Modeling and Simulation of Mixed Traffic on a Prioritized Shared Medium

Jeffrey J. Evans
Email: evanjef@charlie.cns.iit.edu

Cynthia S. Hood
Email: chood@charlie.cns.iit.edu

Department of Computer Science
Illinois Institute of Technology
10 W. 31st St.
Chicago, Illinois 60616

Abstract -- **Network access systems (NAS) such as digital loop carriers (DLC) are increasingly utilizing a shared medium, such as Hybrid Fiber Coax (HFC) to provide point to multi-point access from the public switched telephone network (PSTN) to the end user (consumer). New services, such as direct access to the packet switched network (PSN, WWW) have been added to DLC equipment in such a way as to provide for a prioritized set of services over a shared medium in an effort to take advantage of otherwise unused bandwidth. With the introduction of such services comes the added complexity of traffic analysis and modeling these network access systems, particularly when considering the variability in different service type traffic characteristics. This work identifies a traffic engineering problem of prioritized circuit switched and packet switched (PSTN/PSN) traffic over the same shared medium as it may relate to "perceived" quality of service (QoS). A brief review of the evolution of models for communications traffic is presented. Models are then proposed and developed for the various traffic types being considered, which include voice, modem, and data. These models become the basis for running simulations using the OPNET Modeler simulation software package, modeling a prioritized shared medium with finite bandwidth under varying source and offered load conditions. Simulation results suggest that at high (but typical) concentration ratios, small numbers of data users could experience QoS degradation when combined with sufficient numbers of modem and/or voice users. Finally, recommendations for potential future work are offered.**

I. Background

The individual statistical characteristics of voice, analog modem (modem), and data traffic play a critical role in the performance analysis and design of communications network access systems (NAS). In particular, those access systems that employ point to multi-point, or shared access to the services being made available. Understanding the arrival, service time, and variation characteristics for these sources of communications traffic can help in better designing NAS topologies, and in designing both reactive and pro-active mechanisms for delivering the highest QoS to all users.

Delivery of voice and data services over a shared medium such as hybrid fiber coax (HFC) is becoming a serious competitor to the traditional twisted pair copper lines that have been the mainstay of the telephone network since its inception. Coaxial cable has an advantage of having huge amounts of bandwidth capacity on a single cable. Traditionally this bandwidth has been dedicated to the delivery of broadcast video (cable television, or CATV). In recent years the cable data modem has emerged for the purpose of delivering data service along with the video service, although in some cases separate cables are deployed to deliver the different services. While cable data modems do have large data bandwidth capacity (10Mbps or higher, [7]), they have not been designed to carry real-time, traditionally circuit-switched traffic such as telephone voice traffic. This is due in part to their asymmetrical transport nature (i.e. large downstream capacity, reduced upstream capacity). Emerging transport techniques, such as voice over IP (VoIP) are still in the development stages and are for the most part, experimental, with several QoS and operations related issues outstanding.

A separate class of HFC NAS products has been developed to fill the gap between the high data rate cable data modem and the traditional twisted-pair copper telephone lines. These products possess the capability of delivering traditional CATV service, along with circuit switched telephone traffic (Plain Old Telephone Set or POTS), on the same coaxial cable. They generally use more robust modulation technologies for the POTS transport, in part due to the fact that the medium must be capable of symmetrical (equal downstream and upstream) bandwidth. Hence, they tend to deliver less bandwidth capacity (bits/Hz) than cable data modems.

In some point to multi-point access configurations, whose primary service is to supply access to the PSTN, many end-

users compete for a finite number of "Bandwidth Units" (BU's), or "timeslots" (in the sense of TDM). For POTS services, timeslots normally take the form of a 64Kbps DS0 on a TDM transport from the end-user to a "head-end" (HE). The information is then cross-connected (mapped) to a T1, E1, DS3, OC-X, etc., for transport to and from the PSTN. Figure 1 illustrates such a system. Note that some of the endpoints are represented with both a telephone and a computer. The computer gains access to the packet switched network (WWW) by way of analog modem, but must go through (or at least into) the PSTN first.

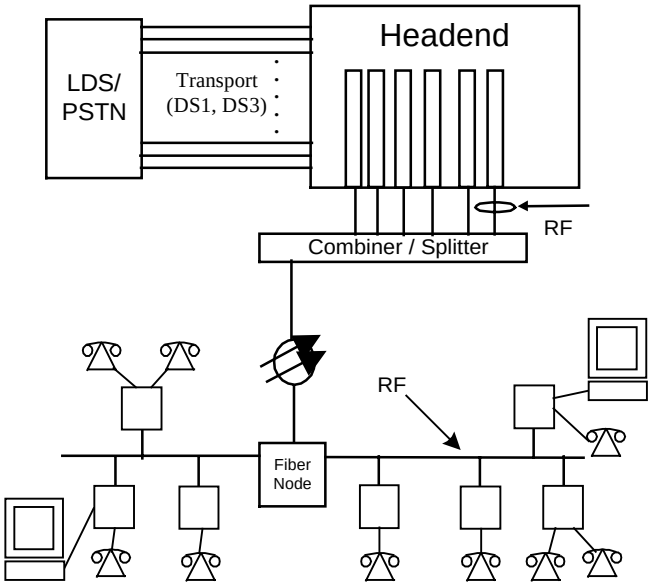

Figure 1. HFC Telephone Network Access System

Recent developments in components making up the HFC portion of the network have resulted in the capability of direct Ethernet-like data access from the end-user to the head-end, then on to the packet switched network (PSN or Internet), similar to that used by cable data modems, but with a twist. Figure 2 illustrates a similar topology to Figure 1, but now direct access to the packet switched network is available. By available we mean that "when BU's are available", a data user can access the shared medium's resources (DS0's) at a reasonably granular level in time. The twist to this notion is that access to the medium is <u>always</u> granted to a telephone user first. In other words, as BU's are consumed by telephone users, available bandwidth unit resources are being depleted, up to the point where if all BU's are consumed, no transport resources are available for data traffic (or new telephone users). In a data only situation this may not be much of a problem, since the nature of data traffic is bursty (section III) so the data user (or application) could reasonably expect that bandwidth would become available fairly soon. As such the end-user may not perceive any abnormal delay.

