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

Ramesh Nagarajan[2], James 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 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. It is found 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.

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 Elect. and Comp. Engg.,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 (QOS), 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 service guarantees to subscribers. Due to the availability of high bandwidth fiber and advanced switching technology, future integrated networks will support a variety of novel applications, such as broadcast TV, voice and high speed data transfers. These applications will require that a subscriber be able to specify for a connection an *apriori* quality of service, expressed as constraints on the end-to-end delay of information units, the acceptable levels of information loss, and so on. The network must then be capable of guaranteeing these performance constraints for the duration of the connection.

A number of recent research efforts have focussed on the problem of guaranteeing QOS. Some of this work addresses the issue of local nodal QOS guarantees, while the rest considers end-to-end guarantees. Let us first consider local nodal guarantees. Guerin and Gun [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. Parekh and Gallager [PG92] propose a new service discipline called generalized processor sharing and describe how *worst-case delay guarantees* can be provided at each node for individual connections based on this service discipline. The algorithm in [PG92] for connection admission (rejection) assumes the availability of a delay value for the new session to be satisfied locally. 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 available. However, QOS requirements for applications are only known on an end-to-end basis. In order that any of the aforementioned 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* (in terms of bounds on distributions of pertinent quantities) 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, none of [FV90, VHN92] address in detail the issue of how to assign this excess to the various nodes. In [FV90], for example, the excess value is simply equally distributed among the nodes along the source-destination path.

Thus there is an important and interesting problem of *apportioning the end-to-end QOS values to local nodes* that 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 exclusively 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 from bottleneck nodes while compensating with more stringent local guarantees at other nodes of the network. However, we also uncover, interestingly, scenarios where it does not pay significantly to "optimally" allocate the end-to-end QOS even when there are significant 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 briefly 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 the 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 we allow for interfering traffic. Section 7 considers, briefly, the impact of alternate QOS metrics on QOS allocation policies. A summary of results for the case when the average delay is the QOS metric is presented. 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 *apriori.* We denote by Ω the set of all routes in this network. The routes in this set will be denoted by $\omega_i, \quad i = 1, 2, \cdots, M$ 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

QOS criteria, denoted as Q. In this paper, we assume homogenous 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. We restrict our attention to policies that 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. 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, connection requests arrive and connections terminate dynamically. 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 will not delve into the details of how the local guarantees, q_i^ω, determine the locally supportable load and hence the network suportable 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.

Prior to addressing the general allocation problem presented above, we digress briefly in the following to discuss some of the stochastic models for the network traffic employed in this paper.

3 Traffic models

In this section, we describe two stochastic 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 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 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 will 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 mean and variance of this bit rate. We will denote these quantities by m and σ^2 respectively. For the aforementioned M2 source parameter values we can compute the mean and variance to be

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

4 QOS criteria and Optimal QOS allocation

In this section, we formulate and answer questions regarding 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 this direction is to develop a better understanding of 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 the delay and the particular value of the end-to-end delay to be satisfied is 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 the smaller available bandwidths. This might enable one to support much larger traffic loads than with the EQ policy. Hence, it 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 could expect a large gain in supportable load with sophisticated QOS allocation policies. 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, 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 particular 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 to $q + \Delta q$ from q. In particular, we are interested in determining $\Delta N/N$, the gain in traffic load when the QOS requirement is changed (made less stringent) from q to $q + \Delta q$ locally. 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,

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

where

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

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 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 hence be sufficient. The RGR is thus an useful indicator of allocation policy performance in a network scenario while being 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:

$$q = G(\rho) = (1-\rho)\rho^K/(1-\rho^{K+1}). \tag{5}$$

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:

$$\Phi(K,q) = \frac{d\rho/\rho}{dG(\rho)/G(\rho)}|_{\rho=F(q,K)}. \tag{6}$$

Hence, we have

$$\begin{aligned} \Phi(K,q) &= \frac{(1-\rho)(1-\rho^{K+1})}{(K(1-\rho)-\rho(1-\rho^K))} \\ &= \frac{(1-F(q,K))(1-(F(q,K))^{K+1})}{(K(1-F(q,K))-F(q,K)(1-(F(q,K))^K))} \end{aligned} \tag{7}$$

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 $q \rightarrow 0$ for a fixed value of K, $RGR \rightarrow 1/K$ 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, one can 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., the source M2, 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 approximate analysis of [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 [GG92] considers two different approximations. The reader is referred to [GG92] and Appendix C for details. We provide a general outline of the techniques in the following.

The first approximation in [GG92] is based on modeling the aggregate bit rate of the superposition as a Gaussian distributed random variable whose mean and variance are easily determined from the individual source characteristics. The reader may refer to Appendix C 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 superposed onto a link of capacity C units. The first approximation in [GG92] yields the following relation between N, the number of sources, and q, the loss probability to be satisfied for the sources:

$$C = Nm + \alpha' \sqrt{N} \sigma \tag{8}$$

where

$$\alpha' = \sqrt{-2ln(q) - ln(2\pi)}. \tag{9}$$

The reader is referred to Appendix C 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

$$x = \sqrt{N} = \frac{-\alpha' \sigma + \sqrt{(\alpha' \sigma)^2 + 4mC}}{2m} \tag{10}$$

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 $N = F(q)$. It is shown in Appendix D that $G(N, C)$ is convex in N for $N \leq L(C, m, \sigma)$ where

$$L(C, m, \sigma) = (\frac{-\sigma}{m} + \sqrt{(\frac{\sigma}{m})^2 + \frac{C}{m}})^2. \tag{11}$$

The RGR for this nodal model may then be computed as

$$\begin{aligned} \Phi(C, q) &= \frac{dN/N}{dq/q} \\ &= \frac{2dx/x}{dq/q} \\ &= \frac{2q}{x}\frac{dx}{dq} \end{aligned} \tag{12}$$

where

$$\frac{dx}{dq} = (\frac{-\sigma}{2mq\alpha'})(-1 + \frac{\alpha' \sigma}{\sqrt{(\alpha' \sigma)^2 + 4mC}}). \tag{13}$$

Simplifying, one obtains

$$\Phi(C, q) = \frac{\sigma/\alpha'}{\sqrt{(\alpha' \sigma)^2 + 4mC}}. \tag{14}$$

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

Figures 3 shows the RGR for this model as a function of the link capacity and the QOS requirement (Figure 4 shows the RGR behaviour in a picturesque fashion). 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 $\Phi(C, q) \to 0$ as $C \to \infty$ or $q \to 0$. We hence conjecture 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.

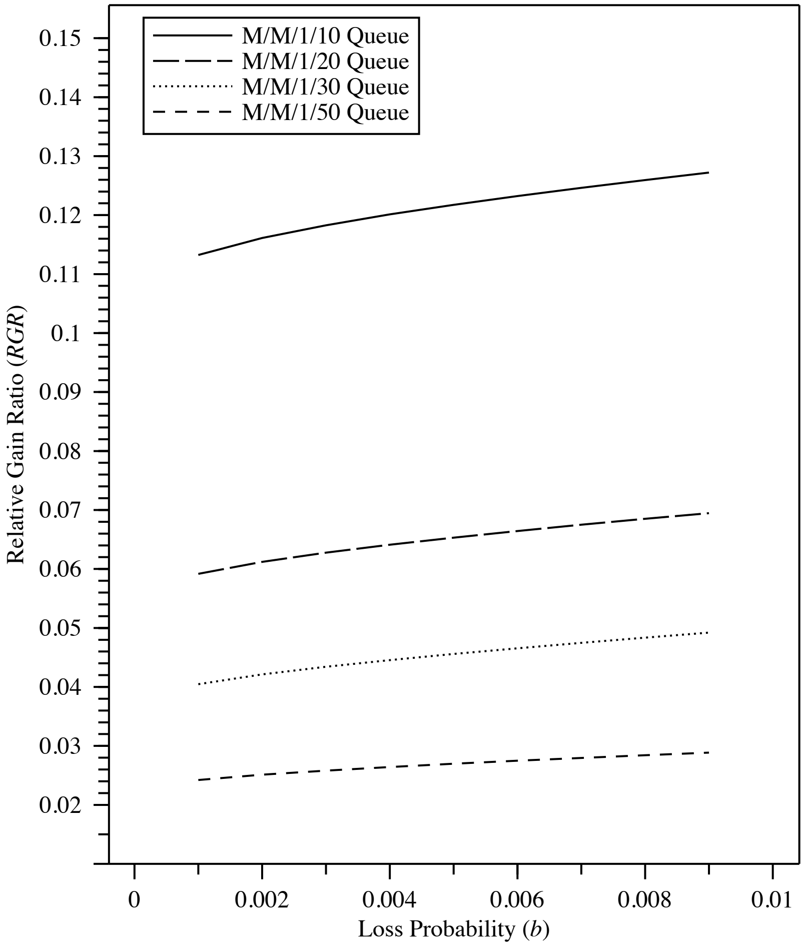

Figure 1: Relative Gain Ratio for the loss QOS metric and the $M/M/1/K$ queue

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.

