# Quantifying the Temporal Characteristics of Network Congestion Events for Multimedia Services

Victor S. Frost, *Fellow, IEEE*

***Abstract*—Effective quality-of-service (QoS) metrics must relate to end-user experience. For multimedia services these metrics should focus on phenomena that are observable by the end user. Once a congestion event occurs in the network it tends to persist, resulting in long bursts of consecutive packet loss. Such an event is observable to the network customer. There is a need to increase our understanding of the temporal characteristics of congestion. It has become increasingly apparent that the temporal characteristics of congestion events have the dominant effect on user-perceived QoS. A rigorous definition of the time between congestion events is given here, as well as an associated prediction methodology. The inter-congestion event time or the rate of congestion events per unit time provides a network quality metric that is easily understandable to network users and is conveniently predicted and measured. The contribution of this paper is the definition of a metric to characterize congestion events and development of an analytic methodology to predict the expected number of congestion events per unit time. The proposed methodology is evaluated for a variety of traffic models.**

***Index Terms*—Network congestion, prediction of quality-of-service (QoS), QoS.**

## I. Introduction

CONGESTION is a state of sustained network overload, where the demands for resources exceed the supply for an extended period of time. Thus, a congestion event will cause a significant number of packets to be lost consecutively. It has been observed that long bursts of packet losses are present in the Internet [15], [16], [22]. For example, measurements have shown [15] that less than 1% of all bursts of lost packets contain 50% of all losses. While quality-of-service (QoS) adaptive [17], [18] or network aware techniques [27], [28] can improve overall performance in some cases, long sequences of packet losses will cause a significant user-perceived impairment, or as discussed in [33] a “performance incident”. For example, when using the G723.1 recommendation for compressed voice over packet networks [19], only slight static and clipping result from one-to-four consecutive packet losses. However, longer bursts of lost packets will significantly degrade the quality of service (QoS) delivered to the user. Recent assessment of voice transmission over the Internet backbone [35] has demonstrated that voice quality is a function of time, which experiences rare intervals of unacceptable performance. Short bursts of lost packets are not significant while multiple consecutive losses typically cause noticeable impairments [20]. Loss concealment techniques attempt to mask the impact of a small number of losses and variations in delay. Further, developers have done a good job of designing applications (by making them adaptive or network aware) to mask short-lived performance problems in the network. Thus, packet/cell loss probability as a global measure of QoS does not necessarily provide a direct indication of user-perceived QoS [8]. Metrics such as delay and loss may have little direct meaning to end-users of rapidly changing multimedia applications. That is, knowledge of the specific coding and/or adaptive techniques is required to translate delay and loss into the user-perceived performance [24]. A congestion event as defined here will impact the user’s QoS independent of the specific coding mechanism or of attempts to mask and/or adaptively compensate for its effect.

It is has recently become understood that the temporal packet loss pattern has critical influence on user-perceived network performance, i.e., the temporal characteristics of the congestion episodes have the dominant effect on user-perceived QoS. Recent work in the Internet Engineering Task Force (IETF) on measurement-based temporal QoS metrics [21] reinforces the increasing importance of packet loss patterns. Models for the temporal dependence of packet loss have recently been derived from measured packet traces [22], [23]. These studies assume that a particular packet loss process models the transitions between different states: a no-loss state and a loss state, e.g., as in a Gilbert model. Parameters of the loss process (usually represented as a Markov process of different order) are estimated from measured packet traces. These models are used to characterize the nature of bursts of lost packets [22], or, as in [23] the length of good runs, i.e., congestion free intervals. In [21] two metrics are proposed, “loss distance” and “loss period.” Taken together, these characterized loss patterns seen by packet flows on the Internet. Previous work [21]–[23] is based on packet traces or knowledge of a packet loss models; such models are also derived from measurements of the loss process. Theoretical methodologies to transform network design parameters, i.e., traffic characteristics and loads, into predictions of the temporal dependence of congestion events have not yet been reported. A metric as defined here, is the time between congestion events that is amenable to analysis. This metric enables the prediction of phenomenon like the loss distance defined in [21]. The rate of congestion events is easily understood by end-users and is inde-

Manuscript received October 12, 2000; revised June 14, 2002. This work was supported in part by the National Science Foundation under Grant ANI-0125410. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation. The associate editor coordinating the review of this paper and approving it for publication was Dr. Chung-Sheng Li.

The author is with the Information and Telecommunications Technology Center, Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045-7612 USA (e-mail: frost@eecs.ku.edu, www.ittc.ku.edu/~frost).

Digital Object Identifier 10.1109/TMM.2003.813276

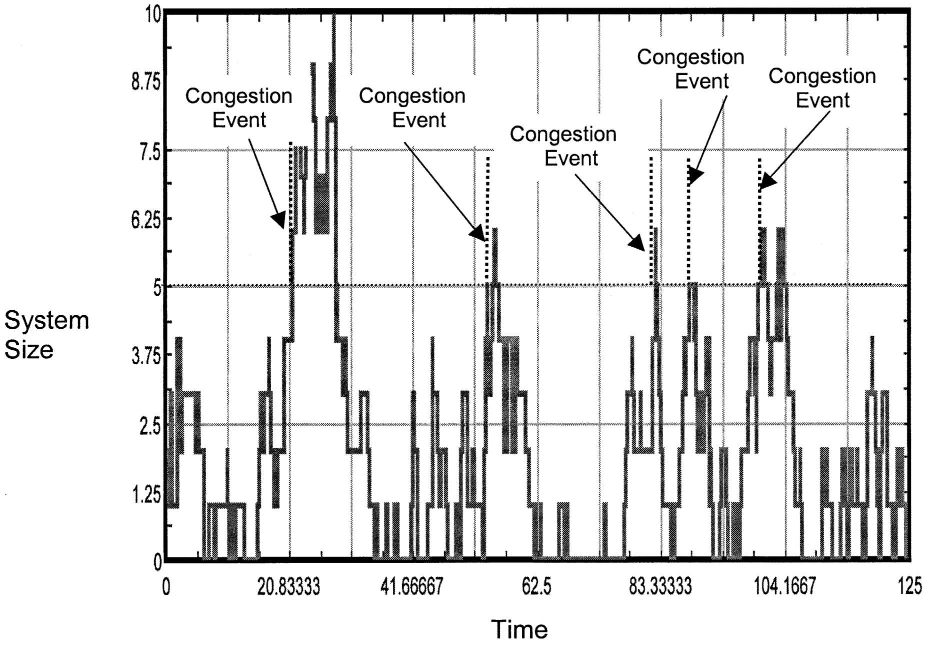


Fig. 1. Definition of Congestion Event with $b = 5$.

pendent of the specifics of the application, that is, it captures the network performance that is observable by network customers.

