voice and modem users as shown in Figure 10. The requested bandwidth significantly exceeds the available bandwidth as the simulation approaches a steady state. At this point, with zero high-speed data users, the medium is becoming saturated, and blocking of modem and voice traffic is observed. Recall that 120 voice only users produced an average bandwidth unit requirement of 15, or 1/2 of the capacity of the medium, matching the values that one would see in a standard call table. Also recall that simulations involving 40 modem users required about 20 bandwidth units on average. The introduction of mixed voice and modem users whose sum is the same number of sources cited above (120) substantially alters the behavior, to the point of congestion on the medium. Considering the average length of voice and modem calls, (4 and 55 minutes respectively) it is not difficult to anticipate that a small number of high-speed data users could expect severe QoS degradation.

As modem users are replaced with data users, the resulting queueing delay behavior resembles that of the behavior observed in Figure 9. Figure 11 shows that as more data users replace modem users in this concentration ratio scenario, the queueing delay is drastically reduced, but not eliminated. In the cases where few data users are in the system (top two traces of Figure 11), the queueing delays are now on the order of a minute or more. The QoS in these cases would most likely be perceived as severely degraded. Even in the case where no modem users are in the system, there are random queueing delays on the order of one second, again orders of magnitude higher that those seen under the reduced concentration ratio scenarios. However, even in this common concentration ratio scenario the simulation results suggest that if modem users are eliminated in favor of high-speed data users QoS ramifications are reduced to the point of likely acceptability.

5.4 High-Speed Data Users Only Simulation

In section 2, comparisons and contrasts were explored between Packet-Switched (PS) and Circuit-Switched (CS) HFC access systems. In [8], it is noted that well over one thousand users can share the medium before perceived QoS degradation occurs. In the CS case, we ran simulations to

determine 'what-if' the medium was being shared by only high-speed data users, with no voice or modem traffic. In this case, 256 high-speed data traffic data sources were simulated (the maximum number allowed by the real-world system). Figure 12 indicates that the aggregate behavior of the traffic was consistent with prior simulations used to verify the behavior of a single user as well as a small number (10's) of users. The top trace (child duration) illustrates the distribution of child processes (sessions) showing that most child processes are of short duration (web page browsing) while exhibiting a reasonable number of longer child processes (file downloads). The middle trace (offered load) shows the aggregate average bandwidth consumed in bits per second. The figure of 400,000bps corresponds to an average of about 7 BU resources consumed during the simulation. Lastly, the queueing delay (bottom trace corresponding to degradation of QoS) is negligible. This result is encouraging since it implies that headroom exists for either more users on a particular shared medium or more bandwidth-consuming applications can be supported by the medium. Both of these topics are areas for consideration of potential future work.

6 CONCLUSION

In this work we have gained an intuition as to the potential of QoS degradation to users of a real world prioritized shared medium. Analytic models of three traffic source classes (voice, modem, and high-speed data) and the resource allocation element (BU Server) were developed. Each entity was independently verified by way of simulation. As part of the independent traffic source model verification we observed that modeling modem traffic using an infinite population Poisson process model such as that used for modeling voice traffic produces unrealistic behavior. Moreover, we also observed that using a finite population, birth-death approach using Poisson arrivals and a heavy-tailed service time distribution yielded results closely approximating analytical estimates. Simulation results have shown that no risk to QoS exists in cases where 120 voice sources are modeled. This result is expected. Additionally, mixed traffic simulation results indicate that the risk for QoS degradation is small as the number of modem users is reduced and approaches zero. Conversely, as the number of modem users rises, particularly in the 4:1 concentration case, all classes of traffic suffer from QoS degradation. Our analysis and subsequent simulations of a small number of modem users go to re-enforce our assertions that

1. Modem use underestimates BU resource requirements during the busy hour when relying on traditional modeling techniques and

2. A reduced population (finite probability space) Poisson process model is required to properly analyze and simulate modem traffic.