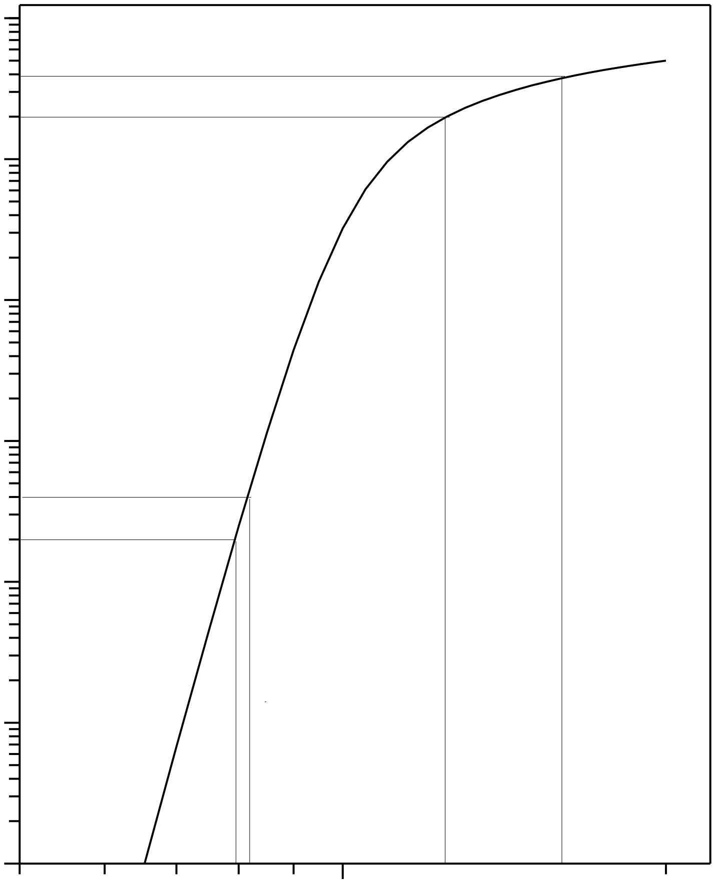

Figure 2: Relative gain in load and its relation to the relative increment (relaxation) of loss in the $M/M/1/30$ queue

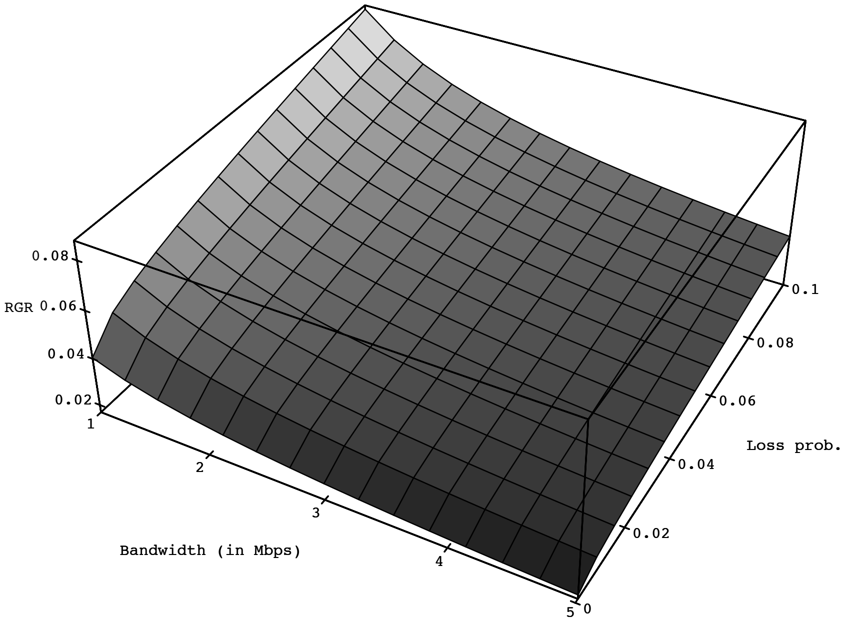

Figure 3: Relative Gain Ratio for the loss QOS metric and voice source multiplexer - Gaussian bit rate approximation

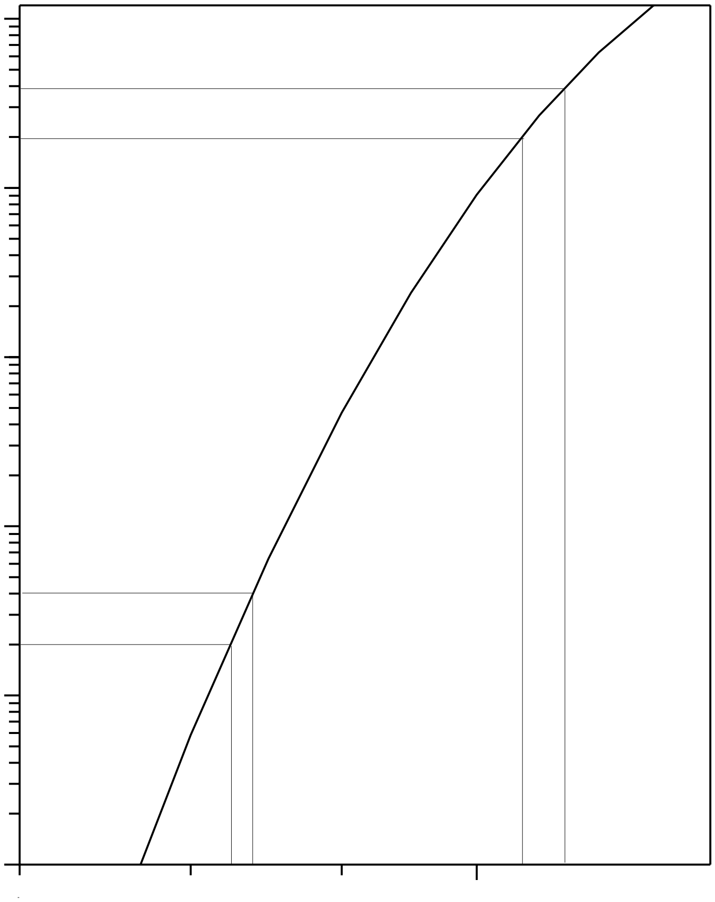

Figure 4: Relative gain in load and its relation to the relative increment (relaxation) of loss in the voice multiplexer with T1 link

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 the essence of the general problem posed in Section 2, and allows us to solve exactly for the number of connections supportable under various policies.

