2009

Understanding the Impact of Self-Similarity on Network Utilization in Multimedia Networks

Han Li

Minnesota State University - Moorhead, li@mnstate.edu

Ashish Gupta

Minnesota State University Moorhead, gupta@mnstate.edu

Understanding the Impact of Self-Similarity on Network Utilization in Multimedia Networks

Han Li
School of Business
Minnesota State University Moorhead
li@mnstate.edu

Ashish Gupta
School of Business
Minnesota State University Moorhead
gupta@mnstate.edu

ABSTRACT

Satisfying Quality of Service (QoS) of multimedia traffic has been gaining more attentions recently. Data, voice and video share the same communication channel but have significantly different traffic characteristics and performance requirements in terms of packet loss, delay, etc. Self-similarity also imposes some new challenges for network dimensioning as more buffer and bandwidth are expected to accommodate the extra burstness of self-similarity of multimedia application. However, previous studies did not investigate how much extra bandwidth is needed to meet QoS requirements of self-similar multimedia traffic. In this study, we model delay and loss requirements of interactive voice as the constraints to simulate the maximum allowable link loads. Results of this study suggest that under a self-similar input traffic, most of the bandwidth resources of slow links will be wasted to accommodate the traffic burstness if quality requirements of voice traffic are to be satisfied in a best-effort manner. For fast links (>10Mbps), self-similarity seems to have little impact on the maximum allowable load, which is above 82%.

Keywords

Self-Similarity, telecommunications, networks, multimedia traffic, utilization.

INTRODUCTION

Recent studies have given convincing evidence that multi-media traffic is self-similar because of the integration of data, voice, and video traffic (Sahinoglu and Tekinay, 1999). "Specially, "bursty" traffic patterns generated by data sources and variable bit rate (VBR) real-time applications such as compressed video and audio tend to exhibit certain degrees of correlations between arrivals, and show long-range dependences in time (self-similar)." Self-similar traffic behaves differently from network traffic traditionally modeled as Markovian process. Multiplexing of different traffic streams will smooth out peaks in traffic modeled by Markovian processes but cannot smooth out burst ness of self-similar traffic. The additional burst ness of the self-similar traffic can aggravate traffic delay and traffic loss. It means more bandwidth and buffer will be needed to achieve designed network delay and loss. Two important concerns are how to guarantee QoS for multimedia application in the context of self-similar traffic and how to incorporate self-similar behavior into the bandwidth design of multimedia traffic network and maximize network utilization.

The primary focus of this study is to simulate the load point above which network will fail to satisfy the delay and loss requirements of multimedia traffic. Specifically, this paper investigated the maximum allowable load under delay and loss constraints imposed by interactive voice traffic in self-similar traffic context since interactive voice has the strictest QoS requirements. Availability of such data can help network operator to schedule bandwidth needed.

The following sections first describe self-similarity in network traffic and its impacts on network dimensioning based on previous studies. This paper then analyzes performance requirements of interactive voice of multimedia application. QoS restrictions imposed by interactive voice are then adopted to simulate the allowable link load of several commonly used links such as 384KB, T1, fractions of T3, T3 in a best effort network. Finally, simulated results are analyzed and summarized in conclusion section.

SELF-SIMILARITY

Self-similarity means the object stays the same when viewed across several scales of time or space. In the context of packet network, self-similarity means network traffic looks similar when viewed across several time scales. Only recently did self-similar processes arouse tremendous interest in telecommunication field although they have long been applied in hydrogeology and geophysics fields. Leland, et al (1994) first applied self-similarity concept in analyzing measured Ethernet traffic. They gathered Ethernet traffic at the Bellcore networks from 1989 to 1992. The measured Ethernet traffic exhibits some extended temporal correlation or long range dependency (LRD). It seems to look similar at different time scales from 0.01 s to 10 s and has scale invariant burst ness (Figure 1 b, c, and d). In comparison, Markovian processes (Poisson model) fail to reproduce such scale-invariant burst ness of real world traffic. For traffic modeled from Markovian processes, burstiness will be smoothed out over large time scale (Figure 1 b' and c'), which suggests only modest-sized buffers are required since buffer can be cleared out over a long time period. Therefore, Markovian processes will overestimate link utilization needed to satisfy certain QoS requirements. Besides Ethernet traffic, self-similarity has been found in other applications such as ISDN, WWW traffic, FTP, FTP and telnet.