Note that QoS mechanisms have not been widely employed, both because of their complexity and their lack of relationship to user-perceived performance. Therefore, it is currently difficult to justify investment in complex QoS technologies while their value to the end customer is unknown. There have been cases where commercial networks have been over-provisioned, resulting in all users, independent of QoS requirement or priority, receiving the same service [33]. Based on these experiences network customers are reluctant to pay extra for QoS. Historically, the telephony network QoS metric, call blocking probability, was not only easily understood but also provided the foundation for network design. Since packet networks experience delay and packet loss in the presence of congestion this metric is not suitable for networks like the Internet. The rate of congestion events per unit time, as defined and analyzed here, provides a direct and easily understandable measure of network performance. A congestion event results in a long burst of lost packets, and thus always produces a noticeable degradation. The metric proposed here is constructed so that once a congestion event has occurred there will be a noticeable impact on the end-user application, thus the length of the congestion episode is considered of secondary importance. It is anticipated that a congestion event rate could become an important component in service level agreements between Internet service providers (ISPs) and carriers.

The contribution of his paper is the definition of a metric to characterize congestion events and development of an analytic methodology to predict the expected number of congestion events per unit time; the methodology is validated for several traffic models. Specifically, the metric has been evaluated in the context of "standard" traffic models, fixed (M/D/1), exponential (M/M/1) and hyperexponential (M/H(2)/1) services times, and for hyperexponential interarrival times (H(2)/M/1). Both a single node and a network of queues have been considered. The next section provides the definition for a congestion event. Section III presents a methodology to predict the time between congestion events. The developed technique is then applied to a single queue as well as to networks in Section IV. Conclusions are discussed in Section V.

## II. Definition of Congestion Events

Let $X(t)$ be a random process representing, at time $t$, the number of packets in the output buffer of a router including any packet in transmission. $X(t)$ can be viewed as the system state. Here, $X(t)$ is the state process in a queueing context. The $j$th congestion event is defined to occur at time $t_j$ if $t_j$ is the first time the process $X(t)$ reaches the level (state) $b$ following the end of the previous busy period containing a congestion event. A congestion episode starts with a congestion event and ends at the start of the next idle period. Thus, the length of a congestion episode is defined as the length of a busy period given the system state $b$ has been reached. Given this definition the process can reenter the state $b$ multiple times during one congestion episode.

Fig. 1 graphically illustrates successive congestion events for a congestion level of 5 ($b = 5$) for a simple M/M/1 system. Note that the system may reenter state $b$ multiple times during one congestion episode. Thus, the probability of a congestion event is not the same as the stationary state probability. Also for this metric the system must go idle between congestion events. The proposed approach makes the following assumptions: 1) the probability of reaching state $b$ is small (rare event), 2) reaching state $b$ would result in a large burst of lost packets, and 3) on average, the length of the busy period containing a congestion event, i.e., a congestion episode, is much less than the time between congestion events.

Two possible QoS metrics are suggested by the above discussion: the rate of congestion events, and duration of the congestion episodes. Run probabilities, as defined in [13], provide an alternate perspective on the duration of a congestion episode. In the context of an M/M/$\infty$ with many bursty sources, both the rate of congestion events and their duration have been considered in [8]–[11]. However, once a congestion event has occurred the damage has been done and the end-user perception of QoS affected; the length of the congestion episode is then of secondary importance.

The rate of congestion events can be used as a QoS metric for multimedia service in packet networks. This metric relates to the quality of real time applications like voice and video where congestion events, i.e., long bursts of lost packets, directly cause distortion. For audio services these impairments are "clicks" and "pops," while video transmission impairments cause color distortion, edge jerkiness, screen artifacts or loss of synchronization. While congestion events are not directly discussed, a related metric—transmission error free intervals—has been proposed as part of a set of standards for multimedia communications in [5]. Error free intervals for audio and video are defined in [5] as the time between noticeable transmission impairments, and the performance expectation for error free intervals is on the order of minutes. In the context of modern fiber based networks (where the principal source of impairments is network congestion and not transmission errors), transmission error-free intervals correspond to congestion free intervals defined here.

Given the above discussion, we approximate the time between congestion events, i.e., intercongestion time, as a first hitting time. Here we define the first hitting time, $\tau$, as the time it takes the system to go from an idle state to the level $b$ for the first time. First hitting times are discussed in [2], [4].

## III. Predicting the Time Between Congestion Events

Let $P_b$ be the stationary state probability that $X(t) = b$. This probability is assumed to be small. From the law of rare events [1] we can assume that the process representing successive excursions of $X(t)$ above $b$ is Poisson with rate $\Lambda$, that is, since the probability of a congestion event is small, the number of congestion events in an interval $(t_1, t_2]$ will approximately follow a Poisson distribution. Define the first hitting time, $\tau$, as the time it takes the system to go from an idle state to the level $b$ for the first time. Let $C_G$ be the sojourn time of $X(t)$ above the level $b$ in general. From the Poisson assumption for the successive excursions of $X(t)$ above $b$ we can write

$$P(\tau > t) \cong e^{-\Lambda t}$$

The average first hitting time is then

$$E[\tau] \cong 1/\Lambda$$

and from the Poisson Clumping Heuristic [2] we can write

$$E[\tau] = E[C_G]/P_b$$

The stationary state probability $P_b$ is known for many cases [1], [4], [29], [30]. The sojourn time above the level $b$ is given for an M/M/1 system in [2].