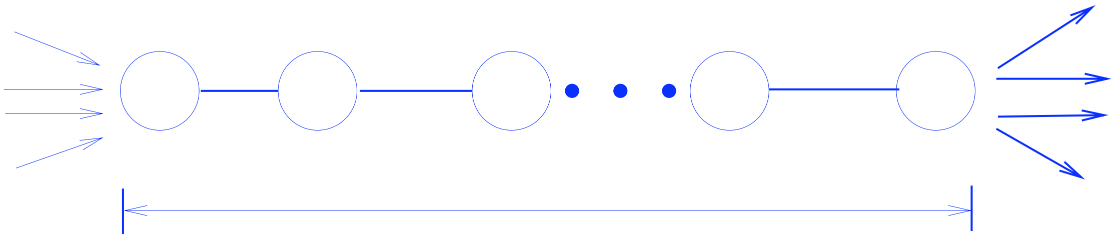

Figure 5: The tandem network model

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

$$\Gamma(q_1, q_2, \cdots, q_h) = Q. \tag{15}$$

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

$$N_{eq} = \text{Min}_{1 \leq i \leq h} \;\; F_i(q, R_i). \tag{16}$$

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 in two different but equivalent ways:

$$\begin{array}{ll} \text{Maximize} & N \\ & \\ \text{Subject to} & \Gamma(G_1(N, R_1), G_2(N, R_2), \cdots, G_h(N, R_h)) \leq Q \end{array} \tag{17}$$

or alternatively,

$$\begin{array}{ll} \text{Maximize} & \text{Min}_{1 \leq i \leq h} \;\; F_i(q_i, R_i) \\ & \\ \text{Subject to} & \qquad\qquad \Gamma(q_1, q_2, \cdots, q_h) \leq Q. \end{array} \tag{18}$$

The analysis in subsequent sections adopts the simpler first formulation to solve for the optimal allocation. The alternative formulation is, however, more natural and potentially more useful in gaining insight into the numerical results in later sections.

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 will 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, hence, represents the highest possible relative gain realizable in the given model with respect to the EQ allocation policy.

In the following derivation of the upper bound, we will assume that the source traffic characteristics are unaffected by the values of the locally-allocated QOS under different policies. For example, when the QOS metric is packet loss it is clear that loss values at upstream nodes reduce the traffic "rate" at downstream nodes. Further, these losses may even result in the downstream nodes receiving traffic whose characteristics are very different from those at the network "edge". We ignore any such changes in connection characteristics. 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]).

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

$$N_\pi = \mathrm{Min}_{1 \le i \le h} \; F_i(q_i, R_i) \tag{19}$$

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

$$\frac{N - N_\pi}{N_\pi} = \frac{\mathrm{Min}_{1 \le i \le h} \; F_i(Q, R_i) - N_\pi}{N_\pi} \tag{20}$$

where N represents the number of connections supportable under a policy that actually realizes this upper bound. Note that no such policy may exist. However, when the resource capacities at all nodes except for a 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. Hence, the upper bound represents the relative gain obtainable in a scenario which is the worst from the perspective of a naive allocation policy such as the EQ policy.

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

$$\frac{F_i(Q,R_i) - F_i(q_i,R_i)}{F_i(q_i,R_i)} \leq \Phi(R_i,q_i)\frac{Q-q_i}{q_i} \tag{21}$$

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

$$F_i(Q,R_i) \leq F_i(q_i,R_i)\Phi(R_i,q_i)\frac{Q-q_i}{q_i} + F_i(q_i,R_i). \tag{22}$$

Substituting the above in equation (20) yields a new upper bound based on the RGR value. In the rest of this paper, we refer to the bound in equation (20) 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 minimize over all nodes as in equation (20) to obtain a looser upper bound.

Before proceeding to the analysis of allocation policies for our model network, we make a comment on the assumption of unmodified connection traffic characteristics underlying our computation of the upper bound above. If we were to allow for modified connection characteristics depending on local QOS values at upstream nodes then the number of connections supportable at a node for a given local QOS value q is no longer simply $F_i(q,R_i)$. It now depends on some of the upstream node parameters and QOS allocations there as well. Consequently, an upper bound is no longer easily derivable.

5.2 Allocating the loss QOS metric

In this section, we take the QOS metric to be the packet loss probability and consider the EQ and OPT allocation policies.

Before proceeding to the analysis, we consider some notation and outline two of the assumptions used in the analysis.

$Q = b$**:** End-to-End loss QOS requirement.

$R_i = (\mu_i, k_i)\ \ i = 1, 2, \cdots, h$**:** Nodal bandwidths and buffer capacities at the nodes respectively.

For the loss metric, we will 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 that the source traffic characteristics are unmodified as the source traffic proceeds through the network i.e., a M1 source remains a M1 source in the interior of the network. However, we will 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, however, 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. First, we consider the M1 source model and subsequently the M2 traffic model.

5.2.1 Poisson traffic

With the source traffic model M1, the node model is the classical $M/M/1/k$ queue and we have for the loss probability:

$$G(\rho) = (1-\rho)\rho^k/(1-\rho^{(k+1)}). \tag{23}$$

Equal Allocation Policy

Since the nodes equally share the burden of providing the end-to-end loss QOS requirement, we have $q_1 = q_2 = \cdots = q$. Then

$$\begin{aligned} b &= 1 - \prod_{i=1}^{h}(1-q_i) \\ &= 1-(1-q)^h. \end{aligned} \tag{24}$$

Solving for q, we have

$$q = 1-(1-b)^{1/h}. \tag{25}$$

Note that

$$q = \frac{(1-\rho_1)\rho_1^{k_1}}{1-\rho_1^{k_1+1}} = \frac{(1-\rho_2)\rho_2^{k_2}}{1-\rho_2^{k_2+1}} = \cdots = \frac{(1-\rho_h)\rho_h^{k_h}}{1-\rho_h^{k_h+1}} \tag{26}$$

where $\rho_1, \rho_2, \cdots, \rho_h$ are respectively the traffic intensities at the nodes. The above equation can then be solved numerically to obtain $\rho_i,\ \ i = 1, 2, \cdots, h$. We then have

$$N_{eq} = \frac{\mathrm{Min}(\rho_1\mu_1, \rho_2\mu_2, \cdots, \rho_h\mu_h)}{\lambda_s}. \tag{27}$$

We now consider the case of source thinning. In this case, determining the maximum number of connections, N_{eq}, can be formulated as:

$$\begin{aligned} &\text{Maximize} \quad \rho_1 \\ &\text{Subject to} \quad G(\rho_i, k_i) \leq q \end{aligned} \tag{28}$$

where

$$\rho_i = f_i(\rho_1) = \rho_1 \frac{\mu_1}{\mu_i} \prod_{j=1}^{i-1}(1-G(\rho_j,k_j)) \ \ i = 1, 2, 3, \cdots, h. \tag{29}$$

It is shown in Appendix B that $\rho_i, \forall i$ is an increasing function of ρ_1. Hence, the above optimization problem is equivalent to that posed in Appendix E. It is shown there that the optimal solution is

$$\rho_1 = \mathrm{Min}_{1\leq i \leq h}\ \ f_i^{-1}(G^{-1}(q, k_i)). \tag{30}$$

A minor technicality is that while $f_i(\cdot)$ is strictly increasing and hence has a unique inverse, it is bounded for downstream nodes. For example, the utilization at node 2 can never exceed μ_1/μ_2, i.e., $\rho_2 = f_2(\rho_1) \leq \mu_1/\mu_2$. Similarly, bounds can be derived for other nodes. Let us denote these bounds by $u_i, \;\; i = 1, 2, \cdots, h$. Then we take

$$f_i^{-1}(x) = \infty \;\; \forall x \geq u_i. \tag{31}$$

Intuitively, the above implies that there exists no finite value of ρ_1 that can cause the loss at node i to exceed the local guarantee under the EQ policy and hence the node plays no role in determining the maximum number of connections under the EQ policy. N_{eq} is now determined as

$$N_{eq} = \frac{\rho_1 \mu_1}{\lambda_s}. \tag{32}$$

Optimal Allocation Policy

We desire to maximize the number of connections that can be supported while satisfying the QOS constraint:

$$1 - \prod_{i=1}^{h}(1 - G_i(\rho_i)) \leq b. \tag{33}$$

In the case that we do not account for upstream losses i.e., we do not thin the M1 source as it proceeds through the network, we have the following relation among the nodal traffic intensities:

$$\begin{aligned} \rho_i &= \frac{\mu_1}{\mu_i}\rho_1 \;\; i = 2, 3, \cdots, h \\ \rho_1 &= \frac{N\lambda_s}{\mu_1} \end{aligned} \tag{34}$$

Hence $1 - \prod_{i=1}^{h}(1 - G_i(\cdot))$ is a function of ρ_1 only. It is shown in Appendix B that $1 - \prod_{i=1}^{h}(1 - G_i(\cdot)) - b$ is an increasing function of ρ_1. Since maximizing N is equivalent to maximizing ρ_1, N is maximized if

$$1 - \prod_{i=1}^{h}(1 - G_i(\rho_i)) = b. \tag{35}$$

The above equation in a single unknown, ρ_1, can be solved numerically. Note that since $1-\prod_{i=1}^{h}(1-G_i(\cdot)) - b$ is an increasing function of ρ_1, there is a unique solution to the above equation.