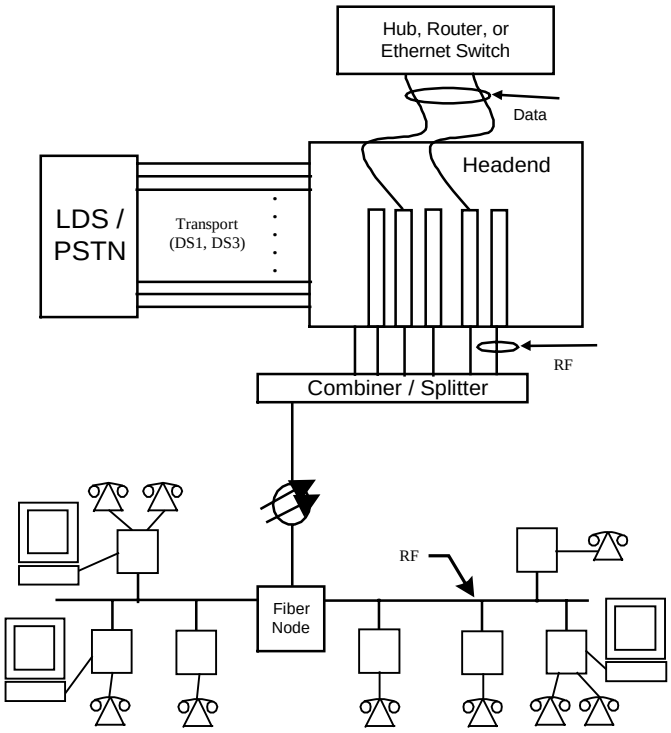

Figure 2. HFC Telephone / Data Network Access System

Additionally, if a telephone user requests service (i.e. goes off-hook or is called) while a data user is in the middle of transmission or reception, the resource required for the telephone user is dynamically re-allocated to the telephone user. This reduces (or eliminates) resources to the data user, resulting in longer burst times (reduced bandwidth case) or added delay due to queuing. Moreover, once resources are consumed (due to circuit switched telephone traffic), they tend to be consumed for minutes at a time. The resources may be consumed even longer if the user is gaining access to the Internet via a modem instead of using the data service. "According to the latest quarterly report from America Online, Dulles, Va., which has nearly 19 million subscribers worldwide, the average user spends 55 minutes per day on the Internet" [6]. These factors, together with issues associated with resource concentration vs. quality of service, further complicate the tasks of network design and growth, given this diversity of offered services and the service provider's desire to provide the best quality of service for the lowest cost.

The design and deployment of network access systems such as that depicted in Figure 1 has in the past been based solely on the principles and techniques used for deployment of telephone services (section III). Modem access tends to complicate call holdup time statistics due to longer call hold times. This is typified by a "hump" in the classic exponential service time characteristic, creating a sort of "bi-modal" effect [2],[5] in the call hold time distribution, further complicating the traffic engineer's network design task.

We will consider a system such as that illustrated in Figure 2 that uses an RF transport medium such as that in Figure 3 for this work.

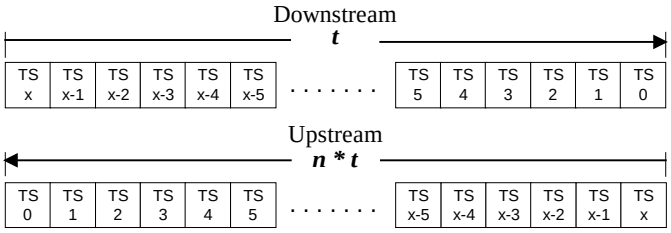

Figure 3. HFC Telephone/Data RF Medium Access Timing

Specifically, the downstream (toward the end user) transport is loosely based on a frame comprised of an integer number of usable payload timeslots (TS), with each timeslot consisting of eight bits. The transmission scheme is such that the aggregate bit rate is on the order of 2Mbps. A subset of timeslots are utilized for non-payload carrying functions such as synchronization (framing), signaling (on-hook/off-hook), and a lower speed data link used by the single point "master" to communicate to the customer premise equipment (multi-point slaves). This data link communication consists of provisioning, 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 slave, then each slave 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 POTS bandwidth is realized over a low-speed data link. This link is again part of a subset of timeslots for non-payload carrying functions. Communication on this link is realized using a common medium access technique, where each slave may attempt to communicate (request bandwidth) at a single point in time (timeslot). We will be focusing on POTS arrival rates on the order of 10 to 20 seconds per arrival, several orders of magnitude beyond which consideration of details of the medium access protocol is required.

A different medium access system is employed by which access to bandwidth is possible for data traffic. The method used employs control bits in the downstream manipulated by the master. These bits inform the listening slaves 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 POTS traffic comes and goes. This dynamic mapping of available payload bandwidth establishes the medium's priority mechanism. The upstream granularity of access is on the order of a millisecond and the volume of data may be as high as the full capacity of the transport bandwidth (on the order of 2Mbps). In the downstream, a single master controls data access, therefore the granularity of access is much finer, on the order of microseconds. Lastly, the medium acts asymmetrically for data access. This means that the master can transmit data in the downstream direction to slave A while slave B (or a different channel on slave A) transmits data in the upstream direction toward the network.

Present deployments of systems using a prioritized, shared medium such as that previously discussed have been designed primarily for telephone voice traffic, using techniques that while valid for voice traffic, fail to consider the ramifications of data traffic, be it via modem or direct high speed (available bandwidth) access. It is likely that the conversion from modem data access to users taking advantage of the potentially higher data rate access offered by the medium will neither occur in mass nor instantaneously. It is more likely that this type of medium will be subjected to various intensities of telephone, modem, and data traffic simultaneously, until such a time as modem access to the PSN becomes truly uneconomical for the end user.

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. Because the medium is not fair, with a bias toward voice and modem users, we wish to determine if data users could reasonably expect substantial delays under typical deployment scenarios with generally accepted voice/modem traffic loads.

We must begin by mathematically modeling the various traffic source types that would request resources (bandwidth) from the prioritized, shared medium described above. These traffic types include voice (POTS), modem (used for access to the PSN), and data traffic. We also need to determine the "Busy-hour, Busy-day (BHBD) characteristics of each traffic type. This is needed to establish any gross phase relationships between the arrival patterns of the different traffic types during the BHBD. Then we need to develop representations of each of the traffic sources and a representation of the medium using a discrete-event simulation tool, such as OPNET Modeler. Lastly, we wish to use OPNET Modeler to run simulations. These simulations would consist of verifying the statistical properties of individual source types, then verifying the properties of multiple instances of the same source type. Simulations can then be developed that combine the different source types in such a way as to 'mimic' actual BHBD traffic on typical deployments.

II. 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 mid-1980's work has been done in this area. It is well understood that voice traffic access to the PSTN follows cyclical patterns. These patterns are different depending on the time scale and can be modeled successfully using simple homogeneous Poisson processes. Designers of

telephone networks typically design to address what is known as the "Busy-Hour, Busy-Day" (BHBD) parameter. The idea is for the system to be capable of handling BHBD traffic with a low probability of blocking (typically less than one percent). More recently, adjustments to call service times have been also modeled to account for the fact that the cost of telephone service has dropped in recent years.

As recently as the early 1990's researchers have been misled by the statistical nature of data traffic. It is now generally recognized that modeling data traffic using Poisson or compound Poisson processes is generally inadequate [10],[11],[13],[15]. 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 queuing theory [4],[12],[16], 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 data access have become outdated [11],[13].

Another area where substantial research has been undertaken is that of Asynchronous Transfer Mode (ATM). One of the main features of ATM is the idea of simultaneously transporting different service types/classes of traffic. In other words, traffic requiring a constant bit rate (CBR) such as telephone voice traffic, can be transported along with traffic requiring a variable bit rate (VBR), such as data traffic, or more specifically, multiplexed compressed audio and video application traffic. The notion of available bit rate (ABR) exists for those data applications tolerant of true "best effort" performance, such as TCP. Over the last decade there has been significant work in the ATM field. Much of the work has been focused on issues specific to the ATM protocol, ATM switches, their control and applications, such as buffer sizing and congestion control algorithms.

For our study there is some similarity between our prioritized shared medium and the notion of mixed CBR and VBR (or ABR) traffic. More importantly, research in ATM has produced mathematical models resulting in algorithms and techniques for generating data traffic that closely resembles that of measured LAN and WAN traffic. Some of these methods [14] have been presented to the ATM Forum as recommendations for "reference load models" for ATM ABR performance testing (ATM FORUM 96-1568). We are interested in the results of this type of work, since these models may have application as data traffic generators in our study.