Then using the approach from [2] the average first hitting time for a M/M/1 system with $\alpha =$ packet arrival rate and $\beta =$ packet service rate is

$$E[\tau] \cong \frac{\beta}{(\beta-\alpha)^2}\left(\frac{\beta}{\alpha}\right)^b$$

Also, as shown in [2], the above approximation captures the dominant terms in the exact expression for the first hitting time of an M/M/1 system. The Poisson Clumping Heuristic [2] has been applied in a networking context in [3], [8]–[11] primarily dealing with M/M/$\infty$ system models to capture the aggregation of identical bursty sources in a fluid flow model. In [11] the Poisson Clumping Heuristic for the M/M/$\infty$ system is rigorously justified. However, the sojourn time above the level $b$ is not generally known.

Here, we combine the Poisson Clumping Heuristic with a simple approximation for the sojourn time above the level $b$ in order to determine the average first hitting for general service time distributions with finite variance. The approximation is based on the observation that the sojourn time of $X(t)$ above the level $b$ is proportional to the average waiting time in the associated M/G/1 queue. Thus, we weight sojourn time above the level $b$ for the M/M/1 system to approximate it for the more general case. Weighting by a descriptor of a "comparable" Poisson process is related to the Poisson Traffic Comparison that was defined and used in [12]. For general service time distribution, the sojourn time above the level $b$ becomes

$$E[C_G] \cong E[C_M](E[W_G]/E[W_M])$$

where

$C_G \cong$ sojourn time of $X(t)$ above the level $b$ in general;
$C_M \cong$ sojourn time of $X(t)$ above the level $b$ for the M/M/1 system;
$W_G =$ waiting time for general case;
$W_M =$ waiting time the M/M/1 system.

It is well known that [4] $E[W_G]/E[W_M] = (1 + C_s^2)/2$ where $S$ is the service time and $C_s^2 = \mathrm{Var}[S]/(E[S])^2$ is its squared coefficient of variation.

Note that $E[C_G]$ can be used as a secondary QoS metric. By combining the above approximation with the Poisson Clumping Heuristic we can write the average time between congestion events as

$$E[\tau] \cong \left(\frac{1}{(\beta-\alpha)}\right)\left(\frac{1+C_S^2}{2}\right)\left(\frac{1}{P_b}\right). \qquad (1)$$

This new approximation can be used to predict the average time between congestion events for general message length distributions with finite variance. Note estimates for $E[\tau]$ using $1/P_b$, that is, the mean recurrence time [34] based on the reciprocal of the stationary state probability, will be in error relative to this approximation. However the mean recurrence time does provides an alternate approach to predicting the time between congestion events. Our results show that the recurrence time approximation overestimates the rate of congestion events and would lead to conservative designs with over-provisioned links as the subsequent experiments demonstrate. The nature of the loss process induced by congestion events also plays a critical role in the performance of TCP; not only the average time between losses but its second order statistics govern TCP throughput [25]. Thus, increasing the understanding of the temporal characteristics of congestion events improves the ability to predict TCP performance.

While (1) is for a single queue, the approach proposed here can be directly extended to a network. Thus, a key attribute of the proposed approach is that it can be applied to predict the rate of congestion events for an end-to-end flow, i.e., a flow that traverses several routers. At each router, $k$, along a route of $M$ routers, the process representing successive congestion events is Poisson with rate $\Lambda_k$. Congestion events at each router are assumed to be independent (which is true when there is sufficient mixing of different traffic flows at each router). Thus, the aggregate congestion event process observed by the customer is the result of merging independent Poisson processes, which produces a Poisson process model for congestion events on the end-to-end path. Given these assumptions each router along the path contributes $\Lambda_k$ congestion events/sec to the total observed by the customer. An end-to-end flow that transverses a set of $M$ successive routers will experience congestion events at a aggregate rate

$$\Lambda_T = \sum_{k=1}^{M} \Lambda_k.$$

Given the nature of the traffic at the output port of each router, the total rate of congestion events $\Lambda_T$ can be predicted for end-to-end flows and used in network design.

## IV. Performance Evaluation

To validate the above approximation a set of experiments was conducted using the commercial simulation package EXTEND [26]. Four different types of traffic are considered: three different service time distributions, fixed-M/D/1 ($C_s^2 = 0$), exponential-M/M/1 ($C_s^2 = 1$), and hyperexponential-M/H(2)/1 ($C_s^2 = 1.5$) and hyperexponential distributed interarrival times-H(2)/M/1. For the hyperexponential distributed service time $k = 2$. For the hyperexponential-2 arrival process, the interarrvial time has a mean of $1/\lambda_1$ with probability $w_1$ and mean of $1/\lambda_2$ with probability $w_2$ where $w_2 = 1 - w_1$. (See [4], [30] for a discussion of the hyperexponential distribution.) Internet traffic has been observed to have significant variability (burstiness) over several time scales [7]. While such variation is often attributed to long-range dependent traffic, several recent studies show both empirically [6] and theoretically [7] that the hyperexponential distribution may also capture the source of the observed variation. The stationary state probability, $P_b$, for these cases can be found in [1], [4], [29], [30]. In each case the average service time was normalized to unity and the arrival rate varied to change the load.

For the single queue case, the following figures show a comparison between the number of congestion events per million-service times measured using simulation and the number predicted using the proposed approximation. The approximation accurately predicts the number of congestion events for the exponential and both hypexponential cases, but is not as accurate for the deterministic case. In all cases the predictions based on approximation developed here are closer to the observations than the simple approach based mean recurrence time, $1/P_b$. As expected, predictions based on mean recurrence over-estimates the rate of congestion events.

