Local Allocation of End-to-End Quality-of-Service in High-Speed Networks[1]

Ramesh Nagarajan[2], Jim Kurose[3] and Don Towsley[4]

Abstract

Quality-of-service (QOS) requirements for applications in high-speed networks are typically specified on an end-to-end basis. Mapping this end-to-end requirement to nodal requirements facilitates providing QOS guarantees and simplifies connection admission. In this paper, we evaluate strategies for such local allocation of the end-to-end QOS. A QOS allocation policy is said to perform better than another when the maximum network load that it can support is greater. A major contribution of this work is the development of a *nodal metric* that predicts the relative performance of QOS allocation policies in a network setting. Computation of the nodal metric and direct evaluation of allocation policy performance for two simple network models yield valuable insight into the choice of allocation policies. Intuitively, one expects significant differences in the performance of allocation policies when there are significant imbalances in nodal resource capacities or traffic loads. It is found, however, that with the *packet loss probability* as the QOS metric, there is *little difference in the performance of allocation policies* in the regime of applications with low loss requirements. From a practical viewpoint, this suggests that a simple allocation policy may be adopted in this scenario with only a small decrease in carried load with respect to an optimal policy. For applications which tolerate *large packet loss* or alternate QOS metrics, however, *QOS allocation policies differ significantly* in their performance. Our results indicate that the development of "optimal" QOS allocation policies is of interest in such cases.

KEYWORDS: Quality-of-Service; High-Speed Networks; Resource Allocation

[1] This research was supported in part by the National Research Foundation under grant NCR-9116183 and the Defense Advanced Projects Research Agency under contract NAG2-578

[2] Dept. of Electrical and Computer Engineering, Univ. of Massachusetts, Amherst, MA 01003

[3] Dept. of Computer Science, Univ. of Massachusetts, Amherst, MA 01003

[4] Dept. of Computer Science, Univ. of Massachusetts, Amherst, MA 01003

1 Introduction

Current packet communication networks offer users very little in terms of a guaranteed quality-of-service beyond a "best-effort" delivery of information. If the goal of providing a variety of services on a single integrated network is to become a reality, future networks will need to provide explicit end-to-end QOS guarantees to subscribers.

A number of recent research efforts have focussed on the problem of guaranteeing QOS. Some of this work addresses the issue of local or nodal QOS guarantees, while the remainder considers end-to-end guarantees. First, let us consider nodal guarantees. Guerin *et al.* [G^{+}91, GG92] propose the notion of equivalent capacity which is the capacity to be allocated to the connection at each node in order to satisfy a *nodal QOS requirement* and depends in general on nodal resources such as buffer space. Nagarajan and Kurose [NK92] consider the issue of appropriate QOS metrics for applications in high-speed networks and approaches to *guaranteeing* these metrics at the *nodal level*. Woodruff and Kositpaiboon [WK90] demonstrate via simulation how *nodal QOS measures* can be satisfied. All of the aforementioned research efforts assume that a *nodal QOS requirement* is specified. QOS requirements for applications are, however, often specified on an end-to-end basis. In order that any of these techniques be applicable, this end-to-end requirement must therefore be mapped to nodal QOS requirements.

We next consider the issue of end-to-end QOS guarantees. Cruz [Cru91] outlines the computation of *worst-case end-to-end delay bounds* for sessions in arbitrary networks. Kurose [Kur92] outlines the computation of *nodal performance bounds* (on distributions of pertinent quantities such as packet delay) on a per-session basis when the session traffic is stochastically bounded over intervals of time. This bounding technique also permits one to compute point-valued worst-case delay bounds. Golestani [Gol90] also provides an *worst-case end-to-end delay bound* for sessions provided a special stop-and-go service discipline is adopted at the network nodes. A straightforward application of the proposed techniques [Kur92, Cru91] in a connection admission algorithm will require the algorithm to compute, at connection setup times, the end-to-end delay bounds for all affected sessions on the new session's route in order to ensure that no QOS guarantees for existing sessions are violated upon admitting the new connection. This can be cumbersome and time-consuming at best. The problem might be alleviated by apportioning the end-to-end guarantee locally for each connection and then simply verifying that the local guarantee, rather than the end-to-end guarantee, is still satisfied. The research efforts of [FV90, VHN92] adopt this approach. In all of these approaches, a "best" possible guarantee is computed at each of the nodes for the new session while satisfying local guarantees for existing sessions. The local guarantees are then aggregated and the "excess" over the required QOS value for the new session is reassigned among the nodes. However, neither of [FV90, VHN92] address in detail the assignment of the excess end-to-end QOS to nodes. In [FV90], for example, the excess value is simply equally distributed among the nodes along the source-destination path.

Thus the important and interesting problem of *apportioning the end-to-end QOS values to local nodes* has been largely ignored in recent research efforts. This issue of the allocation of the end-to-end QOS is the primary focus of this paper. In [WN91], we addressed this issue of QOS allocation under the assumption that nodal QOS guarantees are provided by explicitly allocating resources to each connection at each node. In this paper, we consider a more realistic scenario in which nodal resources are shared among connections. We consider policies for QOS allocation which maximize network efficiency, where network efficiency is measured by the total number of connections or the total traffic load that can be supported by the network while still meeting all QOS guarantees. When there is an imbalance in the network resulting in "bottleneck" nodes where resources are scarce (and thus valuable), we shall see that it pays to meet the required end-to-end guarantees by requiring less stringent local guarantees at bottleneck nodes while compensating with more stringent local guarantees at other nodes of the network. We also uncover, however, scenarios where it does not pay significantly to "optimally" allocate the end-to-end QOS even when there are considerable imbalances in the network.

The remainder of this paper is organized as follows. Section 2 presents a formal statement of the QOS allocation problem. Prior to addressing this QOS allocation problem we discuss network traffic models in Section 3. Section 4 examines certain fundamental aspects of the QOS allocation problem presented in Section 2 and argues that it is feasible to arrive at conclusions about the performance of QOS allocation policies in arbitrary networks by simply examining nodal performance. For purposes of concreteness and to demonstrate the usefulness of the measure of allocation policy efficiency developed in section 4, we consider two particular instances of the general network model in Section 2, with the packet loss probability as the QOS metric. In section 5, we consider a tandem set of queues. We then investigate the actual performance of allocation policies in the context of this model. Next, in Section 6, we consider a tandem set of queues again but now allow for interfering traffic. Section 7 considers, briefly, the impact of alternate QOS metrics on QOS allocation policies. Finally, Section 8 summarizes the paper and discusses open problems for future research.

2 A General QOS Allocation Problem

In this section, we formulate a general QOS allocation problem. While we do not attempt to solve for allocation policy performance in this general model, we do consider two particular instances of this general model in Sections 5 and 6. Further, this general model will motivate our development of an allocation policy performance metric in Section 4.

Consider a communication network of V nodes labeled as $1, 2, \cdots, V$. These nodes are taken to include a set of source-destination pairs that communicate over fixed routes. We do not address routing of connections, implicitly assuming instead that the route is determined *a priori*. We denote by Ω the set of all routes in this network. The routes in this set will be denoted by $\omega_i, \;\; i = 1, 2, \cdots, M$ where M is the total number of routes. Each route, ω_i, in turn is taken to be a string of digits corresponding to the nodes (labels) that the particular route traverses.

Connection requests arrive at the source node of a source-destination pair with a specified end-to-end QOS criteria, denoted as Q. In this paper, we assume homogeneous connections, i.e., identical traffic characteristics and QOS criteria. Each connection is admitted or rejected by an allocation (admission) policy, denoted by π, which guarantees the end-to-end QOS by assigning fractions of the desired end-to-end QOS among the nodes of the connection. When $M > 1$, we require allocation policies to maintain a certain load ratio among the different routes of the network, i.e., we require that the allocation policy support a fraction p_ω ($\sum_{\omega \in \Omega} p_\omega = 1.0$) of the total number of connections on path ω. The above restriction on load ratios is introduced only to prevent the allocation problem for $M > 1$ to degenerate into that for $M = 1$. For example, when $M > 1$ it might be the case that the network load is maximized (when there is no restriction on load ratios) if connections are permitted only on the path with relatively large resource capacities, i.e., the problem has reduced to the QOS allocation problem for this single path. We denote by q_i^ω the locally allocated QOS guarantee at some node $i \in \omega$ of the end-to-end QOS on path $\omega \in \Omega$. Over time, connections arrive and when accepted terminate after a finite duration. In this paper, we do not study this dynamic process, but, rather, study the best that each policy can achieve: for a given policy and network, N_π is the maximum number of connections admitted. More formally, our interest will be in ascertaining N_π such that $\Gamma(q_l^{\omega_i}, q_m^{\omega_i}, \cdots, q_r^{\omega_i}) \leq Q, \ \ \forall \omega_i, \ i = 1, 2, \cdots, M$, where $\Gamma(\cdot)$ denotes an arbitrary function that determines the end-to-end QOS given the local guarantees. We assume in this paper that the QOS value, Q, is in general a scalar real-valued quantity, i.e., we exclude more sophisticated QOS metrics such as those considered in [NK92]. Further, we assume that the function $\Gamma(\cdot)$, while arbitrary, is of a form which requires $q_i^\omega \leq Q$.

We do not delve into the details of how the local guarantees, q_i^ω, determine the locally supportable load and hence the network supportable load, N_π, but will instead defer this discussion to later sections when we consider specific network models and connection types. Our goal will be to compare the different policies based on N_π, which measures how efficiently the total network resources are being used to satisfy the connection requests. In the rest of this paper, we refer to this value (in a qualitative sense) of the total number of connections supportable under an allocation policy as the *performance* of the allocation policy.

3 Traffic models

In this section, we describe two stochastic traffic models that will be adopted in our investigation of the QOS allocation problem.