In [7], the author addresses the notion of performance on a shared HFC medium. The medium in this case has a capacity of 10Mbps. It is not prioritized, and it addresses only data traffic (i.e. cable data modem). Our study addresses both circuit switched voice and data traffic. The techniques used to generate data traffic can be of use to this study as they utilize 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 a reasonable, yet simplifying assumption of a "homogeneous, non-dispersive Web", in an effort to isolate the access portion of the data network from the Internet itself. Their 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.

III. Traffic Characteristics

The is an abundance of studies in the literature that have theorized, measured, and analyzed the nature of both traditional voice and data traffic as it has evolved. It is widely accepted that a Poisson process with negative exponential service times can describe voice (POTS) traffic arrivals. Models using Poisson and Erlang distributions have been used for decades to determine the number of resources (trunks, DS0's) required to guarantee a particular grade of service (QoS, or blocking probability) for a given traffic intensity, given a number (infinite or finite) of traffic sources (originators).

The notion of busy-hour, busy-day (BHBD) has been used for telephone network design purposes since the early days of telephony. The arrival process (users attempting to initiate a phone call) has been measured and characterized over various time scales such as yearly, weekly, and daily and has been shown to be cyclical. Additionally, different 'classes' of users such as business and residential can be seen to produce additional variation in the arrival process [1]. For this study we will be concerned with residential users, since the principle purpose of the shared medium (and most common application) at this time is to provide CATV, voice, and data services. Figure 4 illustrates the daily arrival process of voice traffic for residential users.

More recently, the arrival process of data traffic has been measured in a similar fashion. In ([13], Fig. 1, page 228), the authors show significant variations in the daily connection arrival rate for various connection types (TCP applications such as FTP, Telnet, SMTP, etc.). The main purpose for this figure was to demonstrate that a simple homogeneous Poisson process cannot be used to model the arrival process. However, one can also take from this data another significant point. That is, the period of time that the data arrivals are highest literally overlaps that of the telephone traffic of Figure 4. Therefore, for the purposes of this study we assume that all arrival processes coincide with each other over the course of a BHBD simulation interval.

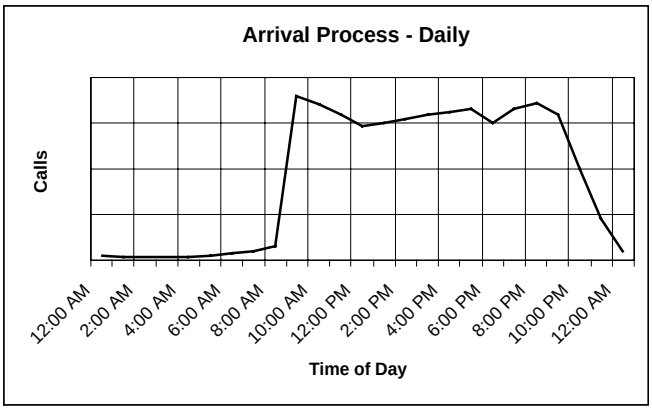

Figure 4. Daily Residential Telephone Traffic Arrival Process

A. Voice Traffic

The probability distribution most commonly used for modeling POTS (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 units of Erlangs and Century-Call-Seconds (CCS). Traffic benchmarks are commonly used around the world to quantify this traffic. 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;

$$\Pr(n,t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (1)$$

If two Poisson processes N_1 and N_2 with arrival rates λ_1 and λ_2 are impinging on a single medium we show that the resulting process is a Poisson process with arrival rate $\lambda = \lambda_1 + \lambda_2$ as follows. Since both processes are Poisson, we can assume independent and stationary increments. Additionally, arrival events are mutually exclusive from one another such that:

$$Pr(N(h)) = Pr(N_1(h)=1) \cap Pr(N_2(h)=0) + Pr(N_1(h)=0) \cap Pr(N_2(h)=1)$$

$$= Pr(N_1(h)=1) * Pr(N_2(h)=0) + Pr(N_1(h)=0) * Pr(N_2(h)=1)$$

$$= \lambda_1(h) + o(h) * 1 - \lambda_2(h) + o(h) + 1 - \lambda_1(h) + o(h) * \lambda_2(h) + o(h)$$

$$= (\lambda_1(h) + o(h) + \lambda_2(h) + o(h)$$

$$= (\lambda_1 + \lambda_2)h + o(h) \quad (2)$$

Expanding this result to *n* sources results in a Poisson process whose arrival rate is the sum of the individual arrival rates;

$$\lambda = \sum_{i=0}^{n} \lambda_i \quad (3)$$

We must now determine the traffic intensity offered to the medium for an individual POTS voice source. From [3] we establish the following call traffic benchmarks from 1997:

- Average Telephone line is in use 57 minutes per day.
- 0.95 Erlangs Per Day (0.95 * 36 CCS/Erlang = 34.2 CCS Per Day)
- 0.114 Erlangs (4.1 CCS) During Busy Hour (Assumes 12% of Per Day Total)

Using an exponentially distributed service time statistic with a mean value μ of 4.0 minutes (240 seconds) we can calculate the per source arrival rate λ_i (from eq. 3) as;

λ_i = offered load per line / offered load per call

offered load per line = 4.1 CCS

offered load per call =

4.0 min. per call / 60 min. per hour = 2.4 CCS

λ_i = 4.1 / 2.4 = 1.7 call per hour (BHBD)

For this study we will increase the call rate slightly in order to have round numbers to work with. Thus, the Poisson process parameters for voice traffic become:

Arrival Rate λ_i = 2.0 call/hour and

Mean Service Time μ_i = 4.0 minutes/call,

which increases the per line intensity from 4.1 CCS to 4.8 CCS. Stated another way, the average line daily usage increases from 57 minutes per day as cited above (1997 data) to 67 minutes per day. The authors of [3] also cite the following traffic intensity growth rate increases from 1994 to 1997;

- 1994 to 1995 1 percent

- 1995 to 1996 3 percent
- 1996 to 1997 6 percent

Extrapolating conservatively from the 1997 data using the same six percent per year growth rate calculates to 68 minutes per day, which is within two percent of our 67 minute per day figure. Lastly, our mean service time value of four minutes per call works well, resulting 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.

B. 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. This has the effect of changing the service time statistic into one that shows signs of an aggregate "bi-modal" characteristic [2]. Studies such as [5] 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)". In [2], the author uses the notion of mixing two Gaussian distributions to form a composite probability distribution of call holding time. Our interest lies in the long call area where a significant portion of the population (approximately 20%) had call holding times in excess of 10 minutes.

From these studies, it is evident that a Poisson arrival process model is probably acceptable for POTS and analog modem arrivals. One significant difference is present however. For POTS voice traffic, it is reasonable to assume that the probability sample space is infinite, due to the combination of originating and terminating calls. We know that the number of originating call sources 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 can conclude that the number of sources is large enough to be considered infinite. This is not true for analog modem traffic however. For analog modem traffic, we assume that all calls are originating calls, as direct-dial bulletin board servers of a decade ago have yielded to "web sites". Therefore, the probability sample space cannot be considered infinite.

For this case we can turn to birth-death theory [9] and develop the following birth and death-rates of the system;

Births (arrivals)

$$\lambda_n = \begin{cases} (M-n)\lambda & (0 \le n < M) \\ 0 & (n \ge M) \end{cases}, \qquad (4)$$

and deaths (service times)

$$\mu_n = \begin{cases} n\mu & (0 \le n < c) \\ c\mu & (n \ge c) \end{cases}. \qquad (5)$$

To illustrate the case where the number of sources is equal to the number of servers we construct the state dependent process shown in Figure 5.