While new mechanisms to allow high-speed data access are being developed to take advantage of unused bandwidth, it is likely that a gradual migration of modem users to high-speed data users will occur as bandwidth-intensive applications become more ubiquitous. The results of this work suggest that there could be undesirable QoS ramifications for high-speed data users, and even voice and modem users if only traditional voice traffic engineering techniques are considered and used to model the BHBD access behavior. The results from this treatment of modeling and simulating of this type of prioritized, shared medium could provide motivation for future work such as:

1. The development of more sophisticated simulation models to better quantify upstream (client) versus downstream (server) behaviors.

2. The development of additional data traffic source models to mimic streaming applications such as 'real video' or 'real audio' that could be included to place more diverse data traffic loads on the system.

3. The development of more comprehensive prediction tools for the design, deployment, and future growth of this type of network access system.

4. Algorithm developments for either reactive or pro-active network access congestion control mechanisms.

Acknowledgments

This work was supported under NSF ANIR 9984811. The authors also wish to thank the anonymous referees for their helpful suggestions to improve the presentation and technical content of this paper.

References

[1] J. Bellamy. *Digital Telephony, Second Edition*. John Wiley & Sons, Inc., 1991.

[2] R. Bolla and F. Davoli. Control of multirate synchronous streams in hybrid TDM access networks. *IEEE/ACM Transactions on Networking*, 5(2):291–304, April 1997.

[3] V. Bolotin. Modeling call holding time distributions for CCS network design and performance analysis. *IEEE Journal on Selected Areas in Communications*, 12(3):433–438, April 1994.

[4] Common Carrier Bureau. Trends in telephone service report. Industry Analysis Division of the Common Carrier Bureau of the Federal Communications Commission (FCC), Washington, DC, 1999.

[5] CableLabs. Packetcable 1.2 archectural framework technical report. Technical report, CableLabs, December 2000.

[6] D. E. Duffy, A. McIntosh, M. Rosenstein, and W. Willinger. Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks. *IEEE Journal on Selected Areas in Communications*, 12(3):544–551, April 1994.

[7] A. Dutta-Roy. A second wind for wiring. In *IEEE Spectrum*, pages 52–60. The Institute of Electrical and Electronics Engineers, Inc., New York, NY, September 1999.

[8] H. S. Fluss. Effective performance of shared bandwidth data channels in hybrid fiber/coax networks. In *the Society of Cable Telephony Engineers*. SCTE, January 1999.

[9] H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9(7):1139–1149, September 1991.

[10] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., New York, NY, third edition, 1998.

[11] R. Jain and S. A. Routhier. Packet Trains - measurements and a new model for computer network traffic. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):986–995, September 1986.

[12] W. E. Leland, M. S. Taqqu, W. Wilinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[13] B. Ngo and H. Lee. Queuing analysis of traffic access control strategies. *IEEE Journal on Selected Areas in Communications*, 9(7):1093–1109, September 1991.

[14] V. Paxson. Fast, approximate synthesis of fractional gaussian noise for generating self-similar network traffic. In *Computer Communication Review*, volume 27, pages 5–18. ACM SIGCOMM, October 1997.

[15] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modeling. In *Proceedings of ACM SIGCOMM '94*, pages 257–268, London, August/September 1994.

[16] R. Schafer, J. Jungjohann, and J. Bezoa. Passive optical networks - is there light at the end of the access tunnel? CIBC World Markets, Inc., January 2001.

[17] Y. De Serres and L. G. Mason. A multiserver queue with narrow- and wide-band customers and wide-band restricted access. *IEEE Transactions on Communications*, 36(6):675–684, 1988.

[18] B. Vandalore, G. Babic, and R. Jain. Analysis and modeling in modern data communications networks. In *Submitted to the Applied Tele-communications Symposium*, 1999.

[19] X. Wang and S. C. Chang. On the performance study of several access control strategies in isdn. In *Conference Record IEEE ICC 88*, volume 1, pages 934–938, 1988.

[20] W. Willinger, V. Paxson, and M. S. Taqqu. *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, chapter Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic. Birhauser, 1998.

[21] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high variability: Statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, April 1997.