The optimal number of connections supportable is then

$$N_{opt} = \frac{\rho_1 \mu_1}{\lambda_s} \tag{36}$$

When we do wish to account for upstream losses, we have the following relation among the nodal traffic intensities:

$$\begin{aligned} \rho_i &= \frac{\mu_1}{\mu_i}\rho_1 \prod_{j=1}^{i-1}(1 - G_j(\rho_j)) \;\; i = 2, 3, \cdots, h \\ \rho_1 &= \frac{N\lambda_s}{\mu_1} \end{aligned} \tag{37}$$

Once again it can be shown that $1 - \prod_{i=1}^{h}(1 - G_i(\cdot)) - b$ is an increasing function of ρ_1 (see Appendix B) and hence the unique real-valued optimal solution can be computed numerically. The maximum number of connections that can be supported is again $N_{opt} = \rho_1\mu_1/\lambda_s$.

Results

In this section, we 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. First, we compute the allocation policy performance when there is no thinning of the sources. Then we compute the upper bound for relative policy performance. Finally, we consider the case of source thinning. 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 two 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 i.e., when there is a significant imbalance in resources in the network one expects a naive policy such as the EQ policy to perform significantly 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 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. This result suggests that the imbalance in buffer space is not large enough that the OPT policy can assign a large portion of the end-to-end loss (for all loss QOS values) to the bottleneck node in order to significantly improve over the EQ policy performance.

Five hop ($h = 5$) example, 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

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

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 arrival rate (see [NT]). The position of the plus sign is determined in the following fashion. We define the bottleneck node as the node which supports the smallest number of connections under the EQ policy. In computing the RGR-based bound, we then compute the bound employing this bottleneck node alone rather than minimize over all nodes (see Section 5.1 and Appendix B for explicit expressions). The range of validity of the RGR-based bound is then approximately $b \leq 1/(k+1)$ where k is the buffer size at the bottleneck node. It can be seen that the actual gain follows the general trend of the upper bound in the case of Figures 6 and 10. In Figure 6 the bound is relatively tight while in Figure 10 it is loose. The looseness in the latter case may be attributed to the convexity of the loss metric rather than the inability of the OPT policy to allocate a large portion of the end-to-end loss to the bottleneck node since the non-RGR based bound is tight. In Figure 8, however, the actual gain falls off in relation to the upper bound suggesting that the OPT policy is unable to allocate a significant portion of the end-to-end loss to the bottleneck node. Table 1 confirms this conjecture.

Before proceeding to the case of source thinning, we remark on the behaviour of the relative gain with increasing number of hops. We noted in the five hop example above that the relative gains in this case were generally higher than for the two hop model. It is hence of interest to determine if the relative gain increases with the number of hops. The upper bound for the relative performance of allocation policies tends to infinity as the number of hops approaches infinity (with the end-to-end loss probability fixed). In fact, the RGR-based bound grows linearly in the number of hops (see Appendix B). Recalling the physical interpretation for the upper bound (Section 5.1), we conclude that the OPT policy will perform infinitely better than the EQ policy in this asymptotic regime provided all nodes except one (the bottleneck) have infinitely large amounts of resources (which is intuitively obvious).

Source thinning

Next, we consider the examples analyzed above but we account for upstream losses. Figure 7 shows the performance of the OPT and EQ policies when we thin the M1 source as it proceeds through the network. Figure 8 shows the relative gain. Similarly, Figures 9 and 10 show the relative performance of the OPT and EQ policies for the case of $h = 5$ when we account for upstream losses. It can be seen, in general, that the EQ policy remains more or less unaffected while the OPT policy shows some improvement for large values of the end-to-end loss probability over the OPT policy when we do not account for upstream losses. This is due to the fact that the bottleneck node is at the head of the tandem queue and the EQ policy does not benefit from any source thinning while the OPT policy does (at least for large values of the loss QOS). We will have more to say on this in a nine hop example below.

Finally, we consider the case of $h = 9$. We consider only the scenario where we account for upstream losses. However, we examine the effect of the relative position of the bottleneck node on the performance of the QOS allocation policies. We consider two extreme scenarios. First, the bottleneck node is taken to be at the head of the tandem set of queues. Then we consider it to be at the end of the tandem queueing model. Figure 11 shows the respective relative gains. The parameter values shown in the figure are for the case of the bottleneck node at the head of the tandem queue. For the other case, the first node is placed at the tail of the tandem with all other nodes in the same relative order. It can be seen that the OPT policy does slightly better than the EQ policy in the case that the bottleneck node is the first node in the tandem. This difference in performance is, however, observed only for large loss QOS values. The upstream losses in the case that the bottleneck node is the last node reduce the connection traffic rate at the bottleneck node. Hence the EQ policy does not incur a large "penalty" for allocating the same QOS to the bottleneck node as to the other nodes and performs somewhat comparably to the OPT policy. Finally, we note that the "notch" in the curve for the case of the bottleneck node at the rear is due to a shifting of the bottleneck for the EQ policy away from the node with smallest amount of resources, i.e., some other node becomes the bottleneck beyond the notch.

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

Here we assume the M2 source traffic model. The nodal performance in this case is specified as (see Appendix C):

$$q = G(N, C) = \frac{1}{\sqrt{2\pi}} e^{-(\frac{C - Nm}{\sqrt{N}\sigma})^2/2}. \tag{38}$$

<u>**Equal Allocation Policy**</u>

Since the nodes share equally the burden of providing the end-to-end loss QOS requirement, we have $q_1 = q_2 = \cdots = q$ (say). Then solving for q, we have as previously

$$q = 1 - (1 - b)^{1/h}. \tag{39}$$

Note that

$$q = G(N_1, C_1) = G(N_2, C_2) = \cdots = G(N_h, C_h) \tag{40}$$

where N_i are respectively the number of connections supportable at each node i given a nodal QOS requirement of q. We then have

$$N_{eq} = \mathrm{Min}(N_1, N_2, \cdots, N_h). \tag{41}$$

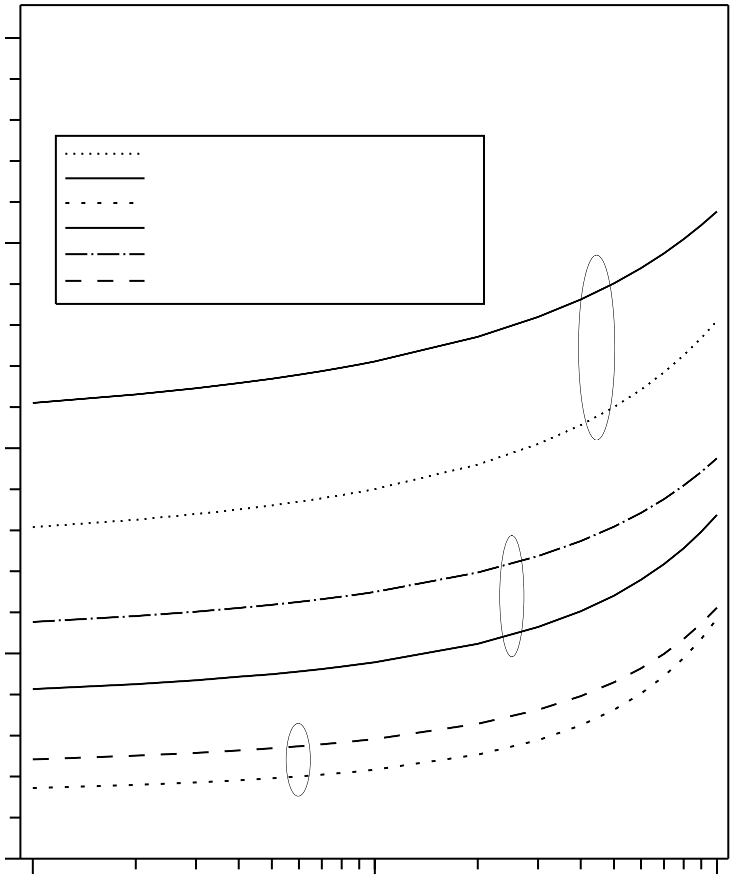

Figure 6: Relative performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two hop model with M1 sources and identical nodal buffer capacities