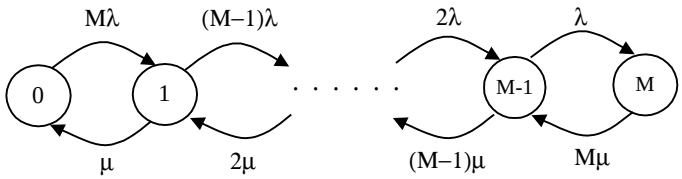

Figure 5. State Dependent Birth-Death Process

In the state-dependent process, M is the number of modem sources and the state represents the number of modem sources in a call. Assuming there are c servers in the system, the state probabilities reduce to [9]:

$$p_n = \begin{cases} \binom{M}{n}\left(\frac{\lambda}{\mu}\right)^n p_0 & (1 \le n < c) \\ \binom{M}{n}\left(\frac{n!}{c^{n-c}c!}\right)\left(\frac{\lambda}{\mu}\right)^n p_0 & (c \le n \le M) \end{cases} \qquad (6)$$

Since this form of p_n does not allow the closed form calculation of p_0, we must calculate each of the coefficients in equation 6 and complete the computation [9] as:

$$p_0 = \frac{1}{1 + a_0 + a_1 + a_2 + ... + a_{M-1} + a_M}. \qquad (7)$$

We are most interested in the average number of modem sources in a call when the system is in a steady-state. To obtain this figure, we use the definition of expected value and get [9]:

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

We also wish to consider the service time distribution and mean value for analog modem traffic. It would be advantageous run simulations using either an exponential service time characteristic or one that could introduce a "heavier" tail as suggested in [5]. This can be achieved in OPNET by using a weibull distribution for the service time. By adjusting the "shape" parameter value to be $0 < \alpha < 1$, we can easily produce service time distributions ranging from exponential to one whose service time distribution possesses a "heavy" tail, with more service times longer than the mean.

We will adjust the source mean service time μ_i, and also the source arrival rate λ_i to compensate for the potentially long service times. For example, from [6] 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. Thus the single source process parameters for analog modem traffic become:

$$\text{Arrival Rate } \lambda_i = 1.0 \text{ call/hour and}$$

$$\text{Mean Service Time } \mu_i = 55.0 \text{ minutes/call.}$$

As suggested earlier, POTS arrivals follow cyclical daily, weekly, and even yearly patterns. In [8],[13] 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 POTS arrival daily cycle. In [8] it was seen 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 POTS, modem and data traffic will overlap.

C. Data Traffic

In [10], the authors demonstrated from their analysis of measured LAN data at M.I.T. a clear deviation from Poisson packet arrivals. 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 description of the "packet train" model, they observed that measured traffic exhibited a phenomenon that "successive packets tend to belong to the same train", such that 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. This notion is also dependent on the utilization of the network. As such, one could expect that under high utilization there would exist less source locality (i.e. more overlap of packets from different source-destination pairs). Subsequent studies argue that LAN and WAN packet arrivals are better modeled using statistically self-similar processes. These observations and those from [11],[13] clearly indicate that Poisson modeling of data traffic is incorrect, or at least outdated.

In [17],[18] the authors derive an elegant set of formulas and tests for the modeling of homogeneous sources using an ON/OFF source model. Their interest however is more focused on the statistical behavior of a stochastic process for large M (number of i.i.d. sources) and T (time). Another ramification of this technique is the generation of self-similar traffic using this approach. The authors explain the relative ease with which long traces (100,000 packets) of self-similar traffic can be generated in linear time - assuming of course a massively parallel (16K processors) computing environment. In [5],[7] simpler techniques are offered to generate self-similar or "near" self-similar traffic behavior. These techniques are based on Pareto and Weibull distributions producing a "heavy-tailed" packet arrival effect. More importantly, a massively parallel computing environment is not required to generate this type of near self-similar traffic.

Some of the more recent studies begin to consider different application types being used on the Internet. For example, HTTP sessions tend to show a "burstier" traffic pattern than a "streaming" session such as "RealAudio" or RealVideo" [7]. The addition of substantially more home businesses and those who telecommute make it increasingly more 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 this study 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 and when aggregated. 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).

In [18], texture plots are used to offer a "visualization" of the packet arrival process at the source/destination level. Later, limit results for aggregate WAN traffic are presented to assume that "sessions" (i.e. FTP, HTTP, TELNET) arrive according to a Poisson process, they then transmit packets deterministically at a constant rate, then cease transmitting packets. The main stochastic element left undefined is the session length, or duration. This is presented as possessing a long-range dependence, or heavy-tailed property.

D. The Medium

The medium itself can be modeled in several ways. The main objective is to avoid having to model the various access methods described earlier. In references [4],[12],[16], the authors chose to model the medium as a multiserver *M/M/N* queue, such that a server would be allotted with each bandwidth unit (or basic bandwidth unit, BBU). Various policies for the handling of blocking were considered. 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. Another approach, which we choose for our initial work with POTS and analog modem traffic, is to model the medium as a single server, infinite capacity queue $M/G/1/\infty$, such that any source asking for any amount of bandwidth (or bandwidth units) will get it. In this way, the resultant simulation data should provide an indication of bandwidth requirements per unit time or over time, needed under various load and source type conditions.

We then improve the medium model with the addition of a data traffic source to be a hybrid of $G/G/1/\infty$ and *G/G/1/N*,

where *N* is of finite capacity (buffers) representing the number of BU's on the medium. In this case, all blocked POTS or modem calls are lost, except when a data burst is in progress. When a POTS or modem call arrives and a data burst is in progress, the data burst BU's are reduced by a single BU, its remaining burst length is calculated, its original termination event is canceled and a new termination event is scheduled based on the remaining burst time. Blocked data bursts are queued, as would generally be the case with Ethernet data traffic, relying on the upper software layers to handle excessive delay scenarios. In the event that a data burst was using a single BU, its remaining burst time (in bits) is placed at the head of the queue, while all other blocked data bursts are placed at the tail of the queue. In this model we develop and keep statistics involving blocking and delay characteristics while maintaining data indicating both the desired and actual bandwidth unit requirements for POTS and modem traffic.

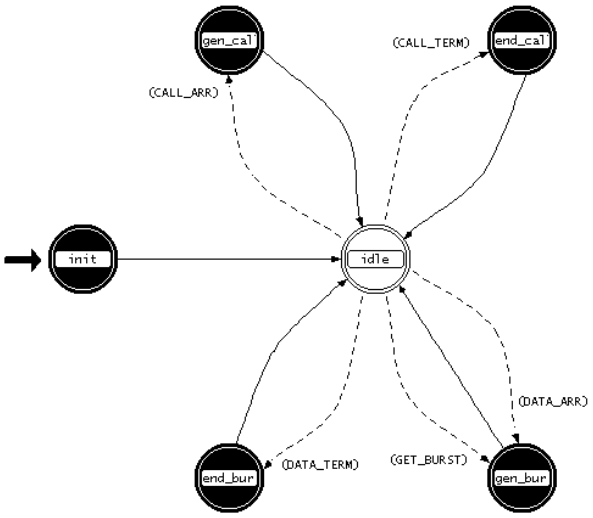

Figure 6. Bandwidth Server FSM Diagram

IV. OPNET Models, Simulations and Results

The OPNET Modeler software package 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.).