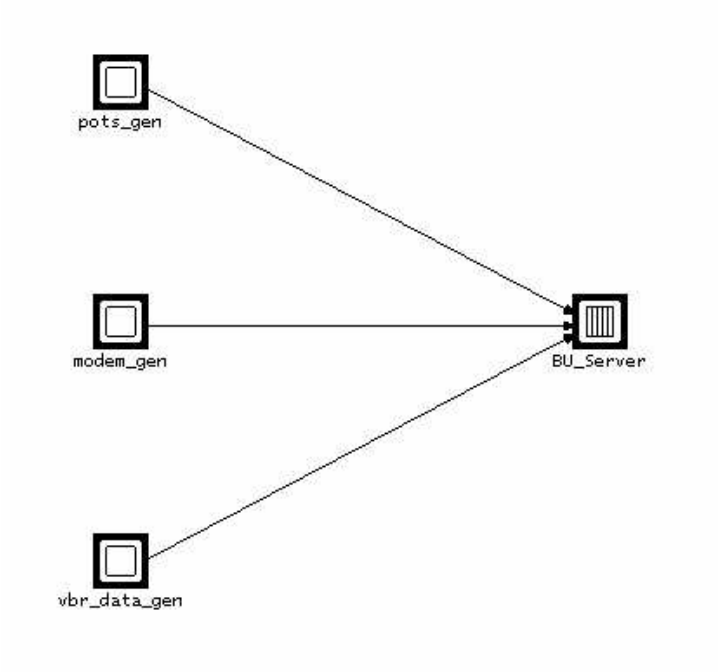

Figure 4: Bandwidth Allocation Node Model

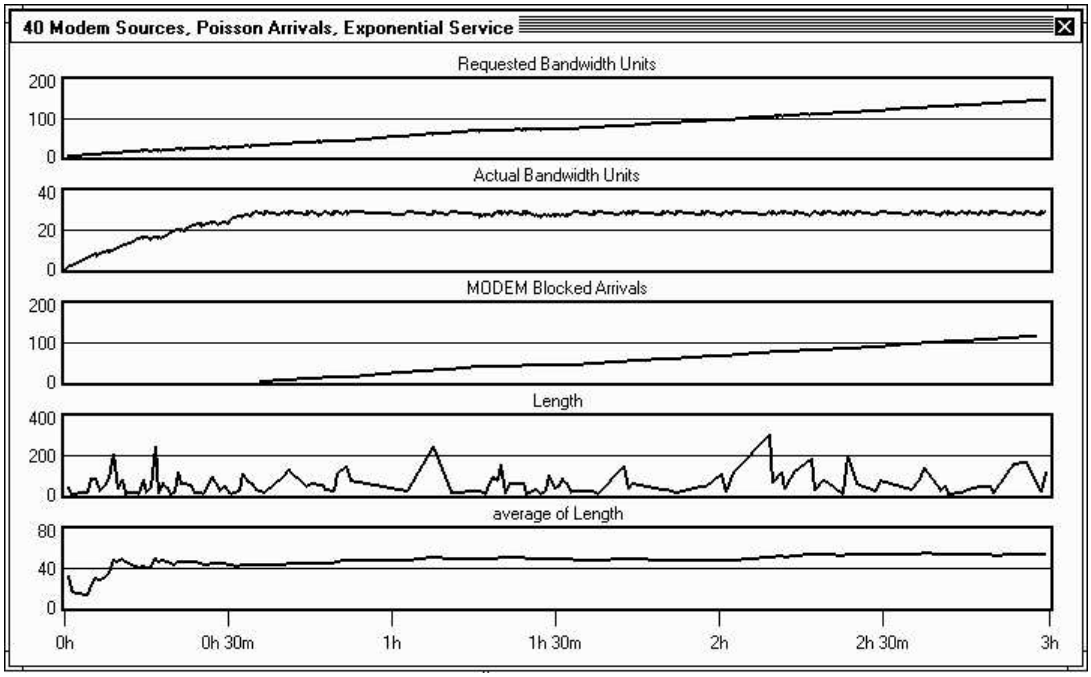

Figure 5: Modem Call Generator Simulation (40 Sources)