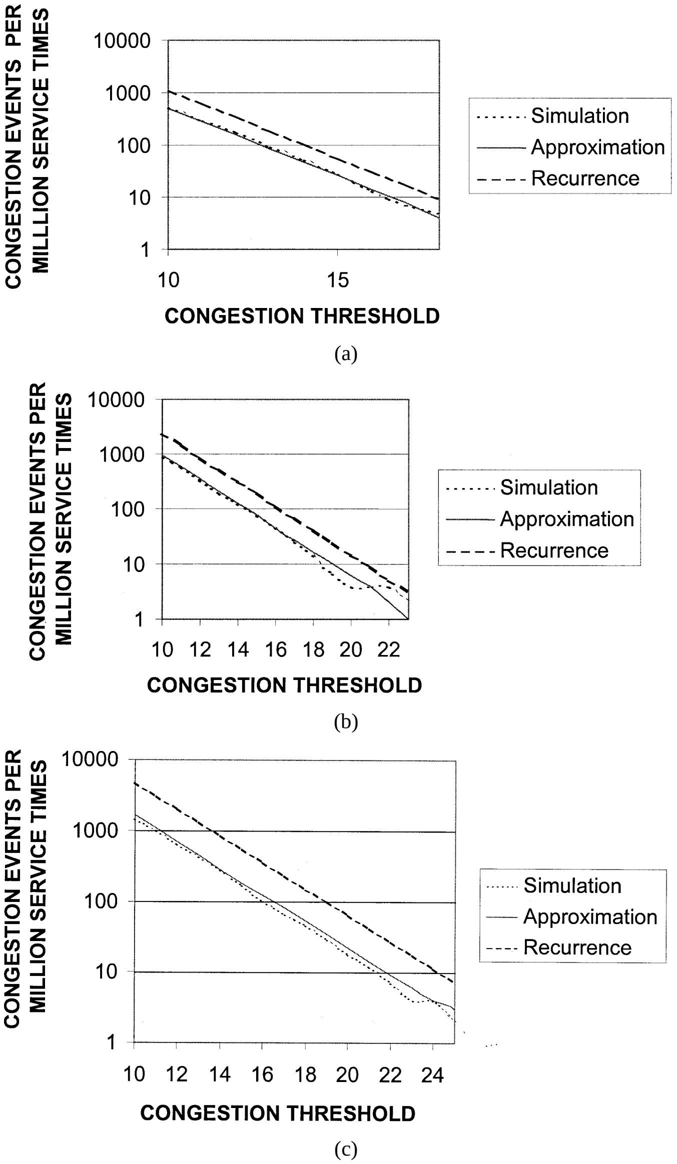
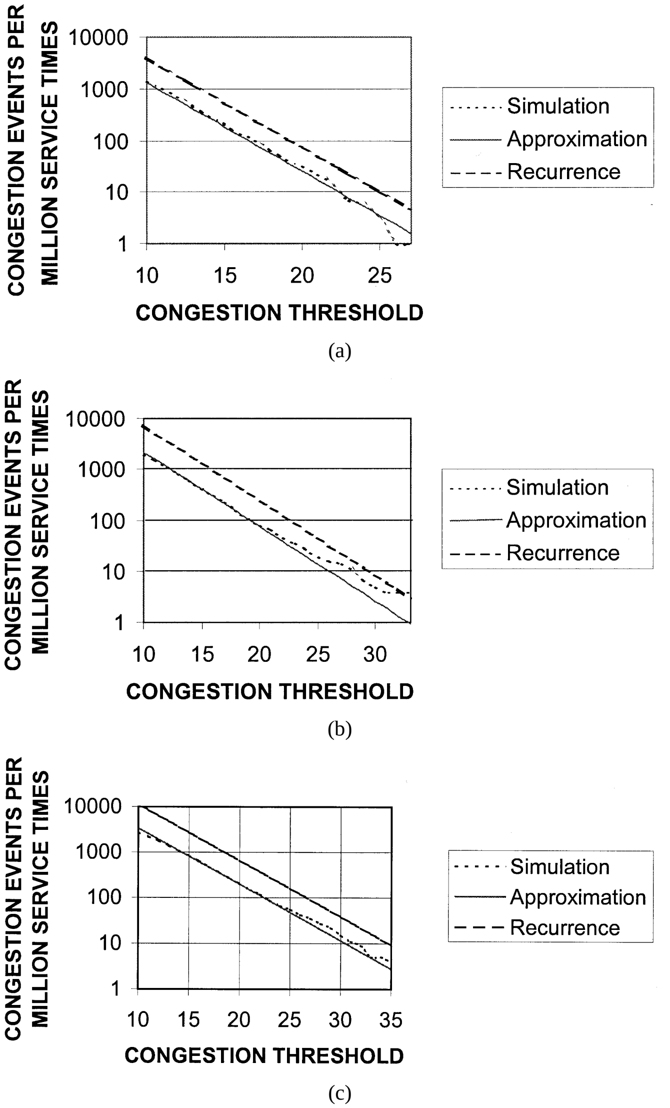


Fig. 2 (a) M/M/1 queue: load 0.55. (b) M/M/1 queue: load 0.60. (c) M/M/1 queue: load 0.65.

Fig. 2(a)–(c) compares the predictions made based on the proposed approximation to observations from simulation for an M/M/1 system. The results for the M/H(2)/1 system are given in Figs. 3(a)–(c) and 4(a)–(c) present the performance of the approximation for the M/D/1 system. The performance of the H(2)/M/1 systems is given in Fig. 5(a)–(c). For the fixed service time case the proposed methodology does not predict the number of congestion events as closely as in the other cases. However, it is still a better predictor than the mean recurrence time. Also, as expected, the number of congestion events decreases as the square coefficient of variation decreases and there