Self-similarity of network traffic has been shown to be second-order self-similar and can be described by several functions. Let X(t) be a second-order stationary time series where t = 0, 1, 2, ... Then, m-aggregated time series $X^{(m)}$ can be defined as $X^{(m)} = \{X_k^{(m)}, k=0, 1, 2, \ldots\}$ by summing the second-order stationary time series over non-overlapping,

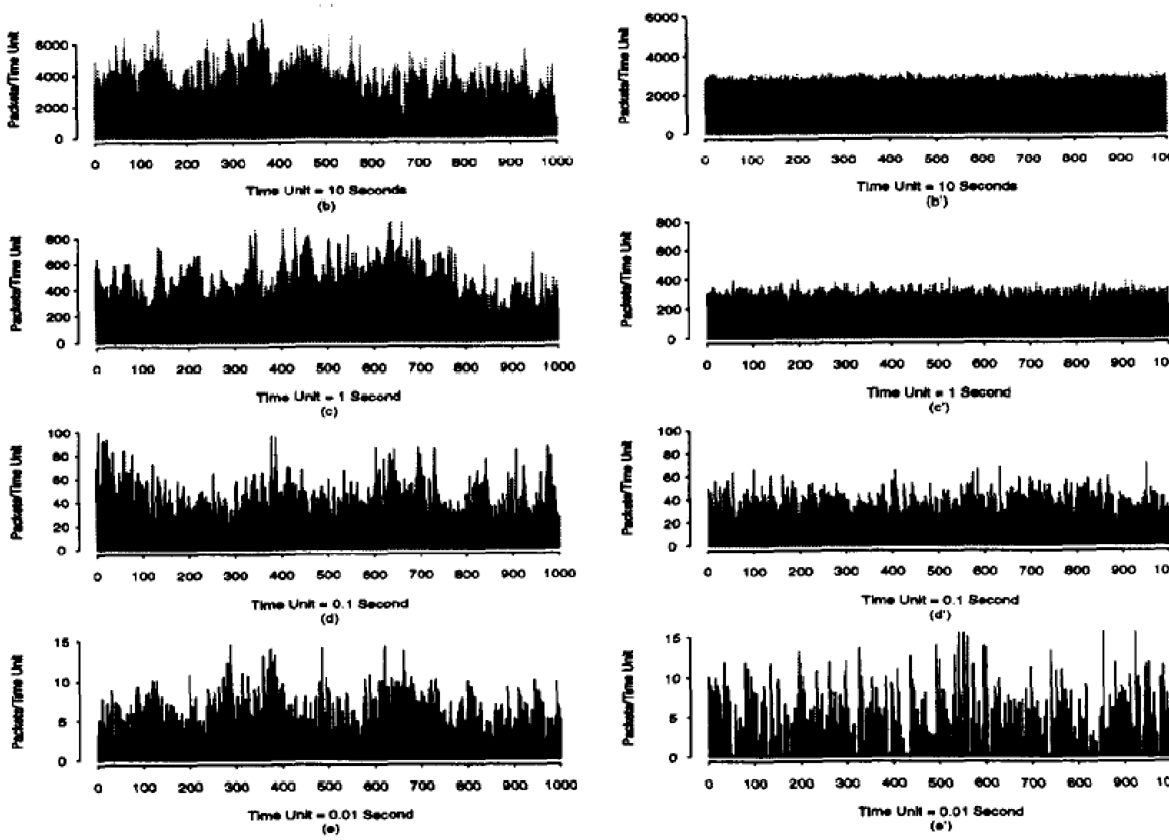

Figure 1. Comparison of Actual Ethernet Traffic (packets per time unit) and Synthetic Traffic from Poisson model at Different Time Scales from 0.01s to 10s (From Leland, et al. 1994).

adjacent blocks of size m (Stallings, 1998). This may be expressed as

$$\mathrm{Var}(X^{(m)}) \sim \frac{\mathrm{Var}(X)}{m^{\beta}} \quad \text{variance}$$

$$R_{X(m)}(k) \rightarrow R_X(k) \text{ as } m \rightarrow \infty \quad \text{autocorrelation}$$

Parameter β can be related to Hurst parameter of a self-similar process as $H=1-k/2$. H is a measure of the degree of self-similarity. The more bursty a process, the larger the H value. The process is self-similar if H is from 0.5 to 1. From the variance formula, we can see variance of a self-similar process decays slowly with the increase of m or the increase of time scale. Autocorrelation for a aggregated self-similar process also does not go to zero as $m \rightarrow \infty$ as opposed to traditional stochastic process in which autocorrelation function will be equal to zero when $m \rightarrow \infty$.