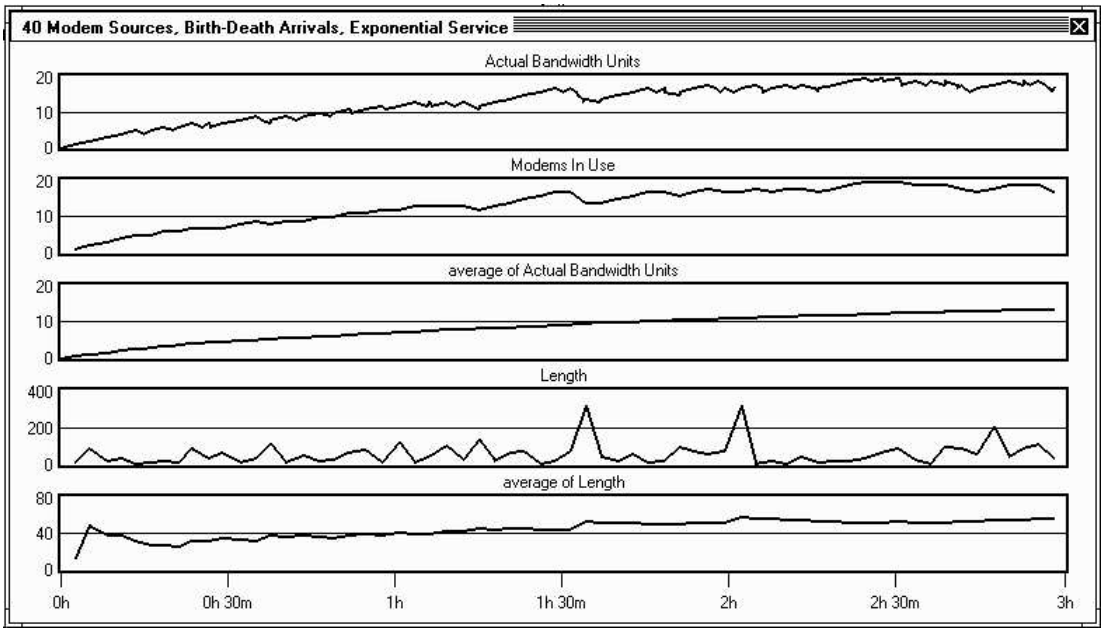

Figure 6: Modem Call Generator Simulation (40 Birth-Death Sources)

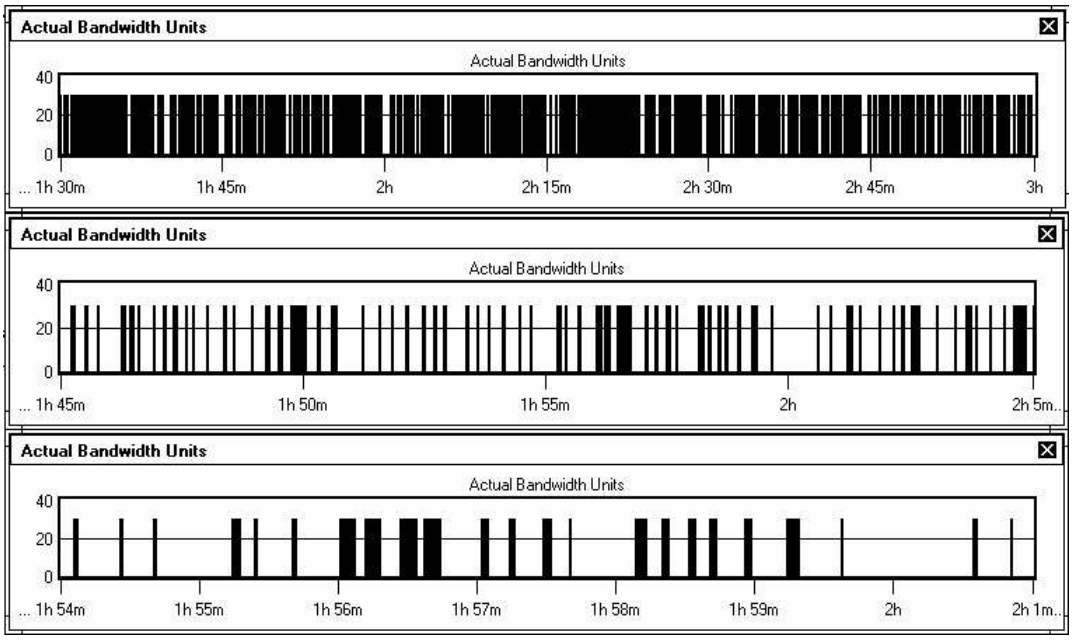

Figure 7: Data (VBR) Traffic Generator Simulation (One Source)