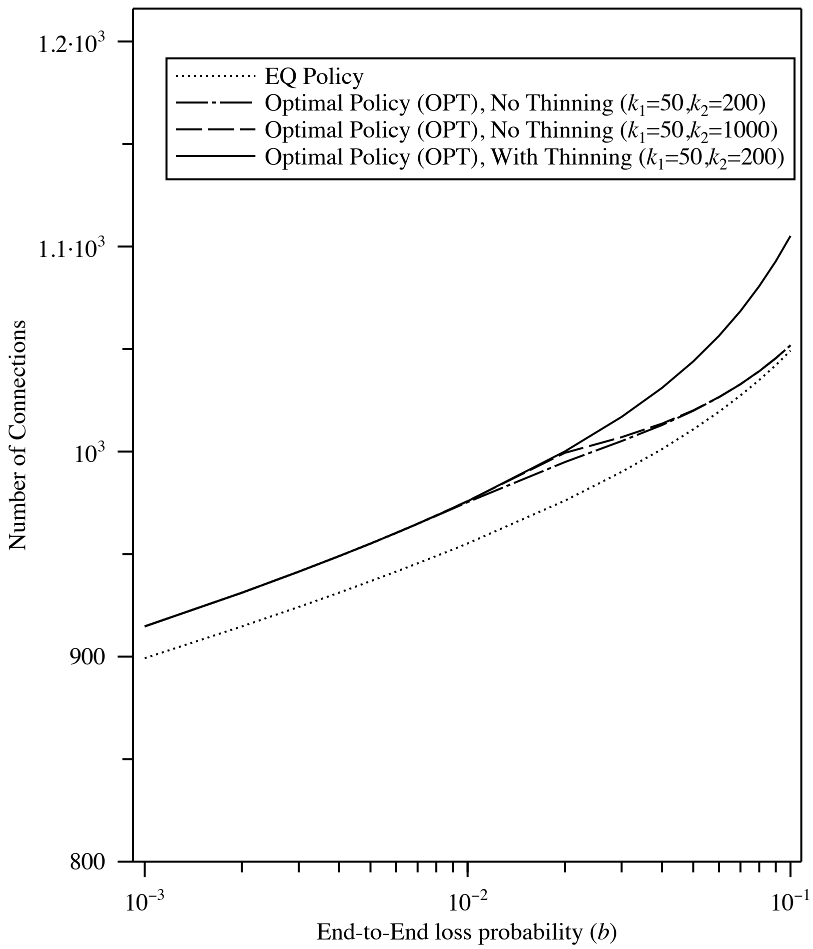

Figure 7: Performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two hop model with M1 sources and identical nodal bandwidths

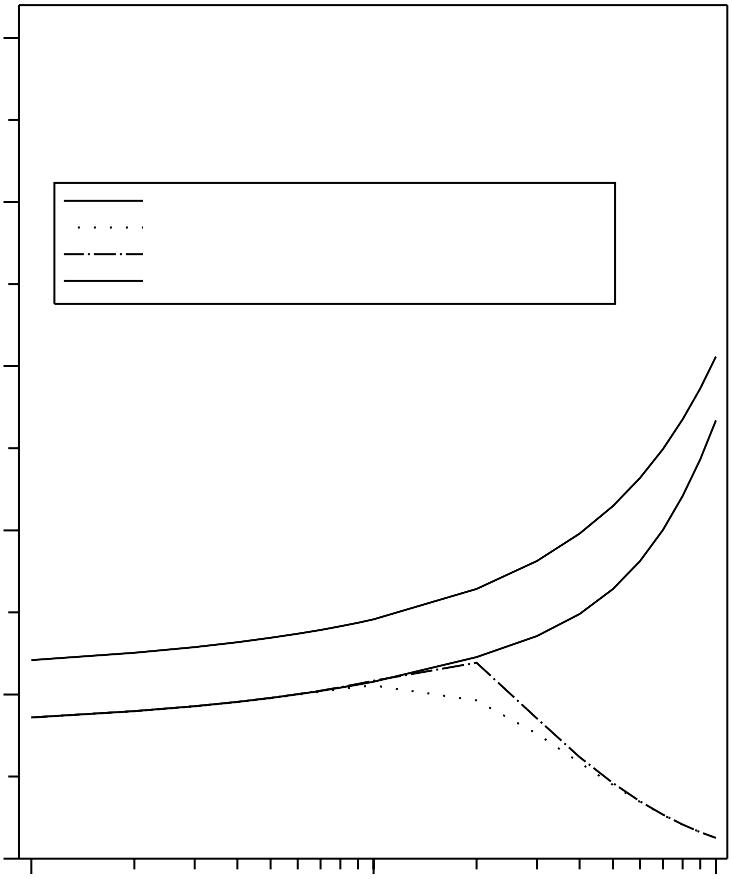

Figure 8: Relative performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Two hop model with M1 sources and identical nodal bandwidths

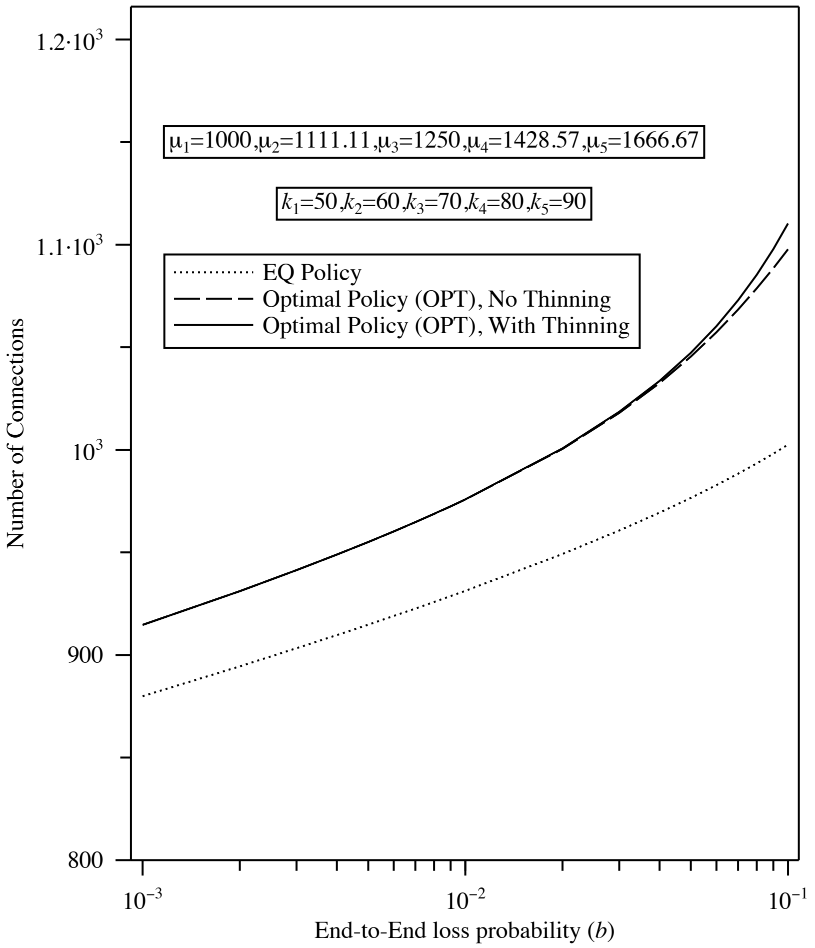

Figure 9: Performance of EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five hop model with M1 sources