The first model assumes that each source generates traffic (packets) according to the classical Poisson process and that the packet sizes are exponentially distributed. We will refer to this source traffic model as M1 and denote its average rate by λ_s.

The second model considers each source as a packet voice source. This standard model has as its basic premise (see, e.g., [DL86, HL86, SW86, NKT91]) that an active voice source periodically generates fixed length packets when a speaker is speaking (talkspurt) and otherwise remains idle. We briefly describe this model here; the reader is referred to the above references, in particular [SW86], for additional details and discussion. The voice packetization period is assumed to be fixed at 16 msec. and the talkspurt is assumed to contain a geometrically distributed number of packets, with mean 22 packets. The mean length of a talkspurt is thus $\mu^{-1} = 352$ msec. The period between talkspurts, known as the silence period and denoted by X, is assumed to be exponentially distributed with a mean length of $\lambda^{-1} = 650$ msec.. The speech activity ratio, which is the fraction of time that the voice source is active, is thus 0.351 and each source generates on the average 22 packets every second. Given the above model, the interarrival times between packets generated by a *single* source form a renewal process. With probability 1/22, the interarrival time is 16 msec. and with probability 21/22, the interarrival time is $16 + X$ msec. [SW86]. We will refer to this source model as M2.

In this paper, we find an alternate *fluid* description of the M2 source more amenable to analysis. It is assumed that the source, when active, transmits information at an uniform rate rather than as discrete packets. The rate of source transmission will be specified in bits per second and hence also referred to as the bit rate. It is then meaningful to talk about the peak, mean and variance of the bit rate. We will denote these quantities by γ, m and σ^2 respectively. For the aforementioned

M2 source parameter values, we can compute these quantities to be

$$\begin{aligned} m &= 11.241 \text{ Kbps} \\ \gamma &= 32.0 \text{ Kbps} \\ \sigma &= 15.276 \text{ Kbps.} \end{aligned} \tag{1}$$

4 QOS criteria and Optimal QOS allocation

In this section, we develop a strategy for addressing the performance of QOS allocation policies in the setting of Section 2 by focussing on an isolated node. Such a strategy obviates the need to analyze the performance of a given allocation policy in myriad network topologies with varied connection routing patterns. A useful first step in developing such a strategy is to better understand the mechanics of QOS allocation and its influence on the load that can be supported in a network.

Consider first a simple QOS allocation policy which we will refer to as the *Equal Allocation* (EQ) policy (see also Sections 5 and 6). The policy simply requires that the burden of providing an end-to-end QOS be delegated equally to all of the nodes traversed by the connection. For example, if the QOS metric is end-to-end packet delay (d), then each node on the source-destination path of, say, n hops might be required to provide a delay guarantee smaller than d/n. This local value of the QOS metric completely determines the traffic load that can be supported at the node and hence in the network. It is intuitively clear, however, that this is not the best possible strategy when there is an imbalance in the capabilities of the nodes. For example, it may be advantageous to allocate more of the end-to-end delay to the nodes with smaller available bandwidths. This might enable one to support much larger traffic loads than with the EQ policy. It hence appears useful to compute some measure of the gain in supportable traffic load due to relaxed QOS requirements at the node. If this gain is large, one might expect a QOS allocation policy that allocates a large fraction of the end-to-end QOS to the node with small resource capacities and a small fraction to nodes with large capacities to perform considerably better than a naive policy such as the EQ policy. Otherwise, a simple policy such as the EQ Policy may suffice. In the following, we propose an useful nodal metric and outline its computation.

We define

$q = G_i(N, R_i)$: A real-valued function for some network node i that indicates the supportable QOS, q, (i.e., the realized performance) at that node while carrying a load of N sources at that node. R_i denotes nodal resources and maybe a multi-component vector including, for example, the bandwidth and buffer space. The notation suggests that $G_i(\cdot)$ is a function of two variables. In this paper, however, we treat R_i as a known and fixed parameter. Its presence in the notation is merely to emphasize the dependence of $G_i(\cdot)$ on R_i. Last, we assume in this paper that $G_i(\cdot)$ is a convex function of N.

Note that while N is an integer-valued quantity, we will treat it, for convenience, as a real-valued quantity in the rest of this paper. An alternate and more natural view is to consider a function $F_i(\cdot)$ such that $N' = F_i(q, R_i)$, i.e., N' is the supportable load at node i when it is required to meet a QOS criteria of q. We will assume in this paper that $G_i(\cdot)$ is a strictly increasing function of N and hence has a well-defined inverse function $G_i^{-1}(\cdot)$. We will then take $F_i(\cdot) = G_i^{-1}(\cdot)$ in the rest of this paper. Finally, we abbreviate the above notation to $q = G(N)$ in the following analysis since only q and N will be of interest.

We are now in a position to formally state our requirements. Consider a particular network node in isolation. Let q be the locally apportioned value of a end-to-end QOS requirement Q under a certain policy (say the EQ policy). Given our earlier discussion, the locally supportable load at the node is now $N = F(q)$. We are now interested in determining the new value of the supportable load, $N + \Delta N = F(q + \Delta q)$, when the local portion of the end-to-end QOS requirement is changed from q to $q + \Delta q$. In particular, we are interested in determining $\Delta N/N$, the relative gain in traffic load due to a change in the nodal QOS requirement. Since $G(N)$ is a convex function of N, we have

$$\begin{aligned} \Delta G(N) &= G(N + \Delta N) - G(N), \\ &\geq G(N) + \frac{dG(N)}{dN}\Delta N - G(N), \\ &\geq \frac{dG(N)}{dN}\Delta N \end{aligned} \tag{2}$$

or alternatively,

$$\begin{aligned} \frac{\Delta N}{N} &\leq \Delta G(N)\frac{dN/N}{dG(N)} \\ &\leq \frac{\Delta G(N)}{G(N)}\frac{dN/N}{dG(N)/G(N)} \\ &\leq \Phi(q, R)\,\frac{\Delta G(N)}{G(N)} \end{aligned} \tag{3}$$

where

$$\Phi(q, R) \equiv \frac{dN/N}{dG(N)/G(N)}|_{N=G^{-1}(q,R)} = \frac{G(N)}{N}\frac{1}{dG(N)/dN}|_{N=G^{-1}(q,R)} \tag{4}$$

will be referred to as the *Relative Gain Ratio (RGR)* and depends, in general, on nodal resources R and the local QOS value q. Note that the RGR is dimensionless, i.e., it is independent of the units of $G(\cdot)$. We could have derived the RGR by considering $F(\cdot)$ rather than $G(\cdot)$ and in this case we require that $F(q)$ is concave in q and the RGR is then

$$\Phi(q, R) \equiv \frac{dF(q)/F(q)}{dq/q} = \frac{q}{F(q)}\frac{dF(q)}{dq}. \tag{5}$$

The choice of the above two expressions to compute the RGR depends on which of $G(\cdot)$ or $F(\cdot)$ is available in closed-form.

We now remark on some important properties of the *RGR*. Note that for $\Delta G(N)/G(N) = 1$, i.e., a unit change (increase) in QOS, the gain in traffic load is bounded by the value of the RGR alone. In other words, the value of the RGR is a bound on the relative gain in traffic load for a unit increase (relaxation) in QOS. Further, for small values of Δq, the above inequality approaches an equality. Hence, large values of the RGR indicate a potential for large gains in traffic load by judicious local allocation of the end-to-end QOS. On the other hand, small values of the RGR suggest small differences in the performance of allocation policies. In this latter case, a simple allocation policy would be sufficient. The RGR is thus a useful indicator of allocation policy performance in a network scenario even though it is computed on a nodal basis only. We next compute the RGR for some sample QOS metrics and source traffic models.

RGR with the loss metric

Consider identical M1 sources at a node with a finite buffer space K and a "first-come,first-served" (FCFS) service discipline. The QOS metric is taken to be the packet loss probability. Hence, we have the $M/M/1/K$ queueing model for the node and the packet loss probability is:

$$q = G(\rho) = (1-\rho)\rho^K/(1-\rho^{K+1}). \tag{6}$$

where, as usual, $\rho = N\lambda_s/\mu$, is the traffic intensity. In [NT] we show that $G(\rho)$ is convex for $\rho \in [0,1]$, $K \geq 4$, which is typically a regime of practical interest.

Now $dN/N = d\rho/\rho$ and we can compute the RGR in this case as:

$$\Phi(q,K) = \frac{d\rho/\rho}{dG(\rho)/G(\rho)}|_{\rho=G^{-1}(q,K)}. \tag{7}$$

Hence, we have

$$\begin{aligned}\Phi(q,K) &= \frac{(1-\rho)(1-\rho^{K+1})}{(K(1-\rho)-\rho(1-\rho^K))} \\ &= \frac{(1-G^{-1}(q,K))(1-(G^{-1}(q,K))^{K+1})}{(K(1-G^{-1}(q,K))-G^{-1}(q,K)(1-(G^{-1}(q,K))^K))}.\end{aligned} \tag{8}$$

Figure 1 shows that the RGR metric is large for small values of the buffer size and large loss QOS values and small otherwise. Note that as $q \to 0$ for a fixed value of K, $RGR \to 1/K$ i.e., the RGR value for small values of the loss QOS requirement is approximately inversely proportional to the buffer capacity at the node.

Figure 2 illustrates the behaviour of the RGR for the $M/M/1/30$ queue in an alternate intuitively appealing fashion. Since Figure 2 is plotted on a log-log scale, identically sized intervals on either axis represent identical relative increments, e.g., $\Delta Q'/Q' = \Delta Q/Q$ where $Q = 2 \times 10^{-04}$ and $Q' = 2 \times 10^{-01}$. Hence, Figure 2 shows that for identical relative increments (relaxation) in the loss the relative gain in load is larger at the higher loss value (Q').