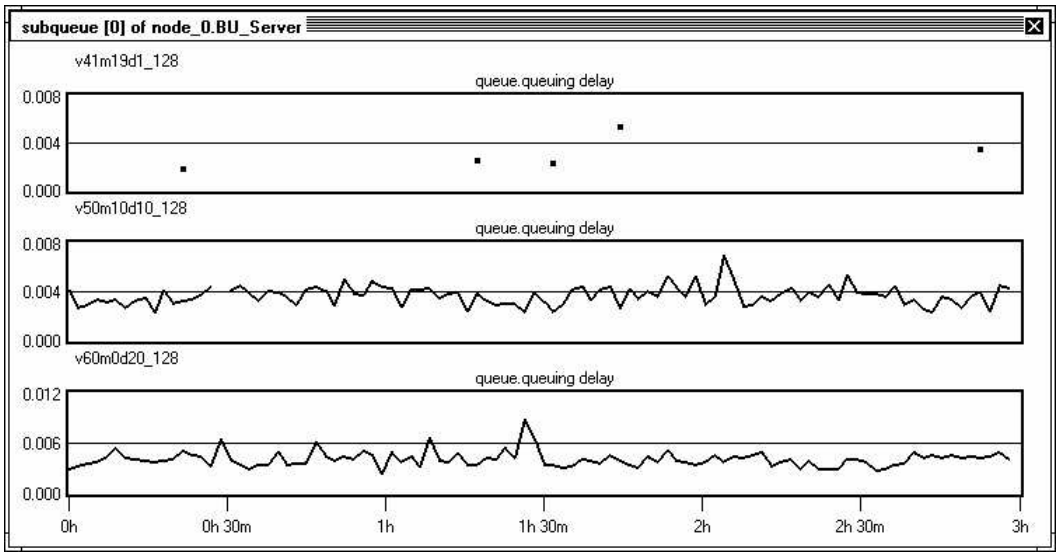

Figure 8: Queueing Delay 2:1 Concentration Ratio)