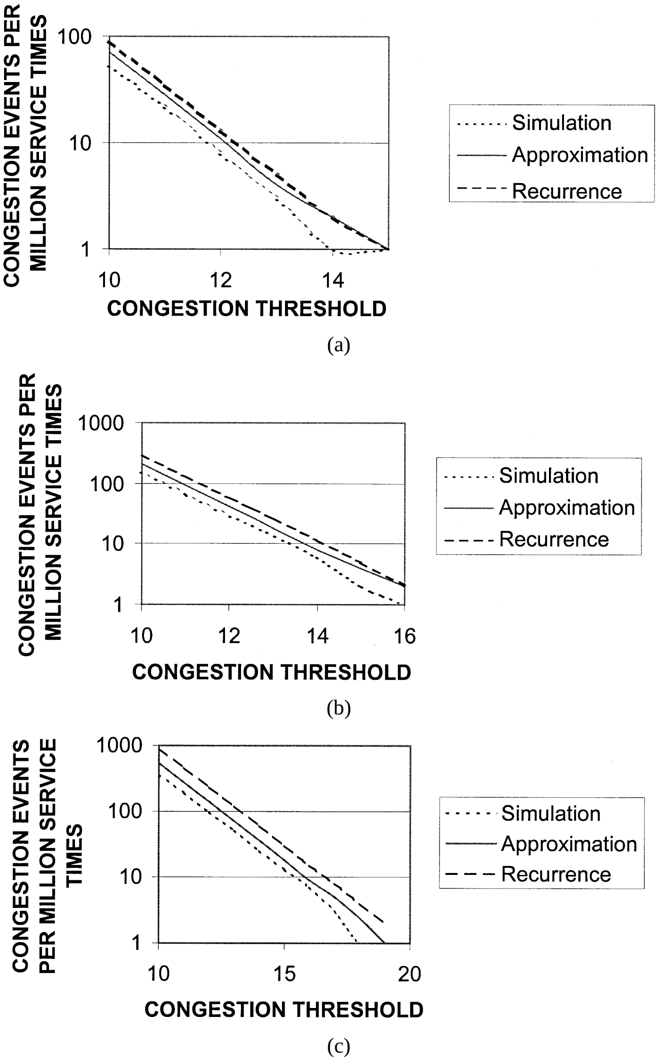


(a)

(b)

(c)

Fig. 3. (a)M/H(2)/1 load 0.55. (b) M/H(2)/1 load 0.60. (c) M/H(2)/1 load 0.65.



(a)

(b)

(c)

Fig. 4. (a) M/D/1 load 0.60. (b) M/D/1 load 0.65. (c) M/D/1 load 0.70.

is a log-linear relationship between the number of congestion events and the threshold level $b$.

To evaluate the potential of the proposed approximation in a network environment a simple network of two queues was considered in which the second queue randomly accepts one half of the traffic that is output from the first queue as well as independent traffic. Note this simple system captures important characteristics of realistic edge/core networks. The traffic into the second queue includes realistic queueing dynamics of the edge node. The second queue is served at the same rate as the first queue as shown in Fig. 6. Fig. 7(a) demonstrates that the proposed approximation provides accurate predictions for tandem M/M/1 systems as expected, while Fig. 7(b) and (c) indicates that the approximation can be used with other message length and interarrival time distributions, in this case the hyperexponential. Again, the proposed approach is not as accurate for the M/D/1 case as shown in Fig. 7(d).

The more complex networks shown in Fig. 8 were also studied, in these networks the traffic flows through multiple queues, as would be the case in edge/core topologies. Results for only the M/H(2) are presented in Fig. 9. For the target flow the load on each node is the same for all three topologies thus the rate of congestion events also remains unchanged. There is some difference between the simulation results and proposed approximation for the tandem network model. However, the prediction of the rate of congestion events is indistinguishable (the lines in Fig. 9 overlap) from the simulation results for the more complex systems.

To illustrate the application of the proposed methodology consider a network architecture that has two distinct hierarchical layers—an edge and a core. The core is the backbone of the network and consists of high-speed elements. The edge portion of the network provides access to the core network and serves to aggregate traffic into the network core. Suppose a multimedia end-to-end flow passes through an edge router, two core routers and then another edge router to arrive at the destination. Assume the edge and core routers are connected by 10 Mb/s and 100 Mb/s links respectively (note these rates could either be set by physical or logical limitations). Congestion occurs when the demands for network resources exceed capacity for an extended period of time, $T_c$ (in seconds). That is, the offered load exceeds capacity by some amount, $\Delta_C$ (b/s), such that the output queue grows approximately linearly as long as the congestion persists.