The RGR values for the $M/M/1/K$ queue hence indicate that only in the regime of large loss QOS values and small values of the buffer capacity, can one expect an optimal policy to perform significantly better than a simple policy such as the EQ policy. We will observe this in the context of network models in Sections 5 and 6.

We now consider a more realistic model of a node in a high-speed network in which the input traffic stream to the node is a superposition of on-off packet voice sources, i.e., M2 sources, and the service discipline is FCFS. The analysis of the voice sources multiplexer is rather complex [NKT91, B$^+$91, HL86, SW86, AMS82], only approximate numerical techniques are available and, in general, no closed-form expression $G(\cdot)$ is available for this realistic scenario. Hence, we adopt the

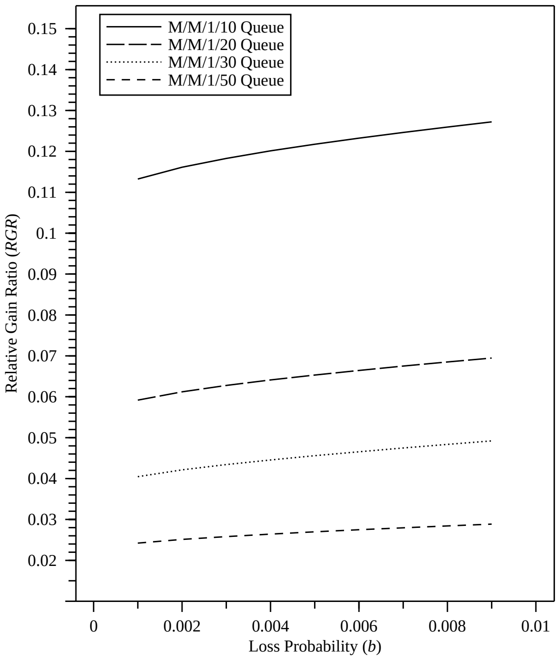

Figure 1: Relative Gain Ratio for the loss QOS metric and the $M/M/1/K$ queue

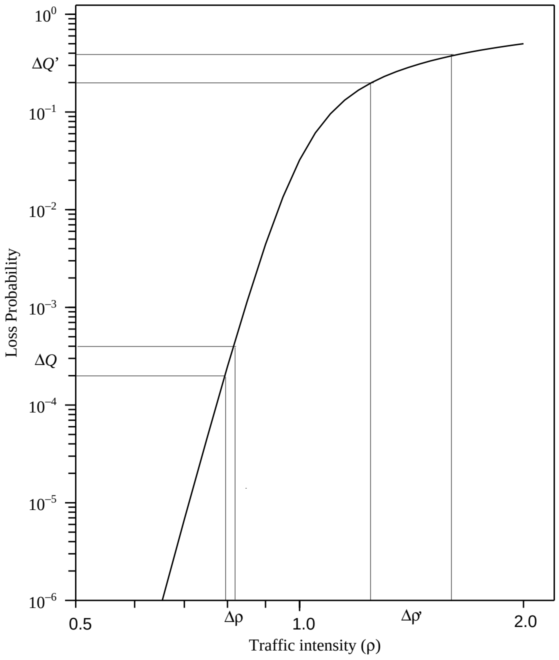

Figure 2: Relative gain in load and its relation to the relative increment (relaxation) of loss in the $M/M/1/30$ queue

approximate analysis of [G+91, GG92] in which closed-form expressions are developed for the so-called *equivalent capacity* - the amount of bandwidth needed to support a given QOS criteria. These closed-form expressions may then be employed in determining the RGR. The work of [G+91, GG92] considers two different approximations. The reader is referred to [G+91, GG92, NKT92] for details. We provide a general outline of the techniques in the following.

The first approximation in [G+91, GG92] is based on modeling the aggregate bit rate of the superposition of sources as a Gaussian distributed random variable whose mean and variance are easily determined from the individual source characteristics. The reader may refer to [NKT92] for a more detailed discussion of the merits of such an approximation. The packet loss probability[5] is taken to be the QOS metric and it is assumed that loss occurs whenever the aggregate bit rate exceeds the channel capacity. The equivalent capacity is then taken to be that real value beyond which the tail of the Gaussian distribution has mass below the required QOS criteria.

We assume N identical M2 sources being multiplexed onto a link of capacity C units. The first approximation in [G+91, GG92] yields the following relation between N, the number of sources, and q, the loss probability to be satisfied for the sources:

$$C = Nm + \alpha^{'}\sqrt{N}\sigma \tag{9}$$

where

$$\alpha^{'} = \sqrt{-2ln(q) - ln(2\pi)}. \tag{10}$$

The reader is referred to [NKT92] for some restrictions under which the above relation holds. Replacing N by x^2 in the above equation, we obtain a quadratic in x which is solved to yield

$$x = \sqrt{N} = \frac{-\alpha^{'}\sigma + \sqrt{(\alpha^{'}\sigma)^2 + 4mC}}{2m} \tag{11}$$

It can be easily shown that the alternate solution to the quadratic equation is non-positive and hence is not a valid solution. The reader may now recognize that we have an expression of the form $N = F(q, C)$. It is shown in [NKT92] that $G(N, C)$ is convex in N for $N \leq L(C, m, \sigma)$ where

$$L(C, m, \sigma) = (\frac{-\sigma}{m} + \sqrt{(\frac{\sigma}{m})^2 + \frac{C}{m}})^2. \tag{12}$$

The RGR for this nodal model may then be computed as

$$\begin{aligned} \Phi(q, C) &= \frac{dF(q)/F(q)}{dq/q} \\ &= \frac{2dx/x}{dq/q} \\ &= \frac{2q}{x}\frac{dx}{dq} \end{aligned} \tag{13}$$

[5] The authors [G+91, GG92] consider the buffer overflow probability and not the packet loss probability but it is believed that the two quantities might be close for the system in consideration [Mit] (see also [SW86])

where

$$\frac{dx}{dq} = (\frac{-\sigma}{2mq\alpha'})(-1 + \frac{\alpha' \sigma}{\sqrt{(\alpha' \sigma)^2 + 4mC}}). \tag{14}$$

Simplifying, one obtains

$$\Phi(q, C) = \frac{2\sigma/\alpha'}{\sqrt{(\alpha' \sigma)^2 + 4mC}}. \tag{15}$$

Figures 3 shows the RGR for this model as a function of the link capacity and the QOS requirement It can be seen that the values of the RGR are relatively low for high link capacities and low loss values, a regime of interest in future high-speed networks. This can be also inferred from the equation for the RGR above, since $\Phi(q, C) \rightarrow 0$ as $C \rightarrow \infty$ or $q \rightarrow 0$. Note that this does not imply that the relative gain in traffic load also approaches zero, it depends on the value of $\Delta q/q$ as well.

The second approximation in Guerin *et al.* [G$^+$91, GG92] relies on the fluid-flow model of [AMS82]. An expression for the equivalent capacity of an on-off fluid source [G$^+$91, Equation. 2] is developed by considering the asymptotic behaviour of the tail of the queue-occupancy distribution. While Guerin *et al.* derive expressions for the equivalent capacity of a fluid source using heuristic (intuitive) arguments, a more rigorous derivation of identical expressions appears in [EM, GH91]; the reader is referred to these papers for details.

Consider identical M2 sources being superposed onto a link of capacity C units and infinite queueing space. The number of fluid sources, N, that can be supported while meeting a buffer overflow criteria (as previously, this may be considered to be an approximation for the packet loss probability), $P(B > x) \leq q$, where B denotes the queue occupancy is

$$N = F((q, x), C) = \frac{C}{\alpha(\zeta)} \tag{16}$$

where $\alpha(\zeta)$ is the equivalent capacity for the single fluid M2 source and is given as [G$^+$91, EM, GH91]

$$\alpha(\zeta) = \frac{\zeta\gamma + \mu + \lambda + \sqrt{(\zeta\gamma + \mu - \lambda)^2 + 4\lambda\mu}}{2\zeta} \tag{17}$$

and $\zeta = log_e(q)/x$. The RGR for this model can then be easily derived and shown to be

$$\Phi(q, x) = \frac{-d\alpha(\zeta)/d\zeta}{x\alpha(\zeta)}. \tag{18}$$

It can be shown that $F(\cdot)$ is a convex function of q and hence the RGR in this case is a lower bound for the relative gain in traffic load for unit gain (increase) in QOS. For small values of Δq, however, one can interpret the RGR as a good approximation for the relative gain (see earlier discussion on properties of the RGR).