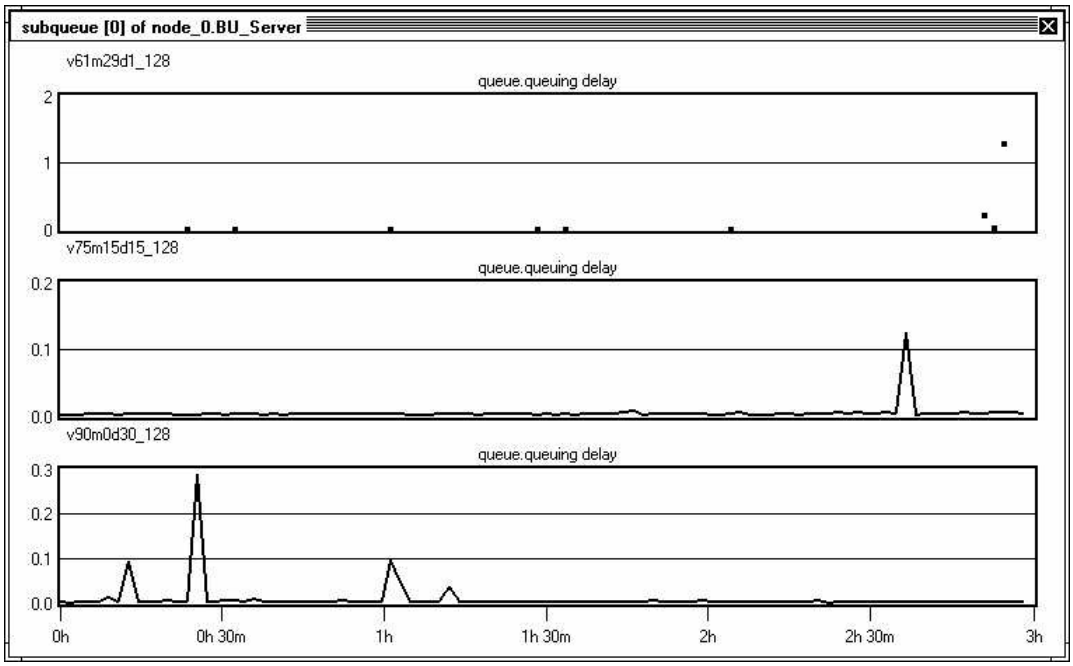

Figure 9: Queueing Delay 3:1 Concentration Ratio

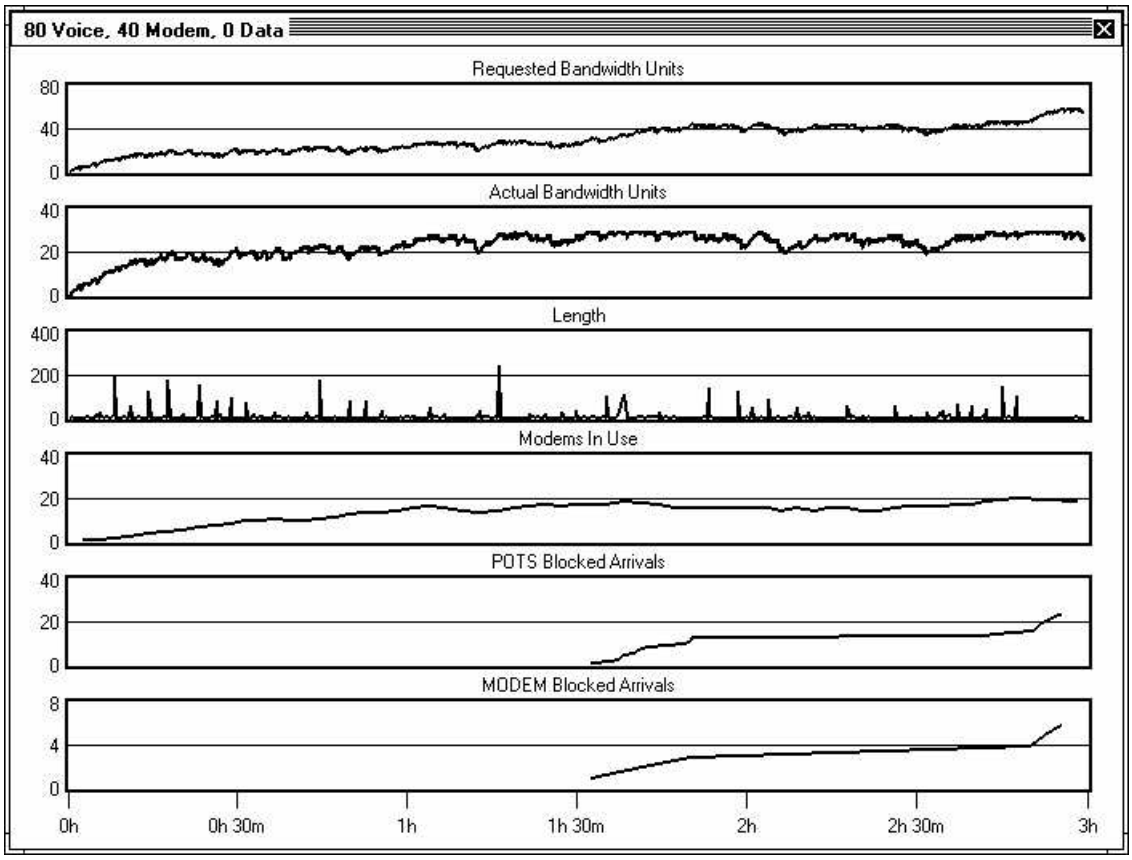

Figure 10: Voice/Modem Simulation Results, 4:1 Concentration Ratio