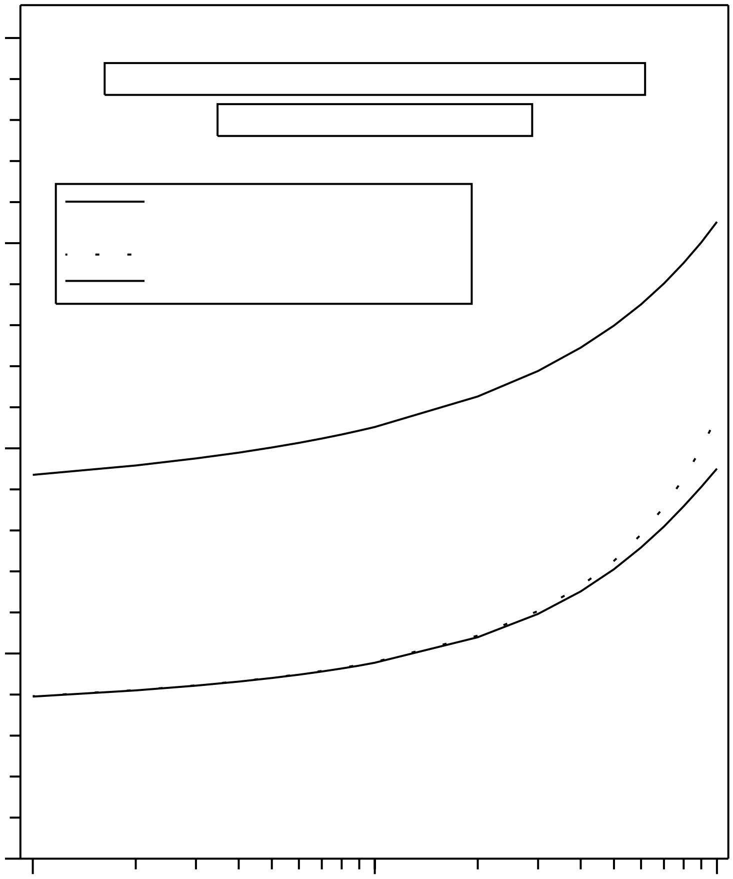

Figure 10: Relative Performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five hop model with M1 sources

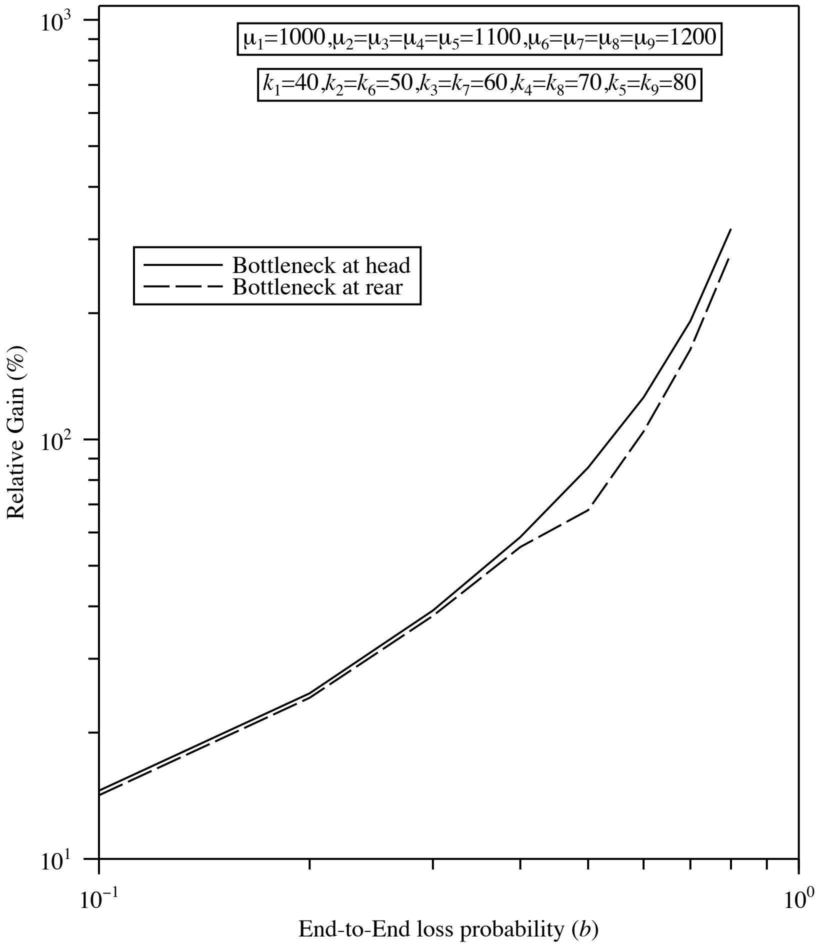

Figure 11: Relative performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement and bottleneck node position (accounting for upstream losses) - Nine hop model with M1 sources

as the number of connections supportable under the EQ policy.

Optimal Allocation Policy

We again desire to maximize the number of connections that can be supported while satisfying the QOS constraint:

$$1 - \prod_{i=1}^{h} (1 - G_i(N)) \leq b. \tag{42}$$

Since the M2 source cannot be thinned in any natural manner, we will consider only the case that we do not account for upstream losses. It is shown in Appendix B that $1 - \prod_{i=1}^{h}(1 - G_i(\cdot)) - b$ is an increasing function of N and hence N is maximized if

$$1 - \prod_{i=1}^{h} (1 - G_i(N)) = b. \tag{43}$$

The above is an equation in a single unknown, N, and can be solved numerically. Again, note the uniqueness of the solution to the above equation since $1 - \prod_{i=1}^{h}(1 - G_i(\cdot)) - b$ is an increasing function of N.

Results

We consider the case of $h = 5$. Figure 12 and 13 show the absolute and relative performance of the OPT and EQ QOS allocation policies for this example. It can be seen that the relative gains are in the order of $5 - 10\%$ for this example and are of the same order of magnitude as in the case of the M1 source model. 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.

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 prevalant 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 will be on the effect of uneven load distribution over the routes of the network. 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 apply to this more general model as well. Indeed, we will observe that the performance of various allocation policies for this model are also in conformance to the general trends predicted by the RGR computation of Section 4.

Figure 14 shows the alternate network model to be considered in this section. We generally adopt the notation of previous sections with minor additions and modifications. One minor addition is necessary when connections travelling 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 14 this would the

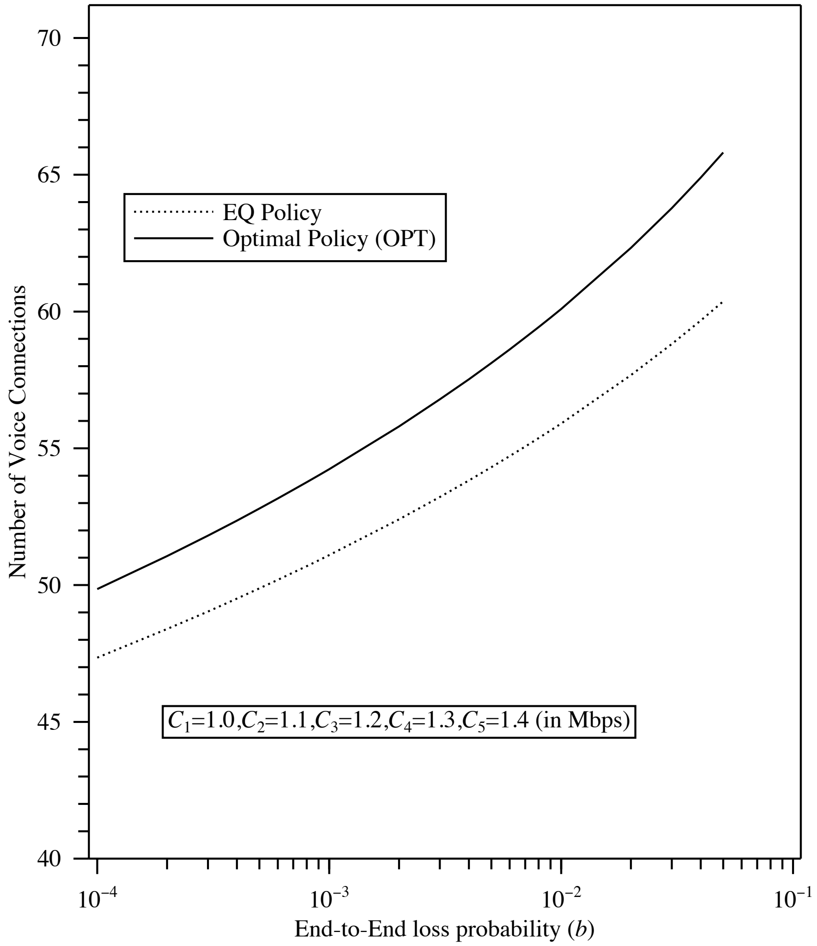

Figure 12: Performance of loss QOS allocation policies as a function of loss QOS requirement - Five hop model with M2 sources