A. Models

The first approach used was to model the bandwidth server (medium) as a *M/G/1/∞* queue. The expected results are to provide verification of the source model statistical characteristics and an indication of bandwidth units required by an infinite capacity medium over time. The resultant OPNET process model for the bandwidth server is pictorially illustrated in Figure 6. It consists of six states. The 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 POTS or modem traffic and the other for data traffic. The remaining two states perform statistics updates and data burst queuing 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. 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.

We then improved the medium model to be a *G/G/1/N* queue, where *N* is a finite capacity (buffers) representing the number of BU's on the medium. This is done by introducing a "Number of Bandwidth Units" parameter which is adjustable at simulation time, and "blocked arrivals" statistics parameters, indicating arrivals (by type, voice/modem/data) that would have been rejected by the medium. The queuing mechanism for blocked data traffic arrivals was also added. This new model allows us to keep statistics involving blocking and delay characteristics while maintaining data indicating the desired bandwidth unit requirements. Some of the statistics kept in the BU server process model are illustrated in Table 1. Individual statistics for the voice source model consist of the number of arrivals generated, call length, offered load, and total offered load, both in Erlangs. Additionally, we take advantage of OPNET's subqueue statistics to acquire data specific to data traffic, such as the size of the queue and the queuing delay.

The voice traffic generator (POTS) source process model consists of an initialization stage and a generator stage shown in Figure 7. The initialization stage prepares variables for use in simulations. Simulation variables 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).

Table 1 - OPNET BU Server Process Model Statistics

Statistic	Description
Requested Bandwidth Units	BU's requested by either POTS or Modem sources
Actual Bandwidth Units	BU's limited by the number of BU's parameter
Modems In Use	
Call Length	POTS or Modem call length
Burst Length	Data Burst length
Blocked Arrivals	Any POTS or Modem blocked arrival
POTS Blocked Arrivals	
Modem Blocked Arrivals	
Data Arrivals	
Blocked Data Bursts	Any blocked data burst
Data Burst Collisions	Blocked data burst due to collision (burst in progress)

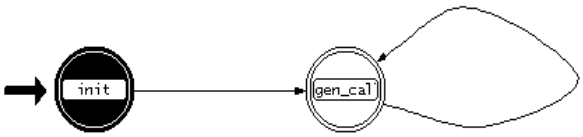

Figure 7. Voice Source Model FSM Diagram

The modem source process model ends up being a near copy of the POTS source. This model also consists of an initialization stage and a generator stage as shown in Figure 7. The initialization stage prepares variables for use in simulations. Simulation variables include arrival distribution (Poisson), arrival rate, hold time distribution (exponential or weibull), mean hold time, number of sources, and a "birth-death" indicator. The birth-death indicator is used with the number of sources and number of bandwidth units parameters to calculate the state probabilities and the average number of sources in the system as developed in section III. The modem call generator has the same individual statistics as that of the voice call generator.

OPNET contains a variable bit rate (VBR) traffic generator consisting of a parent process that creates "child" processes. Each child process is created according to a probability distribution determined at compile time. Within each child process, the length of the child process, packet arrival, and packet length pdf's are all individually configurable, also at compile time and are shown in Table 2. 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 root and child process model finite state machines are illustrated in Figure 8 and Figure 9.

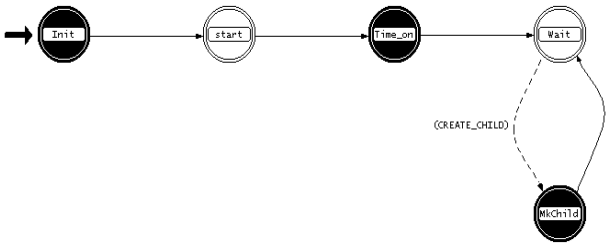

Figure 8. VBR Root Process FSM Diagram

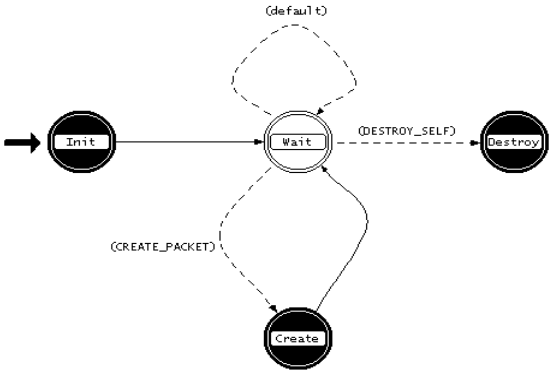

Figure 9. VBR Child Process FSM Diagram

The VBR root process initializes itself, then begins to generate child processes according to simulation parameters determined by the user at compile time. The parameters in Table 2 are used by each child process created by the root process. Each child process then acts independently as a function of the various probability distributions.

Based on the Poisson call generators (voice and modem), the VBR data burst generator and the bandwidth server process model, the bandwidth allocation OPNET node model was created as shown in Figure 10. This node is required as a fundamental building block in order to perform OPNET simulations.

Table 2 - VBR Root Process Model Attributes

Parameter	Description
Start Time	When to begin creating child processes
Packet Format	Set to 'NONE' (user can choose any desired format)
Child Interarrival pdf	Child arrival probability distribution
Child Interarrival arg1	Distribution argument (distribution dependent)
Child Interarrival arg2	Distribution argument (distribution dependent)
Child Duration pdf	Child duration probability distribution
Child Duration arg1	Distribution argument (distribution dependent)
Child Duration arg 2	Distribution argument (distribution dependent)
Packet Interarrival pdf	Packet arrival probability distribution
Packet Interarrival arg1	Distribution argument (distribution dependent)
Packet Interarrival arg2	Distribution argument (distribution dependent)
Packet Size pdf	Packet size probability distribution
Packet Size arg1	Distribution argument (distribution dependent)
Packet Size arg2	Distribution argument (distribution dependent)

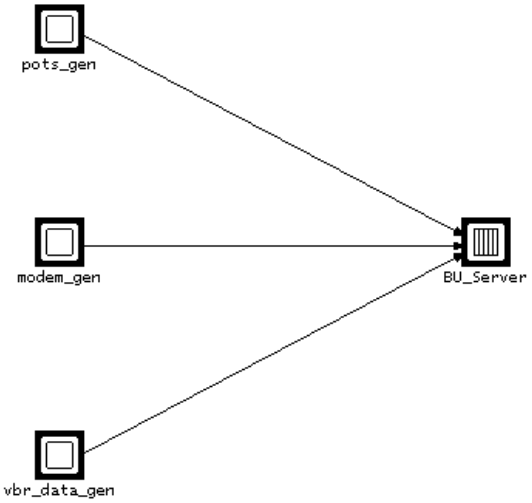

Figure 10. Bandwidth Allocation Node Model

B. Model Verification Simulations

Since the desired end result is to model and simulate a system containing multiple sources of different traffic types trying to access finite bandwidth, we shall establish "deployment rules" to limit the number of sources and their loads presented to the medium. For example, using the bandwidth capacity of that described in section I, (30 DS0's) we could adequately support roughly 135 telephone lines, each providing a load of 2 calls per hour with a mean call hold time of 4 minutes, with a probability of blocking of 0.5% (0.005). 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.

For this study we will consider three "deployment scenarios", based on concentrating the sources on a 30 DS0 medium by factors of two, three, and four. In terms of telephone lines (also used for modem data access) this represents 60, 90, and 120 lines (sources) respectively. Since the benchmark described in the preceding paragraph represents a slightly higher than four-to-one concentration of the medium, we expect a simulation of 120 voice only Poisson sources to require roughly 15 BU's on average.