Many studies have shown that self-similarity of network traffic has a negative impact on network performance, such as longer queuing delay and higher packet loss. Therefore, more buffers and bandwidth are needed than predicted by traditional queuing theory. More complexities are involved in providing QoS to real-time multimedia service due to the strict requirements on both delay and loss from its interactive voice component. An exponential trade-off relationship has been observed between queuing delay and packet loss for traffic with high degree of self-similarity (Park, et al, 1997). Reduced packet loss achieved from increasing buffer size alone is accompanied by significant increase in queuing delay, which is detrimental to the quality of interactive voice. Also, the increase in buffer size does not help too much on packet loss for high-degrees of long-range dependence (Higher values of H) (Sahinoglu and Tekinay, 1999). Therefore, to ensure quality of multimedia traffic in a best-effort network, increasing bandwidth seems to be a more viable approach, which improves network performance with respect to both packet loss and delay.

PERFORMANCE REQUIREMENTS FOR INTERACTIVE VOICE

Internet, as the best-effort network, was originally designed to haul data traffic only, such as email, ftp, etc. Internet traffic is becoming more and more diversified as a result of multimedia applications such as interactive audio and video. Among these different types of Internet traffic, interactive voice traffic presents most of the challenges toward the Internet. Interactive voice traffic differs from best-effort data traffic by putting more constraints on the following parameters.

Latency is time that it takes a packet to traverse the network end-to-end. Large latency can generate echo and talker overlap problems. Generally, it is widely accepted that one way end-to-end delay of voice traffic should be less than 150 ms. There are multiple sources that can cause latency of a voice packet. The major ones include packetization delay at the source, propagation delay and queuing delay incurred in the network during transmission and playback buffer delay at the destination or receiver. Packetization delay is dependent upon type of CODEC used. For example, G729A will need 30ms to generate a voice packet consisting of three voice frames. The packetization delay for nominal operation of any gateway unit should not exceed 30ms (Juniper Networks, 2001). 30ms is a sound approximation for typical packetization delay. When voice packets are transmitted on the network, propagation delay, transmission delay and queuing delay are incurred at each hop. Propagation delay is a function of the traveling distance of the packet. The average propagation delay for intra-continental calls is around 30ms (Karam and Tobagi, 2001). Transmission delay is determined by the packet size and link speed, which is negligible most of the time. Queuing delay during transmission is mainly affected by the link bandwidth and link load. In this study, transmission delay is considered together with queuing delay during transmission. At the receiver end, packets are stored in a playback buffer, which causes playback delay. Playback buffer delay generally is assumed to be equal to the maximum queuing delay incurred in the network (Karam and Tobagi, 2001). From above, the maximum total queuing delay during transmission will be around 45ms. According to Chuah and Katz (1999), the number of hops is typically around 8-12 for packets traveling between west coast and east coast. If queuing delay is assumed to be the same at each hop, the maximum acceptable queuing delay at each hop should be around 5 ms.

Jitter is the variation of delay over a period of time. Jitter is mainly generated by variations in queuing delay in response to short-term changes in network traffic loads. Jitter has tremendous detrimental effect on voice quality. To minimize the effect of Jitter, play-out buffers are used in receiver to store a packet stream. In this study, Jitter is not used as QoS constraints to determine the maximum allowable link load.

Packet loss reflects the probability for packets to be discarded during network congestion. Interactive voice application is much less tolerant to packet loss. On the one hand, real-time applications are based on UDP protocol, which does not have retransmission function in case of packet loss. On the other hand, even if TCP were used for interactive voice application, retransmitted voice packets would arrive too late and miss the playout time at the receiver side. There are different views about the acceptable loss rate of voice packet. The white paper of Juniper Network held that “As long as the amount of packet loss is less than five percent for the total number of calls, the quality generally is not adversely affected.” The study by Chuah and Katz (1997) suggested that voice quality would be intolerable when loss rates exceed 3%. To be conservative, 3% is used to represent the maximum loss rate in this study.