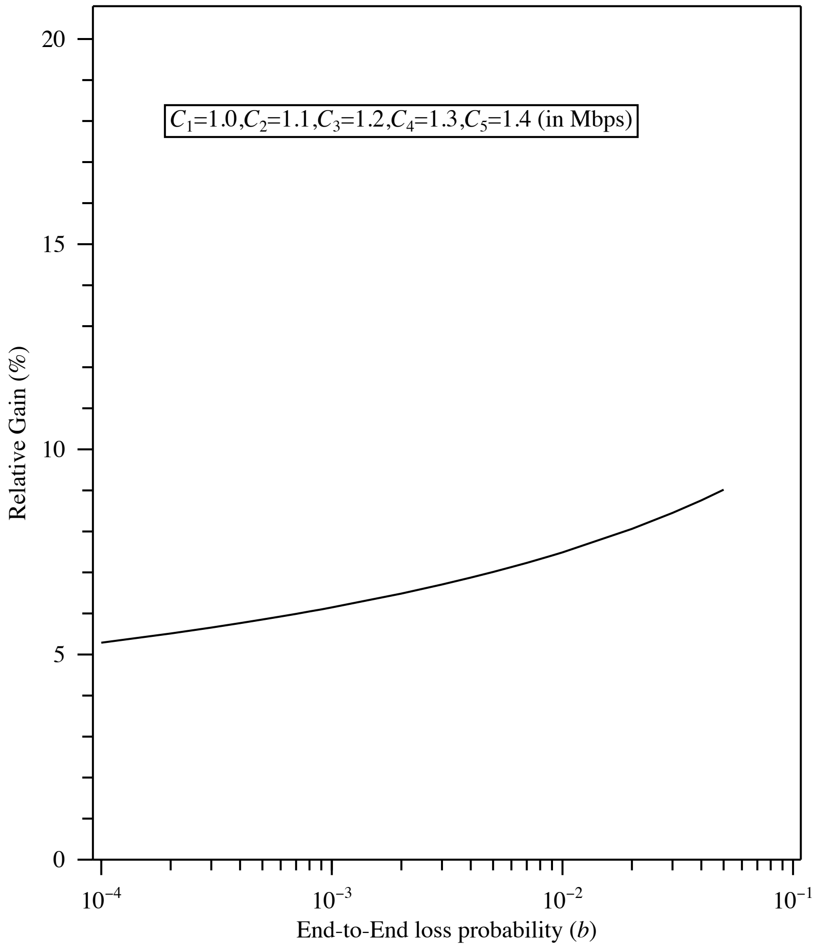

Figure 13: Relative performance of the EQ and OPT loss QOS allocation policies as a function of loss QOS requirement - Five hop model with M2 sources

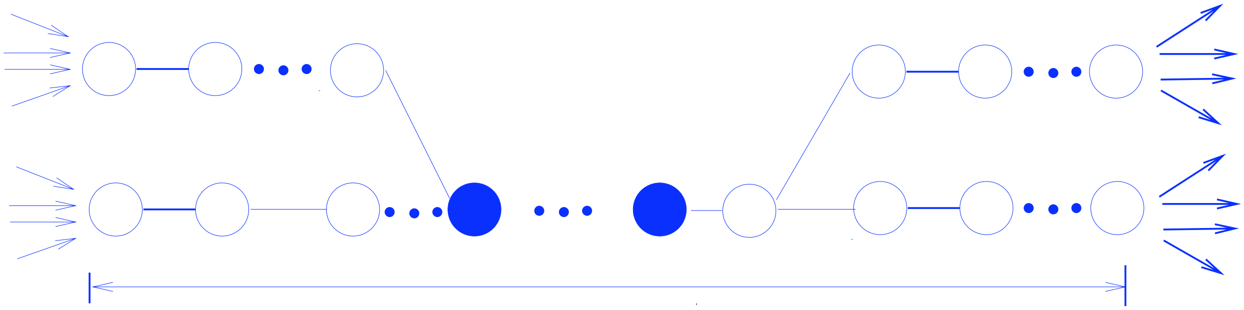

Figure 14: A tandem network model with cross traffic

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 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 a equal portion of the end-to-end QOS Q, i.e., $q_i = q_j = q \text{ (say)}, \forall i, j$. Define

$$N_i = \begin{cases} F_i(q, R_i) & k \leq i \leq l-1 \\ \frac{F_i(q,R_i)}{1-p} & h+1 \leq i \leq 2h-(l-k+1) \\ \frac{F_i(q,R_i)}{p} & 1 \leq i \leq k-1;\ l < i \leq h \\ \text{Min}(\frac{F_l(q,R_l)}{p}, \frac{F_l(q,R_l')}{1-p}) & i = l \\ F_i(q, R_i) & k \leq i < l \end{cases} \tag{44}$$

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 $\sum_{i \in \omega} p_\omega$ 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

$$N_{eq} = \text{Min}(N_1, N_2, \cdots, N_{2h-(l-k+1)}) \tag{45}$$

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

$$\begin{array}{ll} \text{Max} & N \\ \text{Subject to} & \Gamma_1(G_1(pN), \cdots, G_k(N), \cdots, G_l(pN), \cdots, G_h(pN)) \leq Q_1 \\ \textit{and} & \Gamma_2(G_{h+1}((1-p)N), \cdots, G_k(N), \cdots, G_l((1-p)N), \cdots, G_{2h-(l-k+1)}((1-p)N)) \leq Q_2 \end{array} \tag{46}$$

For convenience, we will adopt the notation $F_1(N) = \Gamma_1(\cdot) - Q_1$ and $F_2(N) = \Gamma_2(\cdot) - Q_2$. 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).

We assume that $F_1(N)$ and $F_2(N)$ are increasing functions of N. The optimal solution can then be shown to be (see Appendix E)

$$N_{opt} = \text{Min}(F_1^{-1}(0), F_2^{-1}(0)). \tag{47}$$

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) examples with $k = 2$ and $l = 3$. We denote the two paths by ω_1 and ω_2 and set $p_{\omega_1} = p$ and $p_{\omega_2} = 1 - p$.

Loss Probability

We first consider the M1 source and $M/M/1/K$ nodal model. Subsequently, we consider the M2 source and a set of voice source multiplexers.

The nodal bandwidths are taken to be $R_i = 1000, \forall i$. The buffer capacities at the nodes are taken to be $k_i = 50$, $\forall i$. Hence, we have identical resources at all nodes. The loading factor is fixed at $p = 0.3$. Figure 15 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 at $5{\times}10^{-02}$. Figure 16 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 ps. 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.

The surprising result in Figure 16 is that the OPT policy performs much better than the EQ policy when there is equal loading on the two paths rather than when there is uneven loading on the two paths. To understand this, we need to re-examine the form of the optimal solution for this problem. Let $N_1 = F_1^{-1}(0)$ and $N_2 = F_2^{-1}(0)$. Consider the case of uneven loading on the two paths with $p >> 0.5$, i.e., Path 1 carries a greater portion of the load. We then expect $N_1 < N_2$ (note all nodes have identical resource capacities) and hence $N_{opt} = N_1$. Now N_1 is the optimal solution for the earlier tandem queueing model with a slight difference, viz., all nodes do not carry the same load. The nodes not common to the two paths carry a load of pN while the common nodes carry a load of N. But $pN \approx N$ since $p >> 0.5$. The tandem queue corresponding to Path 1 has, hence, identical resource capacities and near identical loads at each of its nodes. The optimal allocation must, hence, be nearly identical to an equal allocation.