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 more complicated. Refer to Figure 2. Note that at each "household" there is a "black box" which represents the interface between voice (data) appliances and the medium. The economics of developing and deploying these devices is such that it is impractical to deploy a device

with a single voice line interface. In fact, rarely, if ever do telephone service providers run a single twisted pair (one line) to residential households. The norm is two pairs. These factors help produce 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.

- Simulation run time was three hours (simulation time) in order to allow the generator processes to stabilize.

Voice traffic only simulations were performed using the model described in section III. The deployment rules described above were observed. Ten simulations were run, adjusting the seed value for OPNETs random number generator prior to each run. A sample of typical results is presented in graphical form in Figure 11 showing the BU's used, the average BU's used, the service time (length), and average service time. In the 120 POTS voice source case, the results indeed show the average number of bandwidth units required at about fifteen, as expected.

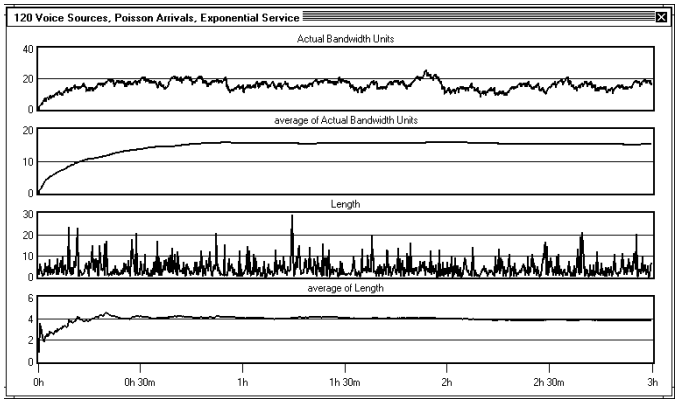

Figure 11. Voice call generator simulation (120 sources)

Modem only **s**imulations were also performed using the model described in section III. The deployment rules described above were observed. 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 OPNETs random number generator prior to each run. The results are presented in Figure 12 and Figure 13 showing the BU's used, the average BU's used, the service time (length), average service time. In the 40 analog modem source case, the results indeed show a marked difference between the simple Poisson process (Figure 12) and the birth-death process (Figure 13).

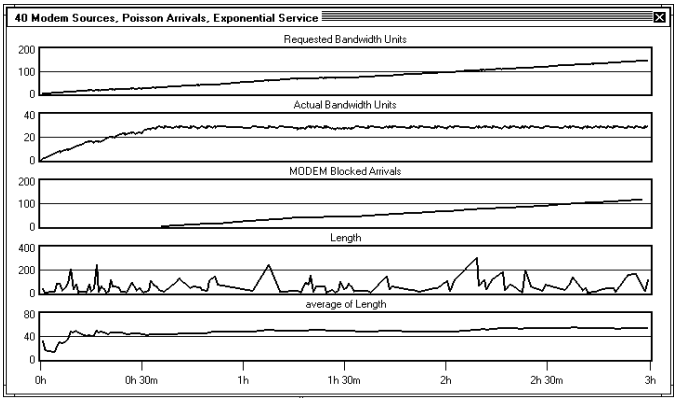

Figure 12. Modem Call Generator Simulation (40 Sources)

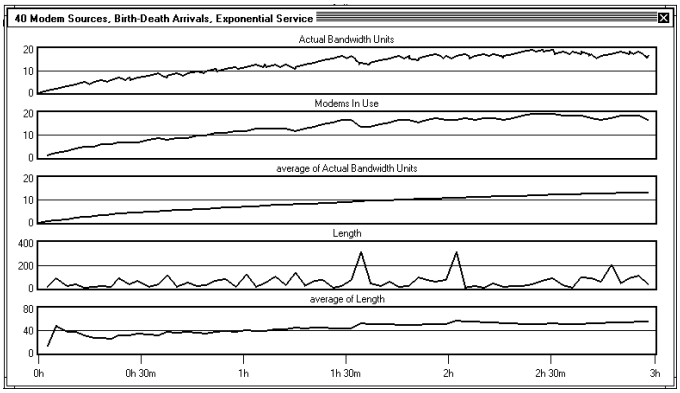

Figure 13. Modem Call Generator Simulation (40 Birth-Death Sources)

Clearly Figure 12 shows that it may be inappropriate to model analog modem traffic in the same fashion as we model voice call traffic. With the birth-death model for modem traffic enabled, we observe that the arrival behavior is limited to roughly one half of the total number of sources over the course of the simulation (Figure 13). This is consistent with equation 8.

Lastly, we illustrate a typical simulation of 40 modem sources using the birth-death model for arrivals (Figure 14). This time however we enable OPNET weibull distribution, using a shape parameter of 0.4, in an effort to produce a heavier tail to the right of the mean, giving us more modem calls in excess of 55 minutes.

Figure 14 shows how we can influence the service time distribution, and hence the mean, for analog modem traffic by using the weibull distribution. As the shape parameter is reduced, we see more arrivals whose service time (length) exceeds the mean, thus producing a "heavy" tail. In simulations that mix all traffic types, we will consider a single value for the shape parameter.

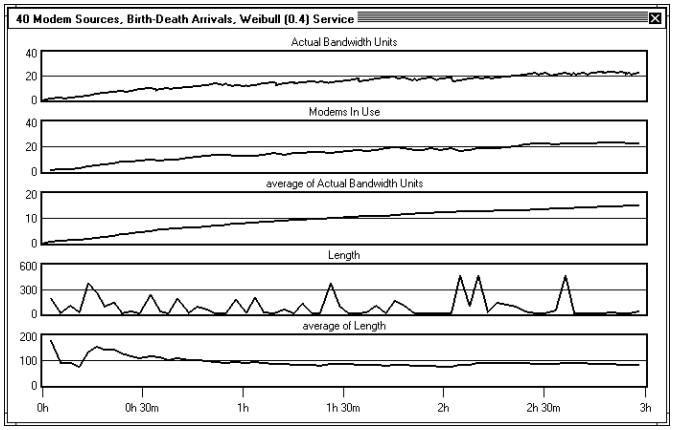

Figure 14. Modem Call Generator Simulation, 40 Birth-Death Sources, Weibull Service

Our verification of the VBR traffic generator consisted of disabling both voice and analog modem arrivals, then observing the behavior of the VBR generator given several scenarios. One scenario was to produce arrivals of child processes at a constant rate, with constant child duration shorter than the child process interarrival time. By stipulating a constant packet size and interarrival rate, we could then produce a constant stream of packets with no expected blocking or overlap. This was important in order to verify the basic functionality of both the generator and the BU server. Subsequent scenarios consisted of manipulating the various process model attributes described in Table 2 per the results and subsequent theorems presented in [18].

One example would be of a packet train with an arrival characteristic that is Poisson, with a fairly short duration and exponentially distributed times or constant times 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. Figure 15 illustrates such a data only simulation (i.e. no POTS or modem traffic) whose parameters were chosen such that the desired result might mimic a single computer user "browsing" the web. In other words, each child process acts like a "mini-session", with arrivals being Poisson distributed with λ = 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 15, the individual graphs show that the traffic is indeed bursty when observed on several time scales, however, the 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 produces a heavy tail such that there will be more arrivals with longer ON periods that represent up to approximately 1M byte of data offered to the medium. This behavior is illustrated in Figure 16.