SELF-SIMILAR TRAFFIC SIMULATION MODEL

OPNET is used as simulation tool in this study. OPNET has a build-in Raw Packet Generator (RPG) which can be used to generate self-similar traffic. RPG utilizes several Fractal Point Process (FPP’s) models proposed by Ryu and Lowen (1998). The default Superposition of Fractal Renewal Processes (Sup-FRP) model is used in this simulation. The Sup-FRP model is constructed as the superposition of M i.i.d. fractal renewal point processes (FRP), where each FRP is completely characterized by the following power-law probability density function (PDF) for inter-arrival times:

$$p(t) = \begin{cases} rA^{-1}e^{-rt/A} & 0 \le t \le A \\ re^{-r}A^{r}t^{-(r+1)} & t > A \end{cases}$$

with 1< r < 2. The Sup-FRP has three parameters (r, A, M). r is a fractal exponent, A is a cutoff parameter, and M is the number of FRPs superimposed. These three parameters can be mapped into the three input parameters (λ, H, T_0) of RPG generator as follows:

$$H = (3-r)/2,$$
$$\lambda = Mr[1+(r-1)^{-1}e^{-r}]^{-1}/A,$$
$$T_0^{\alpha} = 2^{-1}r^{-2}e^{-r}(r-1)^{-1}(2-r)(3-r)[1+(r-1)e^{r}]^{2}A^{\alpha}$$

where $\alpha = 2 - r$. Of these three input parameters, λ is the average number of packets per second and T_0 is the Fractal Onset Time Scale (FOTS).

In this study, Hurst parameter is chosen to be 0.8, which was reported in WWW traffic (Crovella, and Bestavros, 1996). FOTS is the time scale from which self-similar behavior begins to appear. A smaller FOTS corresponds to a higher variance at the small time scales. In this study, FOTS is selected to 0.1 second.

The layout of the network is shown in Figure 2. The network consists two pairs of sources and destinations. Data and Data Destination nodes contain RPG above IP and Ethernet MAC layers. Data node generates self-similar data traffic which is received by Data Destination node. Voice node generates fixed-rate voice packets that are sent to Voice Destination.

The ratio of fixed rate voice traffic and self-similar data traffic is 1:99 in this study. These two sources of traffic are multiplexed in router A. According to the study by Xue and Yoo, the aggregated traffic will still be self-similar with Hurst parameter close to the largest one of the underlying traffic sources. Therefore, the Hurst parameter of aggregated traffic transmitted over the link between two routers should be close to that of data traffic, i.e. 0.8.