Interestingly, the RGR in this case is independent of the link capacity (C). Figure 4 shows the RGR for this model and it can be seen that the values are relatively low with a maximum RGR of about 5% for a loss probability of 1×10^{-03} and $x = 50$ kbits. Finally, the RGR values for the above fluid approximation and the earlier Gaussian approximation for the voice multiplexer suggest

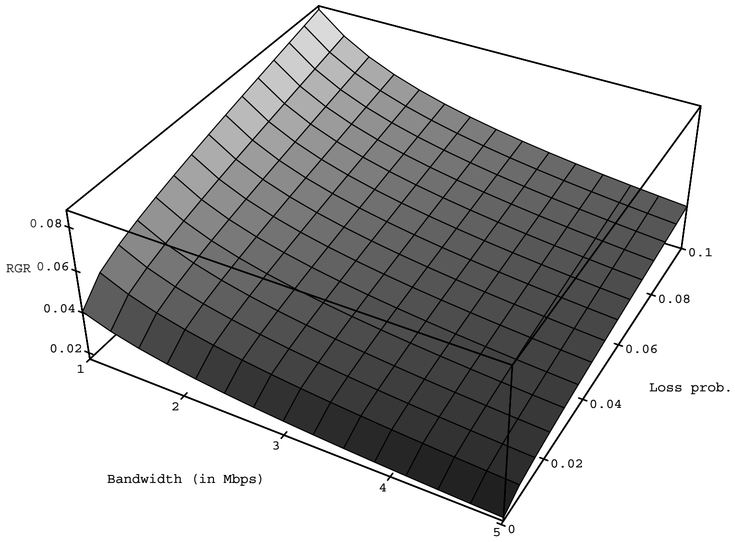

Figure 3: Relative Gain Ratio for the loss QOS metric and voice source multiplexer - Gaussian bit rate approximation

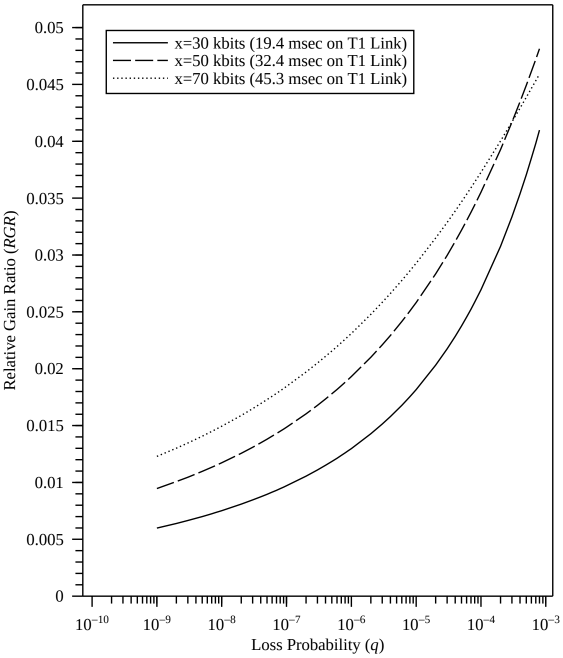

Figure 4: Relative Gain Ratio for the loss QOS metric and voice source multiplexer - Fluid approximation

that the traffic load supportable by a sophisticated QOS allocation policy in a network with voice traffic may not be very large compared to that by a simple QOS allocation policy.

In subsequent sections, we present two network models and evaluate the efficiency of QOS allocation policies for this model. The results for these models help serve as validation of the RGR as a potentially useful metric for evaluating QOS allocation policies.

5 General Model

In this section, we present a simple network model for which we study allocation policy performance. This simple model is a special case of the more general model presented in Section 2.

Figure 5 illustrates our model network consisting of a single source-destination pair of nodes between which connections are setup. The number of nodes in this network is $V = h$ and these nodes are labeled $1, 2, \cdots, h$. Since there is only a single route (path) in this model, we will drop the route notation ω in this section. This simple model captures a number of features of the general problem posed in Section 2, and allows us to solve exactly for the number of connections supportable under various policies.

We will assume in our network models that a connection's traffic characteristics are unaffected as the traffic travels through the network. For example, packet loss at upstream nodes may result in downstream nodes receiving traffic whose characteristics are very different from those at the network "edge". We ignore any such changes in connection characteristics. One exception is the case of the M1 source where we thin the source as it proceeds through the network (see also Section 5.2). The assumption may be justified on the basis of the fact that the small packet delays in future high-speed networks and projected network operation under conditions of low loss will result in near preservation of connection traffic characteristics as it proceeds through the network [O$^+$91] (see also Kelly [Kel91, pp. 12]).

We first consider an equal QOS allocation policy (EQ) which assigns an equal amount of the end-to-end QOS for a connection to each node, i.e., $q_i = q_j = q \ \ \forall i, j$. The value of q is determined by the relation

$$\Gamma(q_1, q_2, \cdots, q_h) = Q. \tag{19}$$

The maximum number of connections supportable under the EQ policy is then:

$$N_{eq} = \mathrm{Min}_{1 \leq i \leq h} \ \ F_i(q, R_i). \tag{20}$$

For this network model, we also determine the the allocation of q_is that maximize the number of connections. We will refer to the (virtual) policy corresponding to this optimal allocation as the optimal policy (OPT) and the number of connections supportable under it, N_{opt}, as the optimal number of connections. We can formulate this problem of determining the optimal number of connections as (the reader is referred to [NKT92] for an alternate but equivalent formulation):

$$\begin{aligned} &\text{Maximize} \quad N \\ &\text{Subject to} \quad \Gamma(G_1(N, R_1), G_2(N, R_2), \cdots, G_h(N, R_h)) \leq Q \end{aligned} \tag{21}$$

For the metrics of interest in this paper, it can be shown that the optimal solution is [NKT92]

$$N_{opt} = \{N : \ \ \Gamma(\cdots) = Q\}. \tag{22}$$

5.1 Upper Bound for Relative Policy Performance

In Section 4, we employed the RGR to make qualitative predictions regarding the performance of QOS allocation policies. In this section, we describe how the RGR may be employed to compute a quantitative upper bound for the relative performance of any QOS allocation policy with respect to the EQ policy. We compute this bound in the context of the network model of the previous section. Typically, analytical bounds are useful when exact analytical computations are either intractable or considerably complex. Our purpose in computing a bound is, however, different. We will first seek to relate the *RGR*, via the bound, to the relative performance of allocation policies in this network model. Second, we will see that the upper bound is indeed realizable when there is a single "bottleneck" node in the tandem model, i.e., that all nodes except the bottleneck have infinite resources. This bound, thus, represents the highest possible relative gain realizable in the given model with respect to the EQ allocation policy.

Consider first an arbitrary allocation policy, π, that assigns q_i of the end-to-end QOS, Q, to node i. The maximum number of connections supportable under this policy is then

$$N_\pi = \mathrm{Min}_{1 \leq i \leq h} \ F_i(q_i, R_i). \tag{23}$$

We are now interested in the maximum relative improvement (over policy π) in supportable load that can be obtained. Since the maximum QOS that can be allocated to any node is bounded by Q, the relative improvement is *bounded* by

$$B_1 = \frac{\mathrm{Min}_{1 \leq i \leq h} \ F_i(Q, R_i) - N_\pi}{N_\pi}. \tag{24}$$

Note that no policy that realizes this upper bound may exist. However, when the resource capacities at all nodes except for a single bottleneck node are infinite, the optimal policy will allocate all of the end-to-end QOS to the bottleneck node and the above upper bound will indeed be realized.

Alternatively, the above upper bound can be computed based on our earlier RGR computation. This will establish a more direct relationship between the performance of QOS allocation policies in "real" networks and the nodal RGR value. Note that

$$\frac{F_i(Q, R_i) - F_i(q_i, R_i)}{F_i(q_i, R_i)} \leq \Phi(q_i, R_i) \frac{Q - q_i}{q_i} \tag{25}$$

when $G(\cdot)$ is a convex function. Hence

$$F_i(Q, R_i) \leq F_i(q_i, R_i) \Phi(q_i, R_i) \frac{Q - q_i}{q_i} + F_i(q_i, R_i). \tag{26}$$

Substituting the above in equation (24) yields a new upper bound, B_2, based on the RGR value. In the rest of this paper, we refer to the bound in equation (24) as the non-RGR-based bound and the above as the RGR-based bound. Finally, we note that for either bound we could choose an arbitrary node rather than minimizing over all nodes as in equation (24); this would yield a looser upper bound.

5.2 Allocating the loss QOS metric

In this section, we take the QOS metric to be the packet loss probability and solve for the maximum number of connections supportable under the EQ and OPT allocation policies.

Before proceeding to the analysis, we consider some modification of the general notation of Section 2 for the particular QOS metric and models considered in this section. We also outline two of the assumptions used in the analysis. We denote the end-to-end loss QOS requirement as $Q = b$. Further, we denote by μ_i and k_i the nodal bandwidth and buffer capacity at node i respectively. Note that $R_i = (\mu_i, k_i)$ is a multi-component vector in this case.

For the loss metric, we consider both the M1 and M2 source models for the source traffic. We assume, for simplicity, that the loss processes at the nodes are independent of each other. We also assume, as noted earlier, that a M1 source remains a M1 source in the interior of the network. However, we thin the M1 source in accordance with the losses suffered at the respective nodes as it proceeds through the network [SR$^+$90]. This thinning is not possible for the M2 source in any reasonable fashion, i.e., without altering the source model itself, and hence the M2 source will retain its exact characteristics as it proceeds through the network. We conjecture that, for the generally low loss probabilities of practical interest, these assumptions will not seriously impact the qualitative nature of the following results.

For both the M1 and M2 models, N_{eq} and N_{opt} can be easily determined as in section 5. For the M1 model, however, it is simpler to maximize the traffic intensity than the number of connections (see [NKT92]). For the case of M1 source thinning, the reader is referred to [NKT92] for details of additional considerations.

5.2.1 Poisson traffic model: Results

We now consider several numerical examples to gain further insight into the actual performance of the EQ and OPT allocation policies when the QOS metric is the packet loss probability. We discuss allocation policy performance primarily for the case of no thinning of the sources. The figures, however, also plot policy performance for the case of source thinning. We notice that there is no great difference in allocation policy performance when we account for thinning and when we do not (at least for small end-to-end loss QOS values). Then we compute the upper bound for relative policy performance. We take $\lambda_s = 1$ in all examples.

Two-hop ($h = 2$) tandem queues, No thinning

We first assume that both nodes have the same buffer capacity, $k_1 = k_2 = k$ but that the bandwidths at the two nodes are $\mu_1 = 1000$ and $\mu_2 = 5000$ units respectively. The relative difference in the performance of the two policies is shown in Figure 6. It can be seen that the difference between the EQ and OPT policies is not that significant. It can be, however, seen that the two policies begin to differ significantly in performance as the end-to-end loss grows in value or as the buffer capacity decreases. It is interesting to note the generally similar performance of the two policies for low loss probabilities in spite of a $1:5$ ratio of available bandwidth at the two nodes. When there is such a significant imbalance in resources in the network one might have expected a naive policy such as the EQ policy to perform considerably worse.