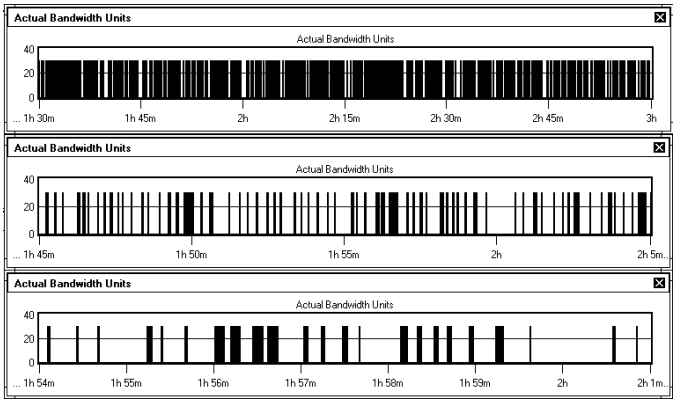

Figure 15. Data (VBR) Traffic Generator Simulation (One Source)

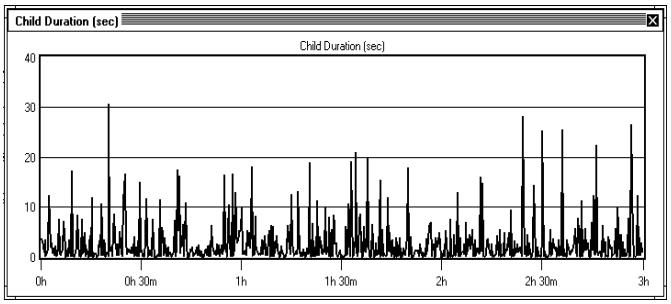

Figure 16. Child Duration (ON Period)

C. Mixed Traffic Simulation Results

For the traffic source combination simulations, three generic concentration ratios were considered, 2:1, 3:1, and 4:1. Within each generic concentration ratio four different 'mixes' were simulated, for a total of twelve different scenarios. 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 where the traffic mix was 40 voice, 20 modem, and zero data users. The bandwidth unit allocation and call length results are illustrated in Figure 17.

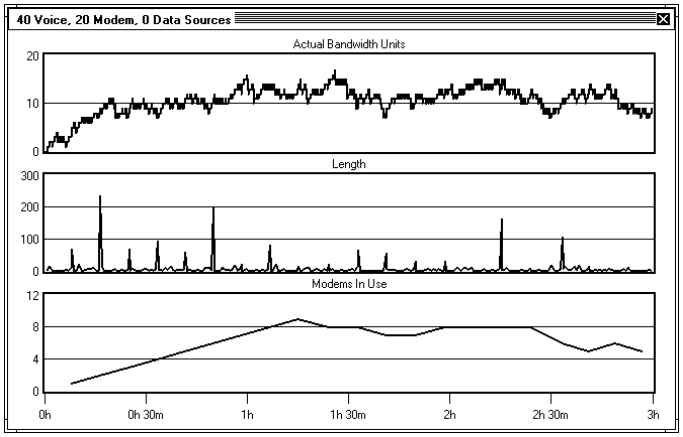

Figure 17. Voice/Modem Simulation Results, 2:1 Concentration Ratio

The traffic source mix then was altered to introduce a single data user. To adhere to our deployment rules the mix became 41 voice, 19 modem, and one data user. 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. Figure 18 illustrates the VBR load offered to the medium for one, ten, and twenty data users, with all other process model parameters remaining constant.

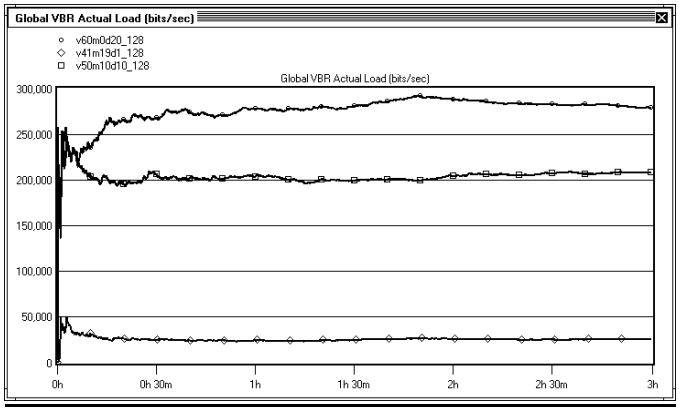

Figure 18. Offered Load (1, 10, And 20 VBR Data Users)

In the 2:1 concentration ratio case we are lastly interested in examining any substantial queuing delay that results from the mixed traffic, which could be perceived as degraded quality of service. Clearly, the 2:1 concentration ratio presents little, if any delay for the data user, as shown in Figure 19. The only difference of note is that as the number of data users increases the small queuing delay is present all the time instead of occasionally (top trace).

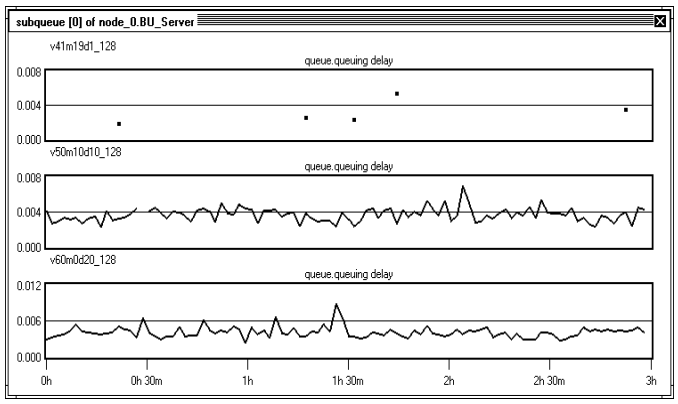

Figure 19. Queuing Delay 2:1 Concentration Ratio

The 3:1 concentration ratio scenarios begin to exhibit more interesting behavior. In the zero data user case the simulation implies that the medium is potentially in danger of becoming congested, as shown in Figure 20. Using different seed values in this scenario produced only a minor variation (1 or 2 bandwidth units) at the end of the simulation. Figure 21 implies that a single data user in a 3:1 concentration ratio could be at risk for a perceived degradation in quality of service, as the queuing delay exceeds one second, at least an order of magnitude greater than when more data users replace modem users.

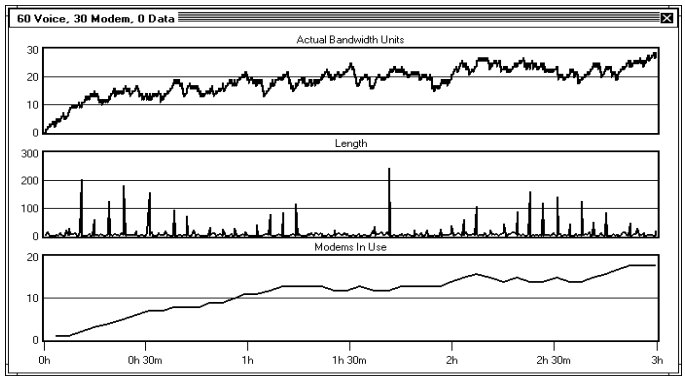

Figure 20. Voice/Modem Simulation Results, 3:1 Concentration Ratio

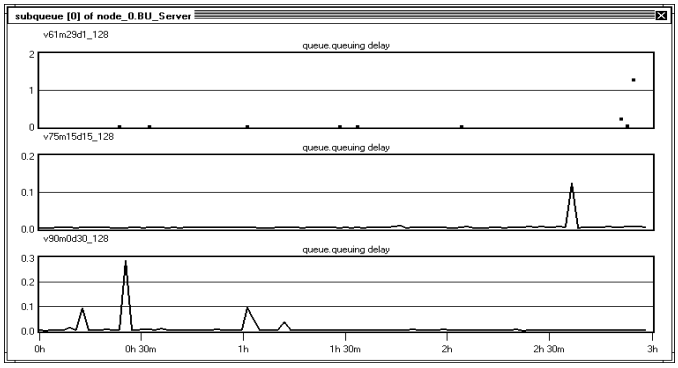

Figure 21. Queuing Delay 3:1 Concentration Ratio

Finally, the 4:1 concentration ratio behavior begins to show not only risk of QoS degradation for data users, but also for voice and modem users as shown in Figure 22. The requested bandwidth significantly exceeds the available bandwidth before the simulation is half completed.