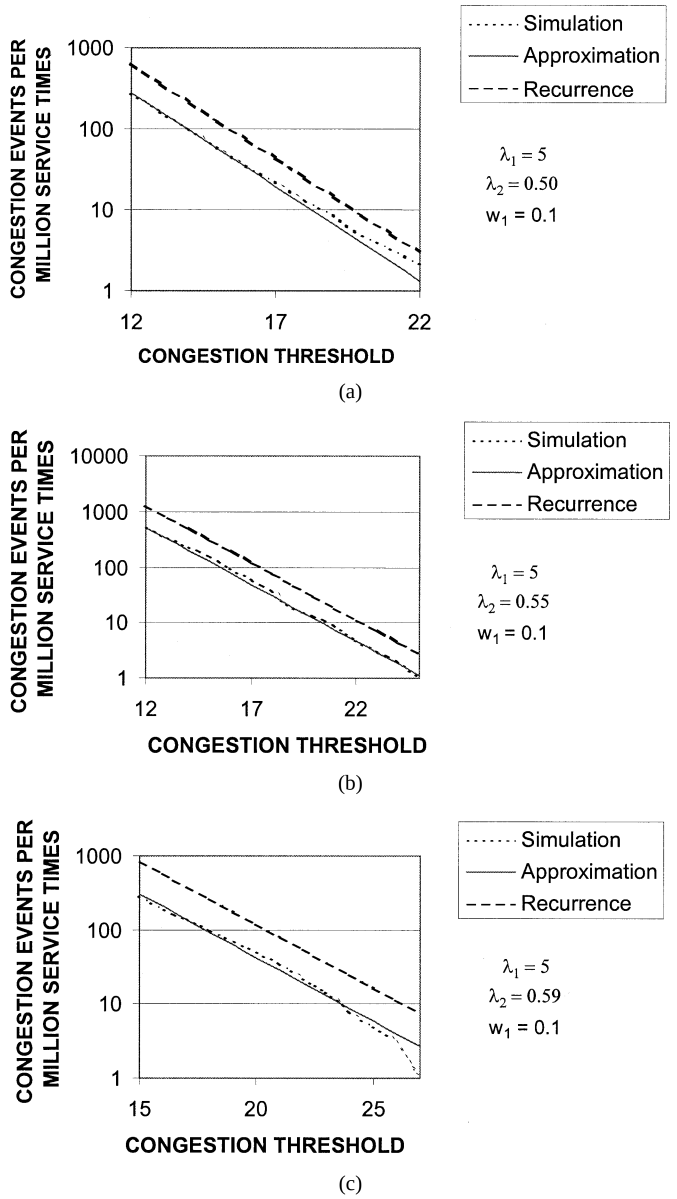


Fig. 5. (a) H(2)/M/1 load 0.55. (b) H(2)/M/1 load 0.60. (c) H(2)/M/1 load 0.65.

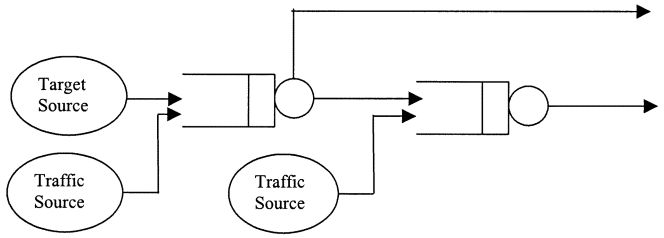


Fig. 6. Tandem network model.

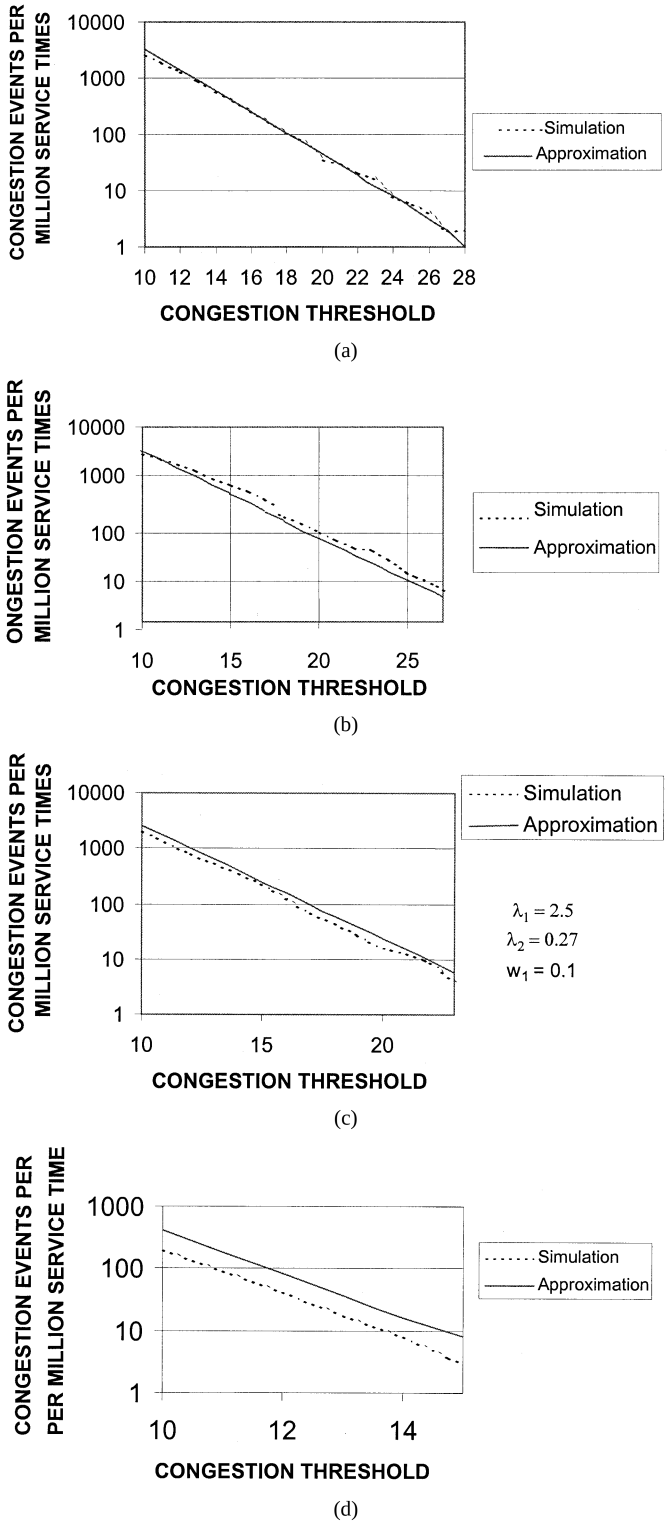
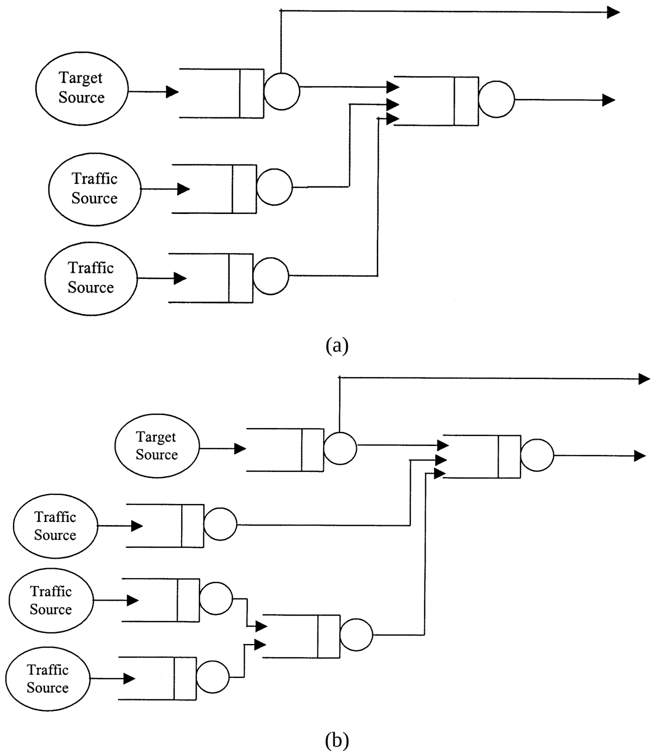


Fig. 7. (a) Tandem M/M/1 load 0.65. (b) Tandem M/H(2)/1 load 0.55. (c) Tandem H(2)/M/1 load 0.60 (d) Tandem M/D/1 load 0.65.