We next consider the case that both nodes have the same link capacity but different buffer capacities. Specifically, we set $k_1 = 50$ and allow k_2 to take on different values. The link capacity

End-to-End Loss (b)	Percent allotted to bottleneck node (OPT)
1×10^{-03}	100.00
5×10^{-03}	99.98
1×10^{-02}	99.98
5×10^{-02}	61.81
1×10^{-01}	53.30

Table 1: Fraction of end-to-end loss allocated to bottleneck node in OPT policy

at both nodes is assumed to be 1000 units. Figure 7 shows the absolute performance of the two policies and Figure 8 the relative gain. It can be seen that the relative gain increases first and then decreases. Table 1 shows the fraction of the end-to-end QOS assigned to the bottleneck node for this example. The table indicates that the imbalance in buffer space is not large enough for the OPT policy to assign a large share of the end-to-end loss (for all loss QOS values) to the bottleneck node and significantly improve over the EQ policy performance.

Five-hop ($h = 5$) tandem queues, No thinning

Figure 10 shows the relative performance of the EQ and OPT policies for the five-hop network model when we do not account for upstream losses. We see that the relative gains in carried load, in general, are somewhat higher than for the two-hop models; the gains being of the order of $4 - 8\%$ in the five-hop case as compared to the $2 - 4\%$ gains in the two-hop case.

Upper bound for relative policy performance

As discussed in the previous section, we can compute upper bounds for the relative gain in load of any policy over the EQ policy. First, consider the RGR-based bound. The upper bound for the above numerical examples are shown in Figures 6, 8 and 10 along with the actual relative gain of the optimal policy over the EQ policy. The plus sign on the curves for the upper bound reflects a constraint on the validity of the upper bound, i.e., the upper bound does not hold (in a theoretical sense) to the right of the plus sign on the curves. The fact that the upper bound curve does lie above the curve for the actual gain to the right of the plus sign is merely fortuitous. The constraint arises due to the fact that the loss probability function is not entirely convex with respect to the number of connections (see [NT, NKT92]). The following are two noteworthy features of the computed upper bound:

- The upper bound values are relatively small, i.e., the relative gain in traffic load for any QOS allocation policy over the EQ policy, even in the worst of scenarios (for the EQ policy), is relatively small. This is a direct consequence of the low RGR values for this model.
- The relative gain in traffic load of the OPT policy over the EQ policy is, surprisingly, close to the upper bound in Figures 6 and 10 even though resources imbalances are not significantly large (non-infinite). The exception is, however, Figure 8 (see earlier discussion and Table 1).

In summary, we have seen that for the M1 model the OPT QOS loss allocation policy does not significantly outperform the EQ policy in a regime of practical interest; the result conforms to

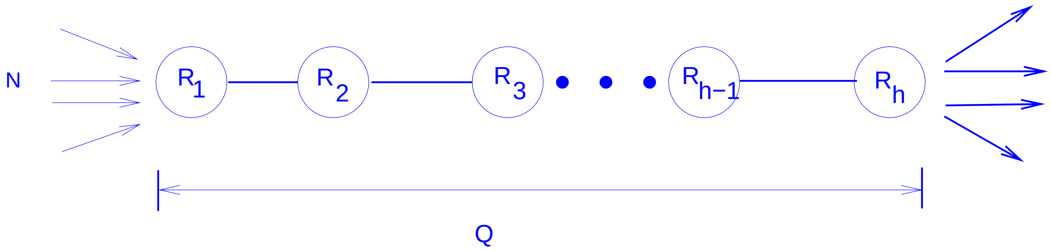

Figure 5: The tandem network model

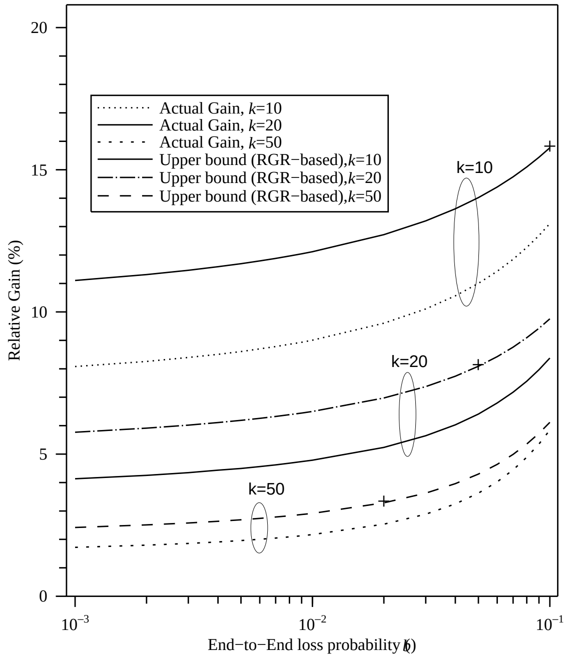

Figure 6: Relative performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two-hop model with M1 sources and identical nodal buffer capacities

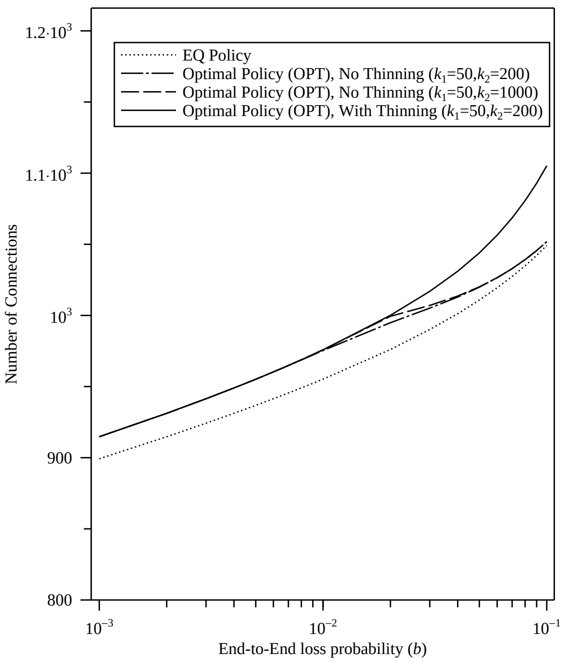

Figure 7: Performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two-hop model with M1 sources and identical nodal bandwidths

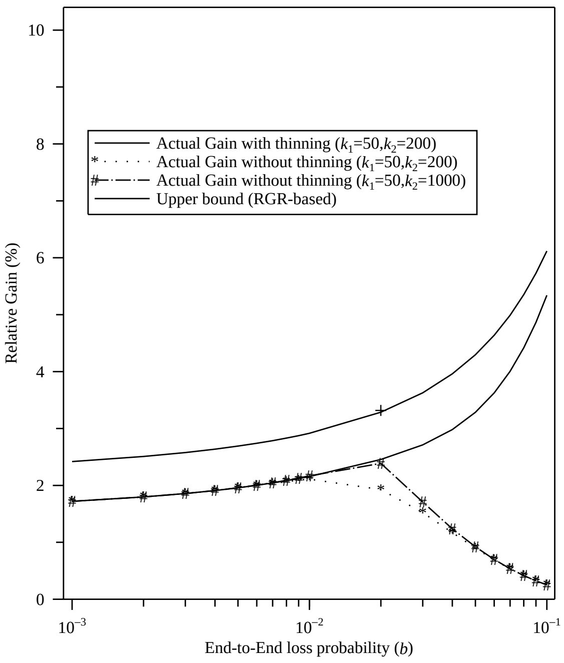

Figure 8: Relative performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two-hop model with M1 sources and identical nodal bandwidths

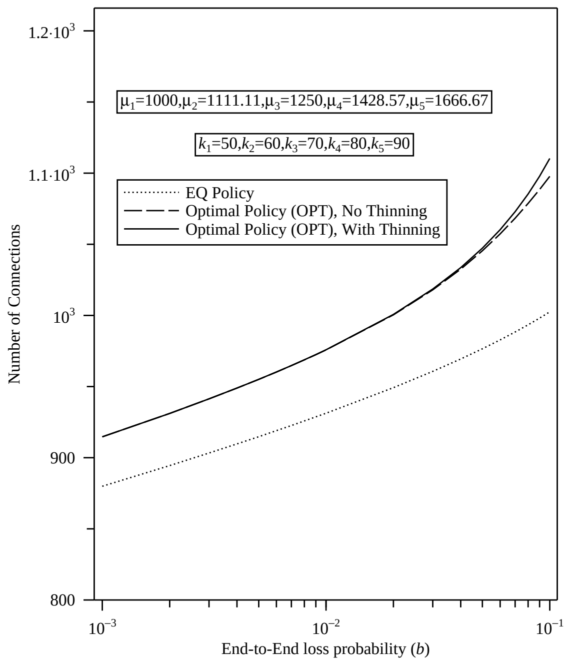

Figure 9: Performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five-hop model with M1 sources

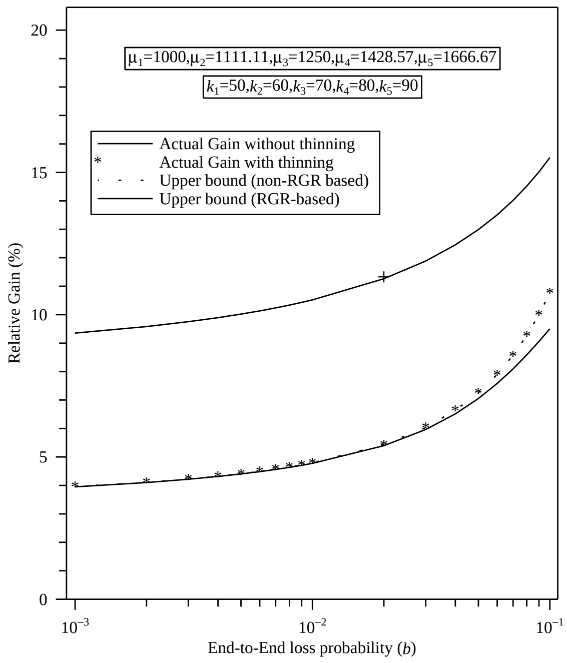

Figure 10: Relative Performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five-hop model with M1 sources

our predictions in Section 4. However, we do observe improvement in the gains in traffic load for the OPT policy with an increase in the number of hops in the source-destination path. Also, an imbalance in buffer capacities appears less detrimental to the performance of the EQ policy than an imbalance in bandwidths at the nodes of the network. Last, we observed in [NKT92] that the OPT policy does somewhat better than the EQ policy when the bottleneck node is closer to the head of the tandem queue than when it is closer to the tail. We now move on to consider the voice traffic model with the same loss QOS metric.