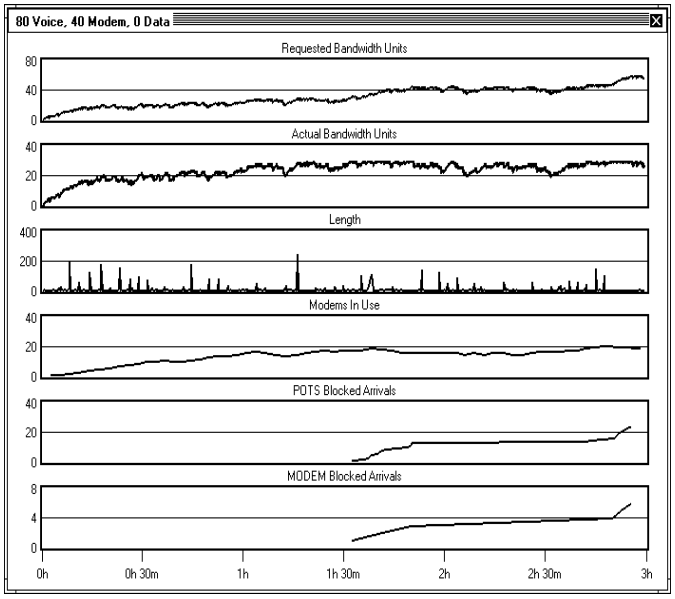

Figure 22. Voice/Modem Simulation Results, 4:1 Concentration Ratio

At this point, with zero data users, the medium is becoming saturated, and blocking of modem and voice traffic is observed. Recall that in Figure 11, 120 voice only users produced an average bandwidth unit requirement of 15, or 1/2 of the capacity of the medium. Peak usage was about 24 bandwidth units, which matches values that one would see in a standard call table. 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.

As modem users are replaced with data users, the resulting queuing delay behavior resembles that of the behavior observed in Figure 21. Figure 23 shows that as more data users replace modem users in this concentration ratio scenario, the queuing delay is drastically reduced, but not eliminated. In the cases where few data users are in the system (top two traces of Figure 23), the queuing delays are now on the order of a minute or more. The quality of service in these cases would most likely be perceived as severely degraded, if not unacceptable. Even in the case where no modem users are in the system, there are random queuing delays on the order of one second, again orders of magnitude higher that those seen under reduced concentration ratios.

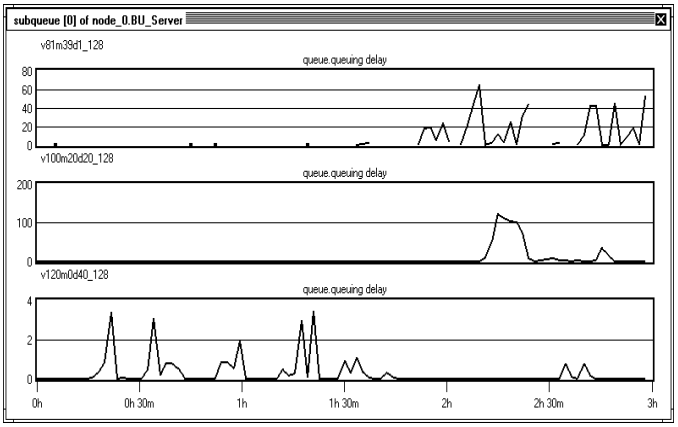

Figure 23. Queuing Delay 4:1 Concentration Ratio

V. Summary and Future Work

One of the main purposes of this work was to attempt to gain an intuition as to the ramifications of servicing mixed traffic types (voice and data) on a prioritized, shared medium. The notion of voice traffic is further complicated by the use of analog modems for access to the data network, thus rendering traditional modeling techniques to potentially underestimate bandwidth unit resource requirements during the busy hour. While new mechanisms to allow higher speed data access are being developed to take advantage of unused bandwidth, the likelyhood of an instantaneous mass demise of modem access to the internet in the near term remains small. This work indicates that there could be undesirable quality of service 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 a more sophisticated simulation model 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.

References

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

[2] Bolotin V. Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis. *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 3, April 1994.

[3] 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.

[4] De Serres, Y. and Mason, L. G. A Multiserver Queue with Narrow- and Wide-Band Customers and Wide-Band Restricted Access. *IEEE Transactions on Communications*, Vol. 36, No. 6, pp. 675-684, June 1988.

[5] Duffy, D. E., McIntosh A., Rosenstein M., and Willinger W. Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks. *IEEE Journal on Selected Areas in Communications*, Vol. 12, No. 3, April 1994.

[6] Dutta-Roy, A. A Second Wind for Wiring. *IEEE Spectrum* (pg. 52 - 60), The Institute of Electrical and Electronics Engineers, Inc., New York, NY, September 1999.

[7] Fluss, H. S. Effective Performance of Shared Bandwidth Data Channels in Hybrid Fiber/Coax Networks. *Presented at the Society of Cable Telephony Engineers (SCTE) '99*, January 1999.

[8] Fowler, H. J., and Leland, W. E. Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management. *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, September 1991.

[9] Donald Gross and Carl M. Harris. *Fundamentals of Queueing Theory*, Third Edition; John Wiley & Sons, Inc., New York. 1998.

[10] Jain, R. and Routhier, S. A. Packet Trains - Measurements and a New Model for Computer Network Traffic. *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 6, September 1986, pp. 986-995.

[11] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson, D.V. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, February 1994, pp. 1-15.

[12] Ngo, B. and Lee, H. Queuing Analysis of Traffic Access Control Strategies *IEEE Journal on selected areas in Communications*, Vol. 9, No. 7, September 1991.

[13] Paxson, V., and Floyd, S. Wide-Area Traffic: The Failure of Poisson Modeling *Proceedings of ACM SigComm* '94, London, August/September 1994, pp. 257-268.

[14] Paxson, V. Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network *Traffic Computer Communication Review*, Vol. 27, No. 5, October 1997, pp. 5-18.

[15] Vandalore, B., Babic, G., and Jain, R. Analysis and Modeling in Modern Data Communications Networks. *Submitted to the Applied Tele-communications Symposium*, 1999.

[16] Wang, X. and Chang, S.C. On the Performance Study of Several Access Control Strategies in ISDN. *Conference Record IEEE ICC 88*, 1988, Vol. 1, pp. 934-938.

[17] Willinger, W., Taqqu, M.S., Sherman, R., and Wilson, D.V. Self-Similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, 1997, pp. 71-86.

[18] Willinger, W., Paxson, V., and Taqqu, M.S. Self-Similarity and Heavy Tails: Structural Modeling of Network Traffic. *A Practical Guide To Heavy Tails: Statistical Techniques and Applications*, Adler, R., Feldman, R. and Taqqu, M.S., editors, Birkhauser, Boston, 1998.