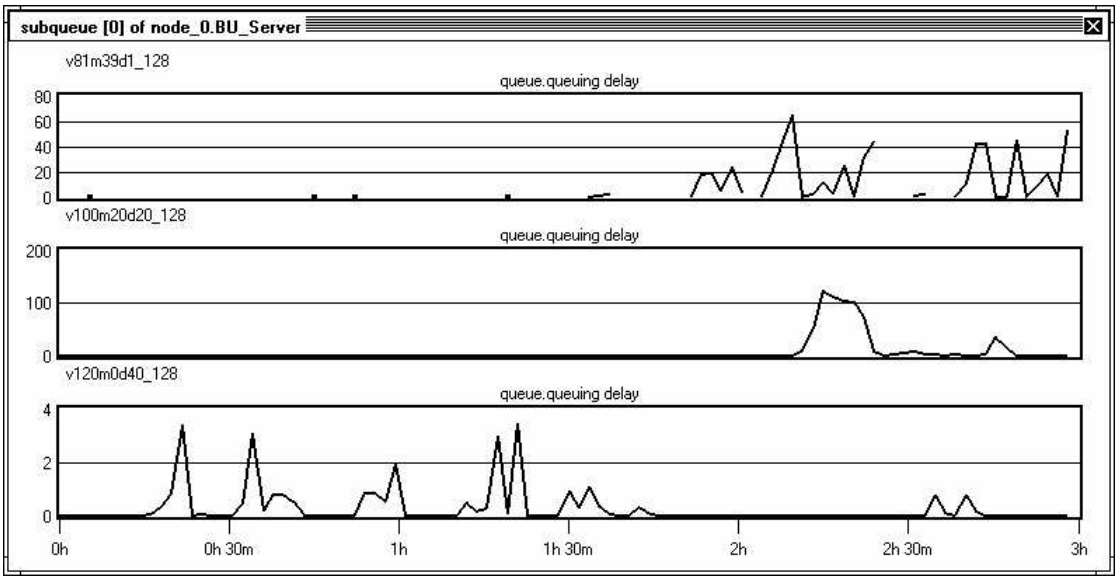

Figure 11: Queueing Delay 4:1 Concentration Ratio

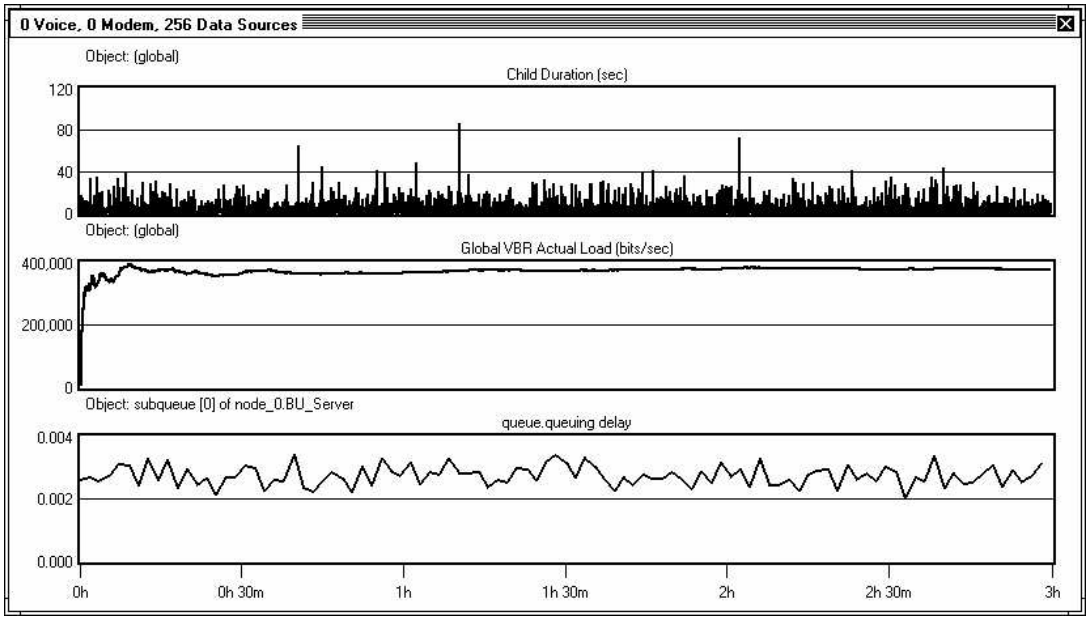

Figure 12: 256 High-Speed Data Users