5.2.2 Voice traffic model: Results

We consider the case of $h = 5$. Figure 11 and 12 show the absolute and relative performance of the EQ and OPT QOS allocation policies for this traffic model. It can be seen that the relative gains are relatively low - in the order of $5 - 10\%$ for this example. Also, the gain increases with increasing values of the end-to-end loss, which is as expected from the RGR values computed in section 4.

Finally, we remark on the behavior of the relative gain with increasing number of hops. We noted in the five-hop Poisson model above that the relative gain was somewhat larger than in the two-hop case. It is hence of interest to determine *the effect of the number of hops on the relative gain over the EQ policy.* It can be easily shown that the bound (equation (24)) on the relative gain approaches infinity as the number of hops approaches infinity for a fixed end-to-end loss probability. Let q be the nodal allocation under the EQ policy. Then as $h \to \infty$, we have $q \to 0$. We now have $N_{eq} \to 0$ and hence $B_1 \to \infty$. Since the upper bound is realizable by the OPT policy (see Section 5.1) when all nodes except one (the bottleneck) have infinitely large amounts of resources, we conclude that the OPT policy will perform infinitely better than the EQ policy in this asymptotic regime with a single bottleneck.

6 An Alternate Network Model

In Section 5, we considered a tandem queueing model to investigate the performance of various allocation policies. However, the model was rather simple and did not account for the effects of cross traffic prevalent in general networks. We consider here such a tandem queueing model with cross traffic. While the earlier tandem model explored the effect of physical resource imbalances on allocation policy performance, our focus here is on the effect of uneven traffic loads at network nodes. Since load imbalances can be viewed as resource imbalances, i.e., the node with a higher load may be thought of as a node with a load identical to other nodes but with smaller amounts of physical resources, we expect the qualitative nature of our earlier results to also apply to this more general model. Indeed, we will observe that the performance of various allocation policies for this model also conform to the general trends predicted by the RGR computation of Section 4.

Figure 13 shows the alternate network model to be considered in this section. We generally adopt the notation of previous sections with minor changes. One minor change is necessary when connections traveling over two different paths traverse a node common to the two paths but do not share the resources at that node. For example, in Figure 13 this would the case for node l. In such cases, we distinguish the nodal resources allocated to connections on one path from those on another path by primed quantities, i.e., R_i and $R_i^{'}$ respectively. The solid nodal circles in Figure 6 indicate nodes at which resources are being totally shared among the connections on the

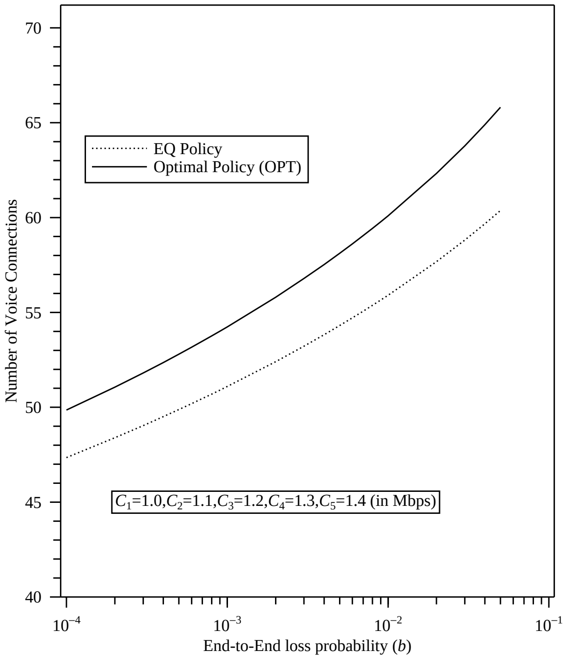

Figure 11: Performance of loss QOS allocation policies as a function of loss QOS requirement - Five-hop model with M2 sources and Gaussian bit rate approximation

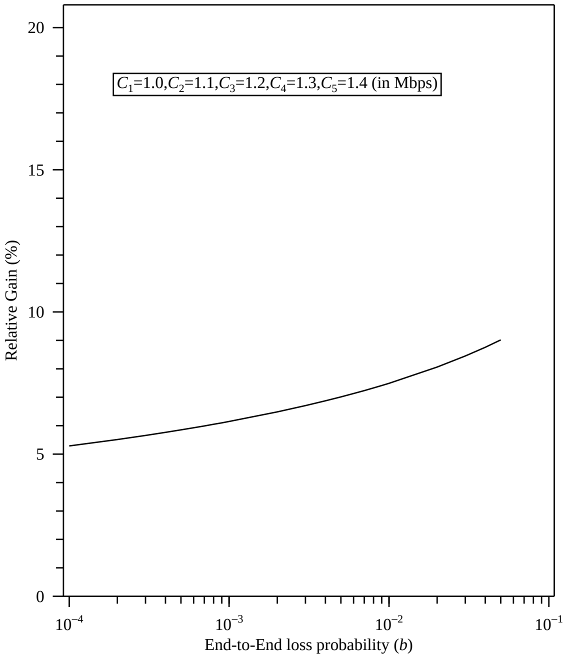

Figure 12: Relative performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five-hop model with M2 sources and Gaussian bit rate approximation

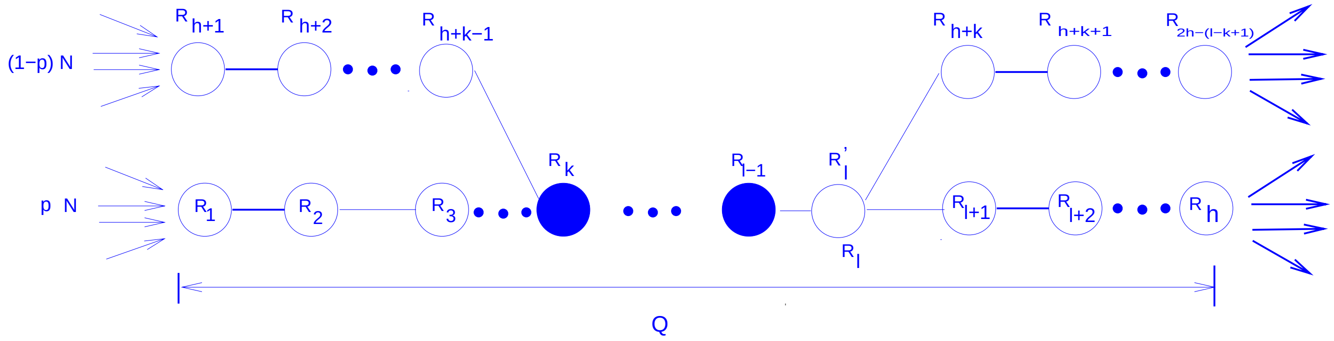

Figure 13: A tandem network model with cross traffic

two different paths. We are now ready to evaluate the performance of the EQ and OPT policies for this model.

Equal QOS Allocation Policy

As previously, we require that each node support an equal portion of the end-to-end QOS Q, i.e., $q_i = q_j = q$ (say), $\forall i, j$. Define

$$N_i = \begin{cases} F_i(q, R_i) & k \leq i \leq l-1, \\ \frac{F_i(q,R_i)}{1-p} & h+1 \leq i \leq 2h-(l-k+1), \\ \frac{F_i(q,R_i)}{p} & 1 \leq i \leq k-1;\ l < i \leq h, \\ \text{Min}(\frac{F_l(q,R_l)}{p}, \frac{F_l(q,R_l^{'})}{1-p}) & i = l, \\ F_i(q, R_i) & k \leq i < l, \end{cases} \tag{27}$$

where N_i is the total number of connections (including those on paths ω_1 and ω_2) that can be supported in the network given the constraints at node i, i.e., a local QOS allocation of q and the fraction $\sum_{\omega: i \in \omega} p_\omega$ of the total number of connections to be supported at node i. The number of connections that can be supported by the EQ policy is then

$$N_{eq} = \text{Min}(N_1, N_2, \cdots, N_{2h-(l-k+1)}). \tag{28}$$

Optimal QOS Allocation Policy

To solve for the optimal allocation policy, we need to find the allocations q_i that maximize the total number of connections N. More rigorously, the optimal policy can be formulated as

$$\begin{array}{ll} \text{Maximize} & N \\ \text{Subject to} & \Gamma_1(G_1(pN), \cdots, G_k(N), \cdots, G_l(pN), \cdots, G_h(pN)) \leq Q_1, \\ \text{and} & \Gamma_2(G_{h+1}((1-p)N), \cdots, G_k(N), \cdots, G_l((1-p)N), \cdots, G_{2h-(l-k+1)}((1-p)N)) \leq Q_2. \end{array} \tag{29}$$

Note that we have allowed for different QOS values on the two paths. In the examples, however, we will consider only identical values of QOS on the two paths. This is because the number of connections that can be supported at a node with FCFS service is constrained by the most stringent QOS requirement of the connections. The problem of multiple QOS classes is, hence, not very interesting in the context of our nodal model (where we assume FCFS service). For the metrics of interest in this paper, the optimal solution for the above problem is (see [NKT92])

$$N_{opt} = \mathrm{Min}(N_1, N_2), \tag{30}$$

where $N_1 = \{N : \Gamma_1(\cdots) = Q_1\}$ and $N_2 = \{N : \Gamma_2(\cdots) = Q_2\}$.