Given the link rate, $\Delta_C$, and $T_c$, the queue threshold level, $b$, can be calculated. Assume the edge and core routers use a $\Delta_C$ of 10% of the link capacity and a $T_c$ of 250 ms and 100 ms respectively. Also assume an average packet length of 1 KB. Thus congestion occurs in the core router if the link has an offered load of 110 Mb/s for 100 ms yielding $b = 125$ (note other combinations of$\Delta_C$, and $T_c$ will also result in a $b = 125$). In [5] a requirement of an error free interval for audio and video multimedia desktop conferencing teleservices is given as 30 min, thus the quality of service criteria will be an average congestion free interval of 30 min. Further suppose that it is required to have fewer congestion events in the core of the network compared to the edge. In this scenario assume that 80% of the congestion events are allocated to the edge while the remaining 20% to the



(a)

(b)

Fig. 8 (a) Network topology 1. (b) Network topology 2.

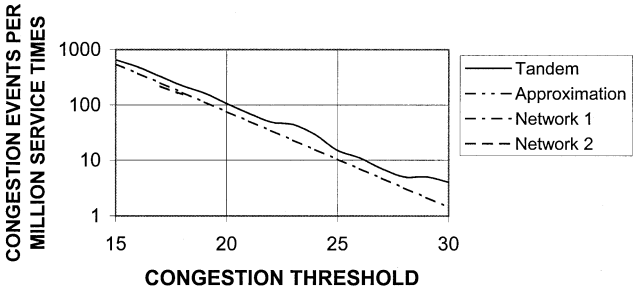


Fig. 9. Networks with M/H(2) Traffic-load 0.55.

core routers. Assuming an M/H(2)/1 model, the average congestion free interval perceived by the end user will be greater than 30 min if the load on the edge and core routers is less than 0.59 and 0.71 respectively.

## V. Conclusion

It has been recognized that the temporal nature of network congestion significantly impacts user-perceived and TCP performance. The acknowledgment of the importance of loss patterns has recently led to the initial definition of new metrics in the IETF. The proposals in the IETF and other previous research have focused on measurement-based approaches and not techniques to predict the loss patterns given the nature of the traffic. Such prediction methodologies are needed for Internet design and engineering.

The contribution of this effort is a new QoS metric based on the number of congestion events per unit time as well as a methodology for its prediction. The premise of this work is that the first hitting time can be used as the theoretical basis to characterize the rate of congestion events in a network context. A rigorously defined first hitting time performance metric is well matched to Internet design and performance analysis problems. Network customers can easily understand the resulting metric.

For network design, the total rate of congestion events for an end-to-end flow can be allocated to different network elements, that is, a congestion event budget can be distributed to various segments of the system. For example, more of the congestion events can be allocated to the edge of the network while the core is provisioned for minimum congestion. Alternately, each hop along a route can be designed to limit the rate of congestion events, and then the number of hops along the path can bound the end-to-end congestion rate. Thus, routing can be used to directly control the QoS.

The results presented here suggest several areas for future research. Additional research is required to improve its accuracy for deterministic service times. Many traffic models have been studied and applied to network analysis [31], [32], only limited set of these have been considered here, thus there is need to generalize the methodology to allow for a wider variety of traffic sources. For example, the suitability of the approximation for traffic models containing significant correlation should be considered, e.g., TES models developed in [14]. The approximation for predicting the time between congestion events presented here can only be applied to traffic with finite variance. Other approaches are needed to apply congestion event analysis to self-similar traffic.

Network events as defined here maybe caused by factors other than the user traffic, for example routing changes may cause outage periods in the order of tens of seconds several times a day [35]. A complete characterization of temporal characterization of network events should also consider these other factors.

## Acknowledgment

The author thanks Dr. N. Akar, Assistant Professor, Electrical and Electronics Engineering Department, Bilkent University, Ankara Turkey, Dr. D. Petr of the University of Kansas, Dr. G. Lazarou of Mississippi State University, Dr. B. Melamed of Rutgers University, and the reviewers for their helpful comments and suggestions.

## References

[1] H. Taylor and S. Karlin, *An Introduction to Stochastic Modeling*. London, U.K.: Academic, 1998.

[2] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springler-Verlag, 1989.

[3] A. Somonian, F. Guillemin, and L. Romoeuf, "Transient analysis of statistical multiplexing on an ATM link," in *Proc. IEEE Globecom'92*, 1992.

[4] R. Nelson, *Probability, Stochastic Processes and Queueing Theory*. New York: Springer-Verlag, 1995.

[5] Multimedia Communications Quality of Service (1995). [Online]. Available: www.mmcf.org

[6] P. M. Fiorini, "Modeling telecommunications systems with self-similar data traffic," Ph.D. dissertation, Univ. Connecticut, Storrs, Sept. 1998.

[7] G. Lazarou, "On the variability of Internet traffic," Ph.D. dissertation, Univ. Kansas, Lawrence, Aug. 2000.

[8] J. Roberts, U. Mocci, and J. Virtamo, Eds., *Broadband Network Teletraffic-Performance Evaluation and Design of Broadband Multiservice Networks: Final Report of Action COST 242*. Berlin, Germany: Springer, 1996.