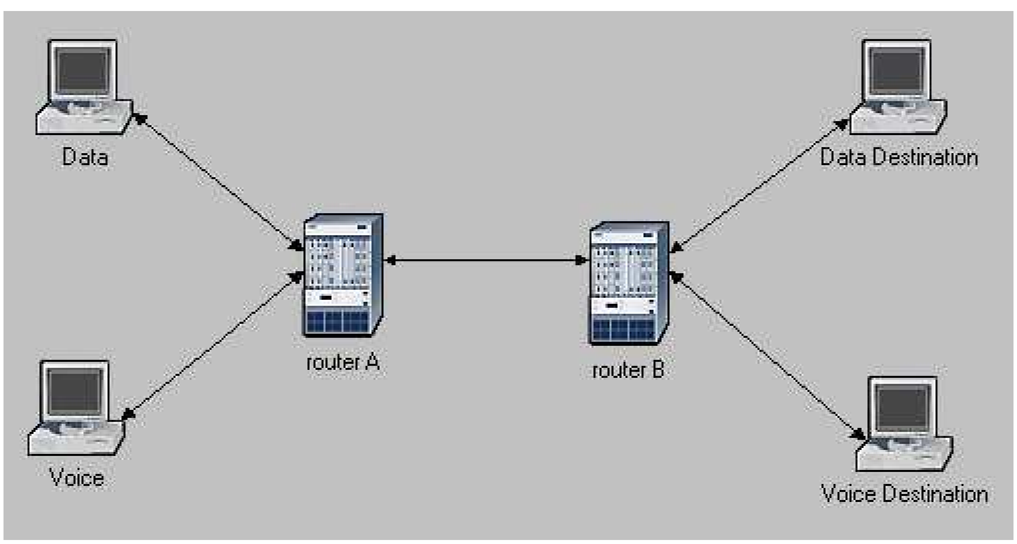

Figure 2. Network Layout

The packet size of self-similar traffic is chosen to be exponentially distributed with a mean of 280 bytes. The total packet size at the data link layer will be 300 bytes with 20 bytes for IP header. This is consistent with the measurement of background traffic from MCI's Internet backbone in 1997. The packet size for voice traffic is chosen to be 30 bytes, which can be used to store three G.729A compressed voice frames. The total packet size at the data link layer will be 58 bytes considering 8 bytes for UDP header and 20 bytes for IP header.

A number of commonly used link speeds are selected for the link between router A and router B, including 384 Kbps, T1 (1.544 Mbps), fractions of T3 (3T1, 6T1 and 9T1) and T3 (45 Mbps). The bandwidth of other links in the network is chosen to be 1Gbps so transmission delays at these links can be omitted. Router processing speeds are set to be 5,000,000 packets/second so processing delay is negligible in gathered end-to-end delay of voice traffic. Therefore, only queuing delay incurred from link between router A and B or one-hop queuing delay is reflected in end-to-end delay gathered at node "Voice Destination".

SIMULATION RESULTS

As analyzed in the section of performance requirements of interactive voice traffic, maximum acceptable one-hop queuing delay is around 5ms for voice packets carried over Internet or WAN. Voice packets are treated as lost packets if their end-to-end delay gathered at Voice Destination node exceeds 5ms. Thus, the maximum allowable link load can be determined from cumulative probability curves of end-to-end delay of voice packet by using 5ms delay and 3% loss rate.

Figure 3 shows the cumulative probability of packet end-to-end delay gathered at node "Voice Destination" under varying traffic load and link speeds. The black dot is used to represent the position where 97% of voice packets have an end-to-end delay less than 5ms and 3% of voice packets have an end-to-end delay above 5 ms. Therefore, curves to the right of this black dot reflect that corresponding link loads are too high to satisfy the performance requirement of voice traffic. Curve extending through this black dot corresponds to the maximum allowable link load for that link speed. In Figure 3(a), it can be seen that the curve of 5% load passes through the black dot. Therefore, maximum allowable link load is only 5% for 384Kbps link in order to satisfy the quality requirements of voice traffic. Apparently, the link will be severely underutilized and 95% of the bandwidth resources will be wasted to accommodate the burstness of similar traffic if voice quality is ensured. For T1 link, a 30 % maximum allowable link load can be identified from Figure 3 (b), which still suggests inefficient link utilization. The adverse impact of traffic self-similarity seems to be aggravated for slow links such as 383 Kbps and T1 links. Better link utilization is achieved when link speed is 4.632 (3T1) and above. Maximum allowable link loads is 66%, 82%, 87% and 95.5% for 4.632 Mbps (3T1), 9.264 Mbps (6T1), 13.896 Mbps (9T1) and 44.736 Mbps (T3) links respectively (Figure 3 (c) to (f)).

Figure 4 plots the different link speeds against their maximum allowable link loads. We can see that maximum allowable link load decreases quickly with the decrease of link speed when bandwidth of the link is less than 10Mbps. However, when link speed is above 10Mbps, maximum allowable link load is higher than 82% and decreases very slowly with the decrease of link speed. Therefore, efficient link utilization can be achieved when bandwidth of the link is higher than 10 Mbps. Self-similarity seems to have less negative impact on network performance with respect to QoS requirements of voice traffic when link bandwidth is above 10 Mbps. Assuming there is enough traffic to fill the link, high-bandwidth communication links (>10Mbps) should be deployed for multimedia traffic. This will be the feasible solution for network backbones.