6.1 Allocating the loss QOS metric: Results

We consider in turn the source models and QOS metrics of previous sections. Our focus in these examples will be more on the effect of imbalances in loading (i.e., disparate p_ωs) than resource imbalances. All of the examples will be four-hop (on each of the two paths) with $k = 2$ and $l = 3$. We denote the two paths by ω_1 and ω_2 and set $p_{\omega_1} = p$ and $p_{\omega_2} = 1 - p$. We first consider M1 sources and a network of $M/M/1/K$ queues. Subsequently, we consider M2 sources and voice multiplexers.

The nodal bandwidths are taken to be $\mu_i = 1000, \forall i$. The buffer capacities at the nodes are taken to be $k_i = 50$, $\forall i$. Further, $\mu'_l = 1000$ and $k'_l = 50$. We hence have identical resources at all nodes. The loading factor is fixed at $p = 0.3$. Figure 14 shows the relative performance of the EQ and OPT policies in this case. We see again that the relative gain values are in conformance to that predicted by the RGR for this model. We next examine the effect of loading with the end-to-end QOS value fixed at $5{\times}10^{-02}$. Figure 15 shows the relative performance of the policies. The piecewise continuous nature of the curve is due to the computation of policy performance for a finite set of p values. We observe that the gain is generally very insensitive to load imbalances but a gain in load of about $5 - 7\%$ is available even with identical nodal resources.

The surprising result in Figure 15 is that the OPT policy performs better than the EQ policy more when there is equal loading on the two paths than when there is uneven loading on the two paths. A detailed explanation is provided in [NKT92]. The essence of the argument is that the nodal load imbalances on the "bottleneck" path, when the path load imbalances are large, are small and vice-versa. Since nodal resources are identical, the OPT and EQ policies perform similarly when path load imbalances are large.

Finally, we consider the M2 source. All nodal bandwidths are taken to be identical and correspond to T1 links. The loading factor is fixed at $p = 0.5$. Figure 16 shows the relative performance of the policies in this case. We see again that the relative gain values are in conformance to that predicted by the RGR for this model. We next examine the effect of loading with the end-to-end QOS value fixed again at $5{\times}10^{-02}$. Figure 17 shows the relative performance of the policies. We observe that the gain is generally very insensitive to load imbalances but there is a gain of about $5 - 7\%$ to be had even with identical nodal resources. Also, note that the general performance of allocation policies in this example is very similar to the previous example when we considered the M1 source.

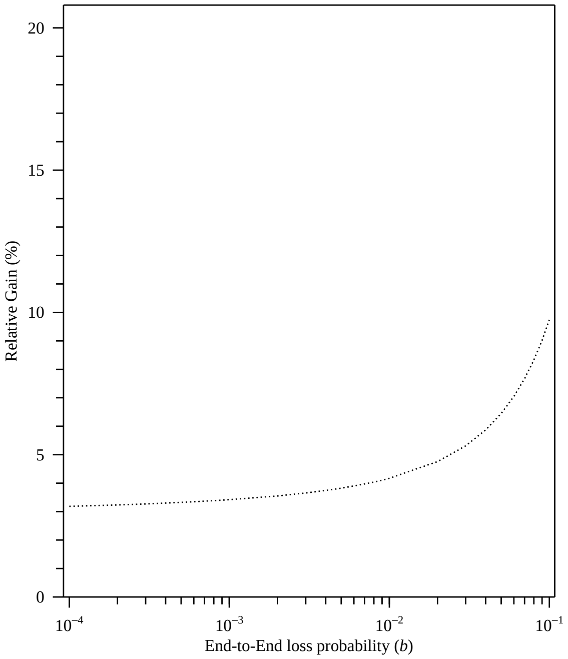

Figure 14: M1 Source Model: Relative performance of the OPT and EQ policies for the alternate four-hop network model with cross traffic

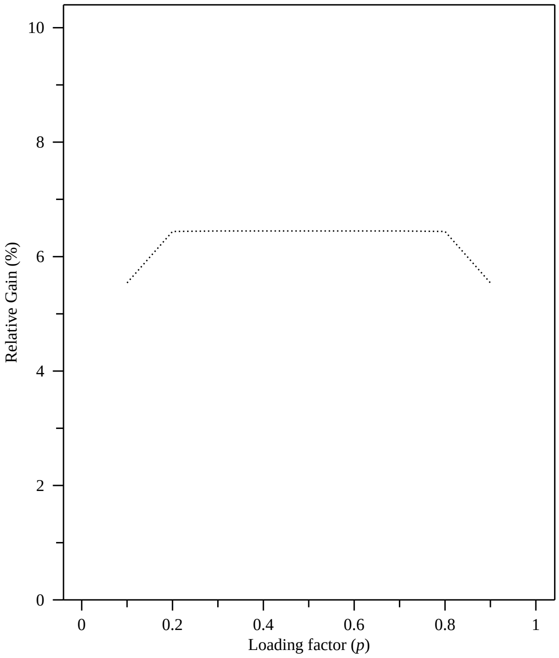

Figure 15: M1 Source Model: Relative performance of the OPT and EQ policies for the alternate four-hop network model with cross traffic

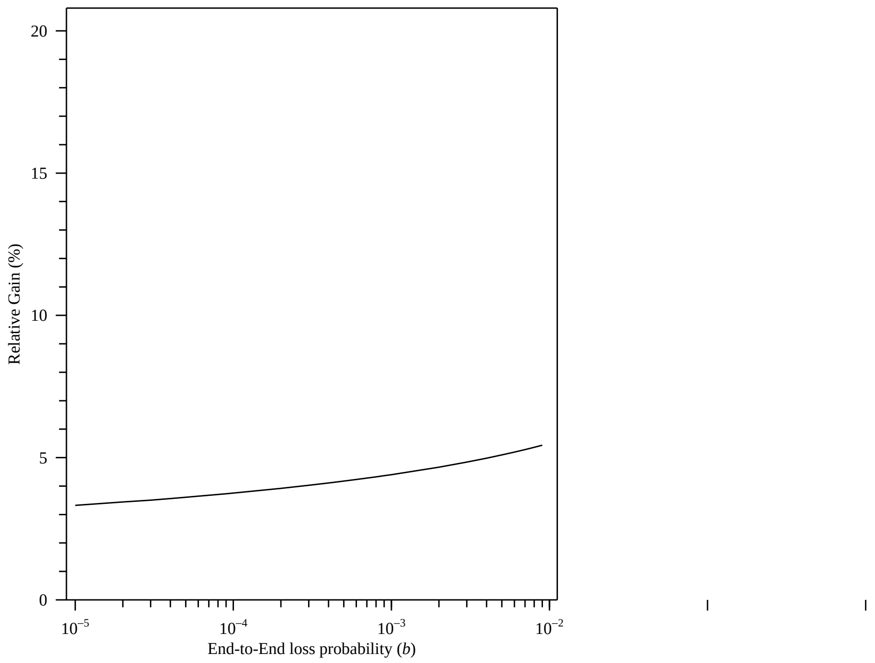

Figure 16: M2 Source Model and Gaussian bit rate approximation: Relative performance of the OPT and EQ policies for the alternate four-hop network model with cross traffic

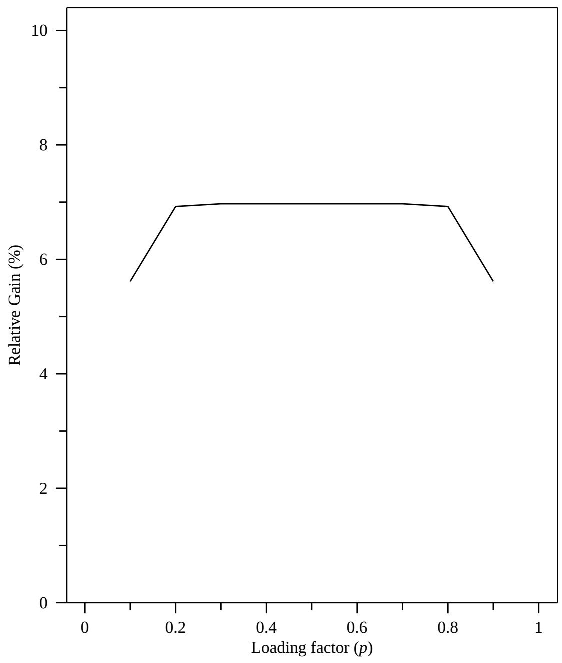

Figure 17: M2 Source Model and Gaussian bit rate approximation: Relative performance of the OPT and EQ policies for the alternate four-hop network model with cross traffic

7 Alternate QOS metrics

The previous sections have exclusively focussed on the packet loss probability as the QOS metric. While the packet loss probability is the QOS metric of interest for future high-speed networks, we briefly digress to consider one other QOS metric, the average packet delay. We see in this case that the *OPT policy significantly outperforms the EQ policy* in the regime of low delay QOS requirements in contrast to our earlier results for the loss metric.