[9] F. Guillemin, G. Rubino, B. Sericola, and A. Simonian, “Transient characteristics of an M/M/$\infty$ system applied to statistical multiplexing on an ATM link,”, INRIA Publ. 874, Oct. 1994.

[10] A. Dupuis, F. Guillemin, and B. Sericola, “Asymptotic results for the superposition of a large number of data connections on an ATM link,”, INRIA Res. Rep. 3010, Sept. 1996.

[11] F. Guillemin and A. Simonian, “Transient characteristics of an M/M/$\infty$ system,” *Adv. Appl. Prob.*, vol. 27, pp. 862–888, 1995.

[12] D. L. Jagerman and B. Melamed, “Burstiness descriptors of traffic streams: Indices of dispersion and peakedness,” in *Proc. 1994 Conf.Information Sciences*, vol. 1, Princeton, NJ, 1994, pp. 24–28.

[13] ——, “The run probabilities for the TES processes,” *Commun. Statist-Stoch. Models*, vol. 10, no. 4, pp. 831–851, 1994.

[14] ——, “The transition and autocorrelation structure of TES processes part I: General theory,” *Stoch. Models*, vol. 8, no. 2, pp. 193–219.

[15] M. Bolella, D. Swider, S. Uludag, and G. Brewster, “Internet packet loss: Measurement and implications for end-to-end QoS,” in *Proc. Int. Conf. Parallel Processing*, Aug. 1998, pp. 3–15.

[16] V. Markovski and L. Trajkovic, “Analysis of the loss episodes for video transfers over UDP,” in *Proc. SPECTS’2K*, Vancouver, BC, Canada, July 2000, pp. 278–285.

[17] D. W. Petr, L. A. DaSilva, and V. S. Frost, “Priority discarding of speech in integrated packet networks,” *IEEE J. Select. Areas Commun.*, vol. 7, pp. 644–656, June 1989.

[18] L. A. DaSilva, D. W. Petr, and V. S. Frost, “A class-oriented replacement technique for lost speech packets,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1597–1600, Oct. 1989.

[19] Int. Telecommun. Union (1996). [Online]. Available: http://www.itc.int

[20] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec, and J. Mahler, “Real-time voice over packet-switched networks,” *IEEE Network*, Jan./Feb. 1998.

[21] R. Koodli and R. Ravikanth, “One-way loss pattern sample metrics,” Task Force, IP Performance Metrics Working Group, Internet Draft, Internet Eng., Nov. 2000.

[22] W. Jiang and H. Schulzrinne, “QoS measurement of internet real-time multimedia services,” in *Proc. NOSSDAV*, Chapel Hill, NC, June 2000.

[23] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, “Measurement and modeling of the temporal dependence in packet loss,” in *Proc. Conf. Computer Communications (IEEE Infocom)*, New York, Mar. 1999, pp. 345–352.

[24] *Voice Quality Recommendations for IP Telephony*. Draft 6 - PN4689, to be published as EIA/TIA/TSB-116.

[25] C. Barakat, “TCP/IP modeling and validation,” *IEEE Network*, vol. 15, no. 3, pp. 38–47, May/June 2001.

[26] EXTEND, Simulation Software [Online]. Available: http://ww.imaginethatinc.com/

[27] Y. I. Wijata, D. Niehaus, and V. S. Frost, “A scalable agent-based network measurement infrastructure,” *IEEE Commun. Mag.*, Sept. 2000.

[28] B. L. Tierney, D. Gunter, J. Lee, M. Stoufer, and J. Evans, “Enabling network-aware applications,” in *Proc. Tenth IEEE Int. Symp. High Performance Distributed Computing*, San Francisco, CA, Aug. 2001, pp. 281–302.

[29] COM2—Computational Methods for the Performance Analysis of Broadband Communication Networks [Online]. Available: http://keskus.hut.fi/tutkimus/com2/

[30] L. Kleinrock, *Queueing Systems: Volume I: Theory*. New York: Wiley, 1975.

[31] A. Adas, “Traffic models in broadband networks,” *IEEE Commun. Mag.*, pp. 82–89, July 1997.

[32] H. Hichiel and K. Laevens, “Teletraffic engineering in a broadband era,” in *Proc. IEEE*, vol. 85, Dec. 1997, pp. 2007–2033.

[33] P. Seveik, “QoS: Shoe me the money,” presented at the Panel on Internet QoS: Technology, Status, Deployment, IEEE ICC 2002, New York, Apr. 2002.

[34] H. J. Larson and B. O. Shubert, *Probabilistic Models in Engineering Sciences*. New York: Wiley, 1979, vol. II.

[35] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, “Assessment of VoIP quality over Internet backbones,” in *Proc.INFOCOM 2002*, New York, June 2002.

**Victor S. Frost** (S’75–M’82–SM’90–F’98) received the B.S., M.S., and Ph.D. degrees from the University of Kansas, Lawrence, in 1977, 1978, and 1982, respectively.

He is currently the Dan F. Servey Distinguished Professor of Electrical Engineering and Computer Science and Director of the University of Kansas Information and Telecommunications Technology Center. His current research interest is in the areas of integrated broadband communication networks, communications system analysis, and traffic and network engineering and simulation. He was principal investigator on the University of Kansas MAGIC gigabit network research effort and ACTS ATM Internetwork (AAI). His research has been sponsored by NSF, DARPA, Rome Labs, NASA, Sprint, NEC America, NCR, BNR, NEC, Telesat Canada, AT&T, McDonnel Douglas, and COMDISCO Systems. From 1987 to 1996, he was the Director of the Telecommunications and Information Sciences Laboratory.

Dr. Frost received the Presidential Young Investigator Award from the National Science Foundation in 1984. He is a member of the Senator Pat Roberts Task Force on Information, Telecommunications and Computing Technology. He received an Air Force Summer Faculty Fellowship, a Ralph R. Teetor Educational Award from the Society of Automotive Engineers, and the Miller Professional Development Awards for Engineering Research and Service in 1986 and 1991, respectively.