At present, slow links are commonly deployed as access links and it may not be cost effective to increase their bandwidth above 10Mbps. At least in foreseeable future, the bandwidth of access network will still be bottleneck of the entire network and self-similarity will continue to aggravate the loss and delay of packets incurred at these slow links. These slow access links will be major determinant on the quality of multimedia traffic. Since bandwidth of access links may not be increased to accommodate the burstness of self-similar multimedia traffic, there exist more needs to implement priority queuing in boundary routers. QoS mechanisms will be crucial to increase the utilization of bandwidth resources of access link while at the same time satisfy the performance requirements of multimedia traffic. Voice and data should be treated differently with higher priority assigned to interactive voice traffic.

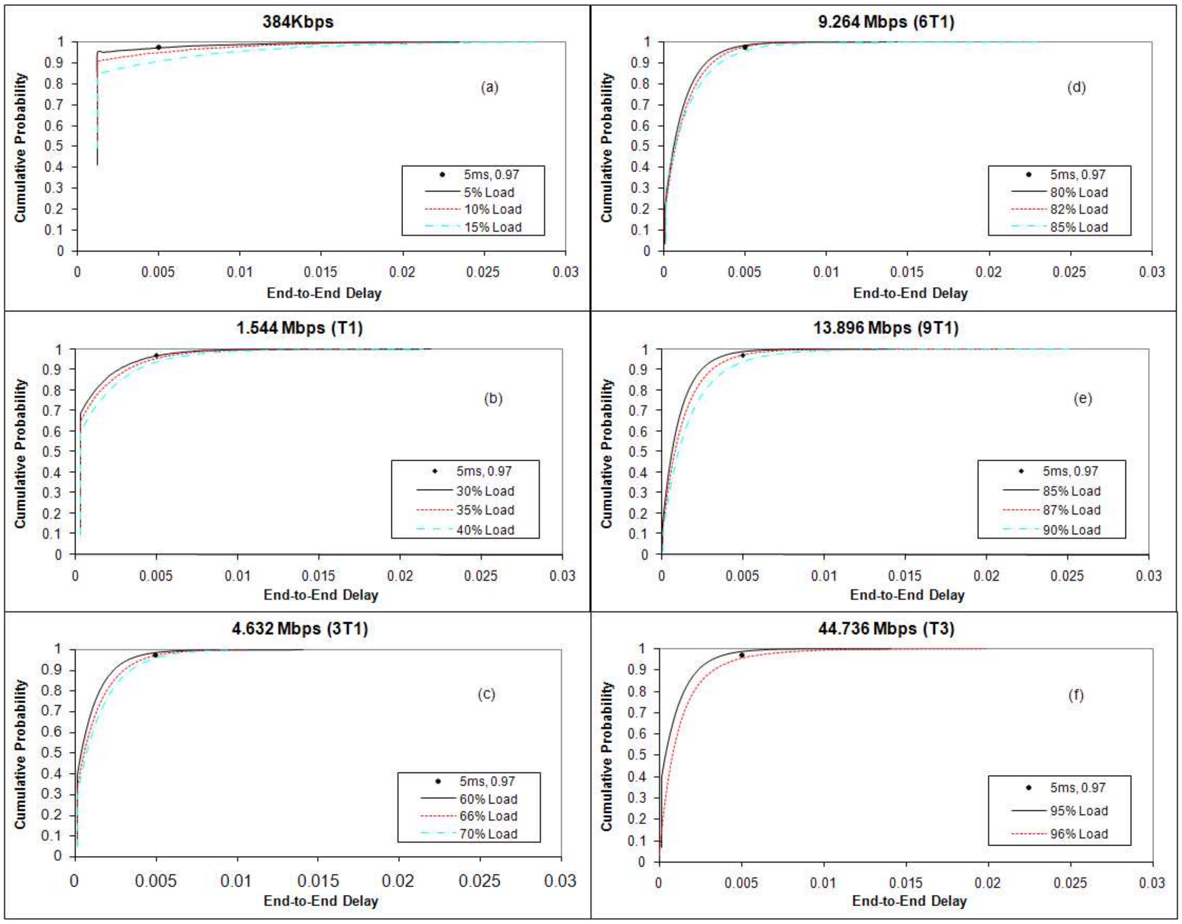

Figure 3. Cumulative Probability of End-to-End Delay of Voice Packet (The black dot represents the position where 97% of the voice packets have an end-to-end delay less than 5ms.)