The computation of the RGR for the average delay QOS metric and the M1 source model and the performance of the EQ and OPT allocation policies for the network models of Section 5 and 6 is outlined in [NKT92]. We present only a sample computation here. Consider the tandem network model of Section 5 with five hops, i.e., $h = 5$. Let the link bandwidths at the nodes be $1000, 1100, 1200, 1300$ and 1400 units and $\lambda_s = 1$ unit respectively. Figure 18 shows the relative performance of the EQ and OPT allocation policies. It can be seen, as expected, that large gains in carried load (over the EQ policy) are available for small delay QOS values while for large delay values only small gains are available. Recall that in the case of the loss QOS metric, the OPT policy outperformed the EQ policy only when the *loss QOS values* were *large*. We also find, in the context of the network model with cross traffic, that the relative gain of the OPT policy over the EQ policy is highly sensitive to the loading factor. This is again in contrast to the relative insensitivity of the performance of loss QOS allocation policies to the loading factor. Finally, the RGR values, in this case, accurately predict, as previously, the performance of the EQ and OPT policies.

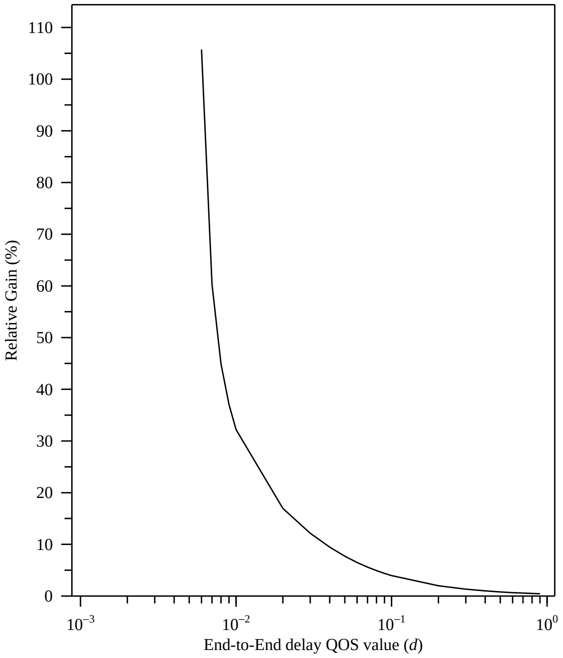

Figure 18: Relative performance of EQ and OPT delay QOS allocation policies as a function of delay requirement - Five-hop model with M1 sources

8 Conclusion

In this paper, we have investigated in detail the allocation of the end-to-end quality-of-service to individual network nodes. We considered two different QOS metrics and traffic models. The primary contribution in this work was to present the RGR as a useful nodal metric for predicting the performance of QOS allocation policies in arbitrary networks. The RGR itself was formulated by appealing to certain fundamental aspects of the allocation problem. Subsequently, it was verified to be a useful metric by investigating the performance of two QOS allocation policies in two simple network models. From a practical standpoint, the following insights into the QOS allocation problem were gained:

- The relative performance of QOS allocation policies is heavily dependent on the particular QOS metric.
- When the average delay is the QOS metric of interest, judicious allocation of the end-to-end QOS yields substantial improvements in carried load over naive allocation policies in the regime of connections with stringent delay requirements.
- When the loss probability is the QOS metric of interest, only small differences in the performance of loss allocation policies were observed in the regime of connections with low loss requirements.
- The relative performance of allocation policies was found to depend to a lesser extent, in comparison to the dependence on the particular QOS metric and its value, on the particular resources that were in imbalance, the number of hops on the source-destination route, the position of the bottleneck node on the source-destination path, the interaction between load and resource imbalances, etc.,.
- The relative performance of QOS allocation policies is of interest only when there are resource or load imbalances in the network.

A number of open interesting issues remain for future research. The complexity of the general problem necessitated a number of simplifying assumptions. One of particular note is that of unmodified connection characteristics in the network models of this paper. While we argued that this might be reasonable for future gigabit networks, it is of interest to evaluate the effects of this assumption on the conclusions of this paper in a "non-asymptotic" low-speed regime (preliminary work [Y^+] appears to indicate that the approximation holds, reasonably, in this scenario as well). In this context, it is useful to view the network nodes as having different characterizations, $G(\cdot)$, for their performance. Hence, these nodes might have non-identical RGRs for identical parameter values. However, the value of the RGR at each node can still be expected to give a useful indication of the expected gain in load at that node due to lower QOS requirements.

The above discussion also suggests a dynamic scheme for QOS allocation. Note that the analysis in this paper considers only a static QOS allocation problem where connection characteristics and routing patterns are known *a priori*. However, it is a dynamic scheme which is of ultimate interest since QOS allocation decisions have to be made in real-time at connection set-up instants. One possible approach is to partition the end-to-end QOS among the nodes on the source-destination path in accordance to the current traffic load, available resources and the corresponding RGR value (for that load and resources). While the former two quantities represent the operating point of the node, the RGR value represents the sensitivity of that point to changes in the QOS allocation. Hence, for example, when two nodes have identical operating points, the node with the larger RGR value will be assigned a larger (looser) portion of the end-to-end QOS. The details of such a scheme remain a topic for further investigation.

Other areas for future research include considering alternate source models such as the (σ, ρ) characterization of Cruz [Cru91]. Finally, similar techniques (in particular the RGR) may be applied to a slightly different allocation problem discussed in [SR$^+$90] that considers the local allocation of an end-to-end deadline for real-time applications.

References

[AMS82] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell System Technical Journal*, 61:1871–1894, 1982.

[B$^+$91] Andrea Baiocchi et al. Loss performance analysis of an ATM multiplexer loaded with high-speed on-off sources. *IEEE J.Select.Areas Commun.*, 9(3):388–393, April 1991.

[Cru91] Rene Cruz. A calculus for network delay, part II: Network analysis. *IEEE Transactions on Information Theory*, 37:132–141, 1991.

[DL86] John N. Daigle and Joseph D. Langford. Models for analysis of packet voice communications systems. *IEEE J.Select.Areas Commun.*, SAC-6:847–855, 1986.

[EM] Anwar Elwalid and Debasis Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks. Submitted to ACM-IEEE Transactions on Networking, July 1992.

[FV90] Domenico Ferrari and Dinesh Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE J.Select.Areas Commun.*, 8:368–379, April 1990.

[G$^+$91] Roch Guerin et al. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J.Select.Areas Commun.*, 9(7):968–981, 1991.

[GG92] Roch Guerin and Levent Gun. A unified approach to bandwidth allocation and access control in fast packet-switched networks. In *INFOCOM'92*, pages 01–12, 1992.

[GH91] R. J. Gibbens and P. J. Hunt. Effective bandwidths for multi-type uas channel. *QUESTA*, 9:17–28, 1991.

[Gol90] S. J. Golestani. Congestion-free transmission of real-time traffic in packet networks. In *IEEE INFOCOM'90*, pages 527–536, June 1990.

[HL86] Harry Heffes and David Lucantoni. A markov modulated characterization of voice and data traffic and related statistical multiplexer performance. *IEEE J.Select.Areas Commun.*, SAC-4:856–867, September 1986.

[Kel91] F. P. Kelly. Effective bandwidths at multi-class queues. *QUESTA*, 9:5–16, 1991.

[Kur92] Jim Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *ACM SIGMETRICS'92*, pages 128–139, June 1992.

[Mit] Debasis Mitra. Personal Communication, October 1992.

[NK92] Ramesh Nagarajan and Jim Kurose. On defining, computing and guaranteeing quality-of-service in high-speed networks. In *IEEE INFOCOM'92*, pages 8C.2.1–8C.2.10, May 1992.

[NKT91] Ramesh Nagarajan, Jim Kurose, and Don Towsley. Approximation techniques for computing packet loss in finite-buffered voice multiplexers. *IEEE J.Select.Areas Commun.*, 9(3):368–377, April 1991.

[NKT92] Ramesh Nagarajan, Jim Kurose, and Don Towsley. Local allocation of end-to-end quality-of-service in high-speed networks. Technical Report TR 92-77, COINS Dept., UMASS, Amherst, November 1992.

[NT] Ramesh Nagarajan and Don Towsley. A note on the convexity of the probability of a full buffer in the $M/M/1/K$ queue. Submitted to Operations Research Letters, October 1992.

[O$^+$91] Yoshihiro Ohba et al. Analysis of interdeparture processes for bursty traffic in ATM networks. *IEEE J.Select.Areas Commun.*, 9(3):468–476, April 1991.

[SR$^+$90] Henning Schulz-Rinne et al. Congestion control by selective packet discarding for real-time traffic in high-speed networks. In *INFOCOM'90*, pages 543–550, June 1990.

[SW86] Kotikalapudi Sriram and Ward Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J.Select.Areas Commun.*, SAC-4:833–846, September 1986.

[VHN92] Carsten Vogt, Ralf Guido Herrtwich, and Ramesh Nagarajan. HeiRAT: The Heidelberg resource administration technique, design philosophy and goals. Technical Report 43.9213, IBM Germany, 1992. To be presented at *Kommunikation in Verteilten Systemen*, Munich, March, 1993.

[WK90] Gillian M. Woodruff and Rungroj Kositpaiboon. Multimedia traffic management principles for guaranteed ATM network performance. *IEEE J.Select.Areas Commun.*, 8:446, April 1990.

[WN91] Abel Weinrib and Ramesh Nagarajan. Guaranteeing end-to-end quality of service in connection-oriented packet networks. Technical Report TR 91-51, COINS Dept., UMASS, Amherst, June 1991.

[Y$^+$] David Yates et al. Call admission policies for future integrated services digital networks. Computer Science Dept., University of Massachusetts at Amherst, Manuscript in preparation, December 1992.