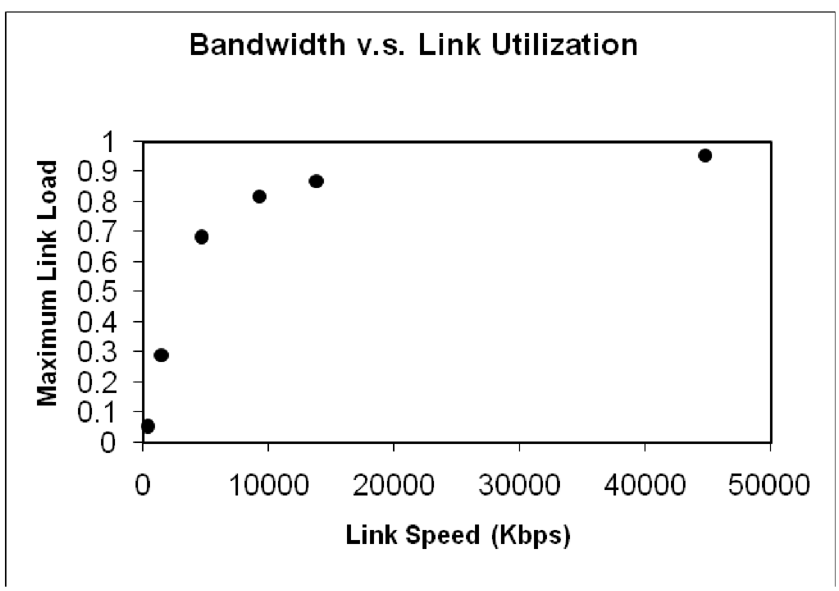

Figure 4. Link Speed Versus Maximum Allowable Link Loads.

CONCLUSIONS

In this paper, the QoS constraints of interactive voice traffic are utilized to determine the maximum allowable link load for self-similar multimedia traffic. Three major conclusions can be drawn from the study. First, when link bandwidth is above 10 Mbps, self-similarity seems to have less impact on network performance. The maximum allowable link load is above 82% for fast links without sacrificing latency and loss requirements of interactive voice traffic. Therefore, there is less need to capture self-similarity in traffic modeling for fast links. Second, for slow links, self-similarity seems to aggravate the loss and latency of voice packets. These slow links will be seriously underutilized if quality of interactive voice traffic is to be satisfied in a "best-effort" manner. Third, high bandwidth links (>10Mbps) should be deployed for multimedia traffic in backbone network while implementation of priority queuing at boundary routers might be a viable solution to increase the maximum allowable load for those slow access links.

REFERENCES

1. B. K. Ryu and S. B. Lowen, "Point process models for self-similar network traffic, with applications, " Stochastic Models, 14(3), 1998, 735-761
2. C. Chuah and R. Katz, "Network Provisioning and Resource Management for IP Telephony," Report No. UCB/CSD-99-1061, August 1999.
3. Fei Xue and S. J. Ben Yoo, The effect of aggregation of self-similar traffic, http://www.eng.auburn.edu/department/csse/classes/comp8700/papers/OPNETWORKPAPERS/
4. K. Park, G. Kim, and M. Crovella, "On the effect of traffic self-similarity on network performance, " Proc. SPIE Int'l. Conf. Perf. and Control of Network Sys., 1997, pp. 296-310.
5. M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," ACM SIGMETRICS, 1996, pp 160-169.
6. M.J. Karam and F.A. Tobagi, "Analysis of the delay and jitter of voice traffic over the Internet", IEEE INFOCOM 2001, vol 2, 2001, pp 824 –833.
7. Voice over IP solutions—White paper, Juniper Networks, 2001, http://www.juniper.net/techcenter/techpapers/200011-04.html.
8. W. Leland. M. Taqqu, W. Willinger, and D. Wilson, "On the self-similarity of Ethernet traffic (extended version)," IEEE/ACM Transactions on Networking, vol. 2, no. 1, Feb. 1994, pp. 1-15.
9. W. Stallings, High speed networks TCP/IP and ATM design principles, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1998
10. Z. Sahinoglu, S. Tekinay, "On Multimedia Networks: Self-similar Traffic and Multimedia Performance," IEEE Comm. Magazine, January 1999.