Finally, we consider the M2 source. The nodal bandwidths are taken to be identical and correspond to T1 links. The loading factor is fixed at $p = 0.5$. Figure 17 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 18 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 Poisson source.

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

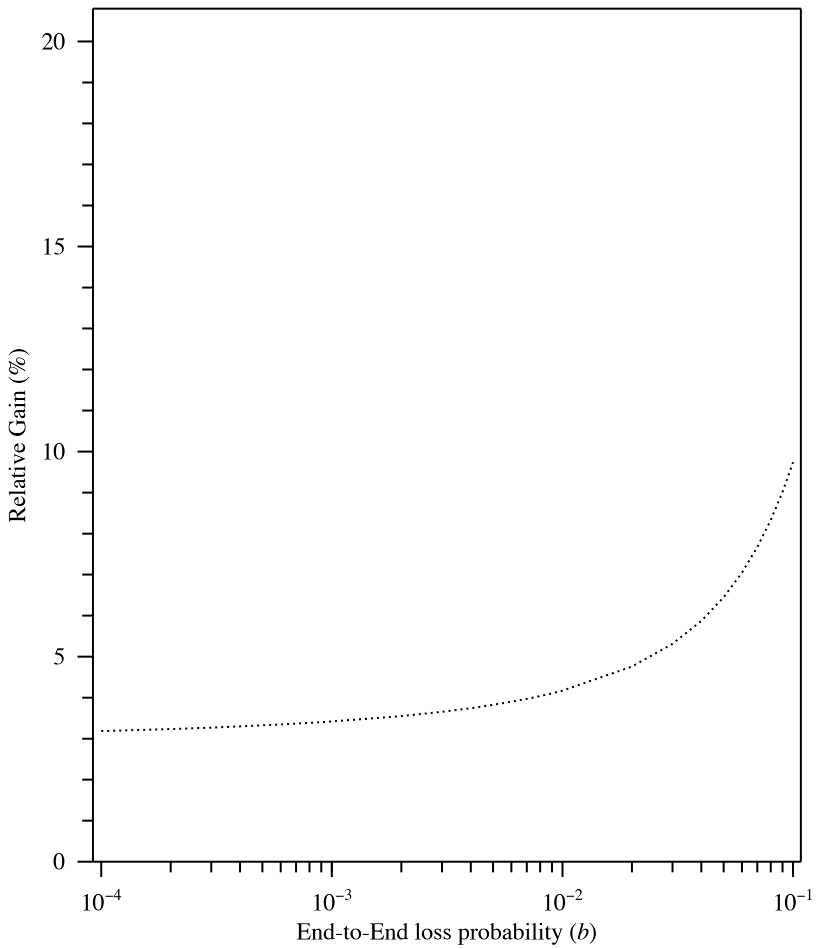

Figure 15: M1 Source Model: Relative performance of the OPT and EQ policies for the alternate four hop network model with cross traffic

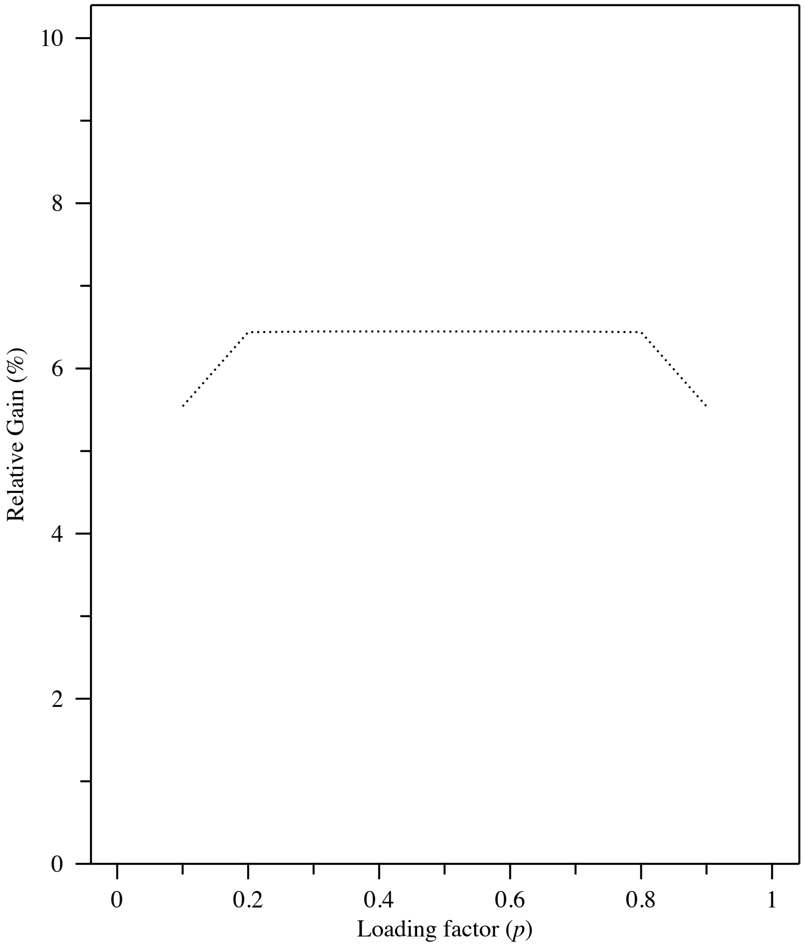

Figure 16: M1 Source Model: Relative performance of the OPT and EQ policies for the alternate four hop network model with cross traffic

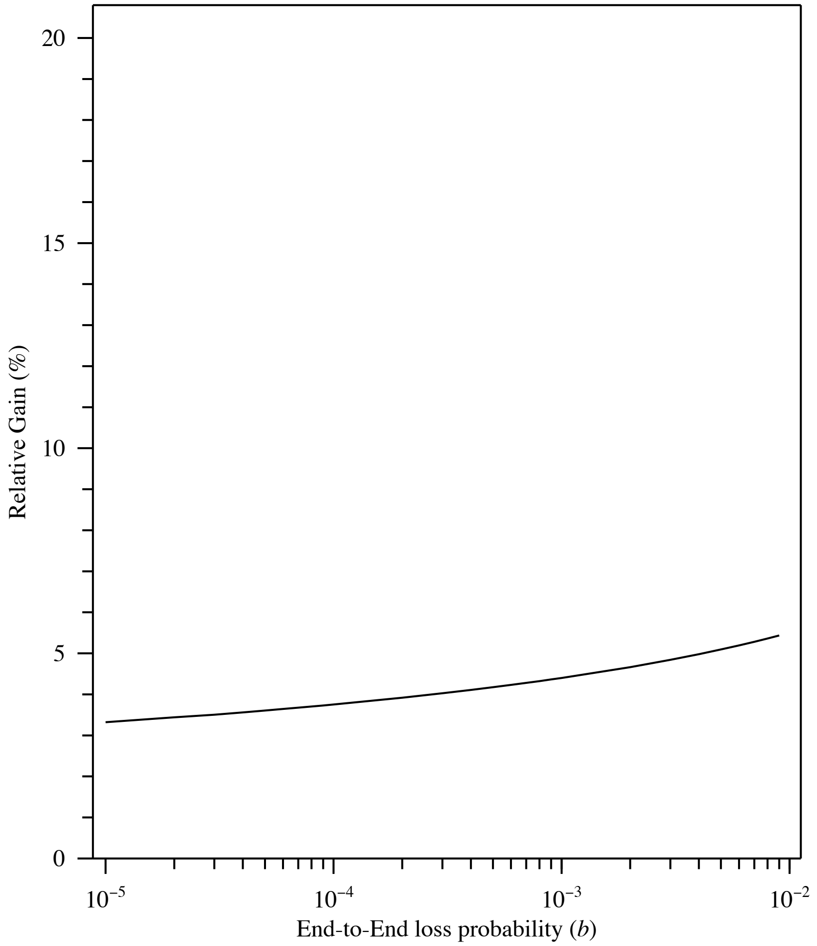

Figure 17: M2 Source Model: Relative performance of the OPT and EQ policies for the alternate four hop network model with cross traffic

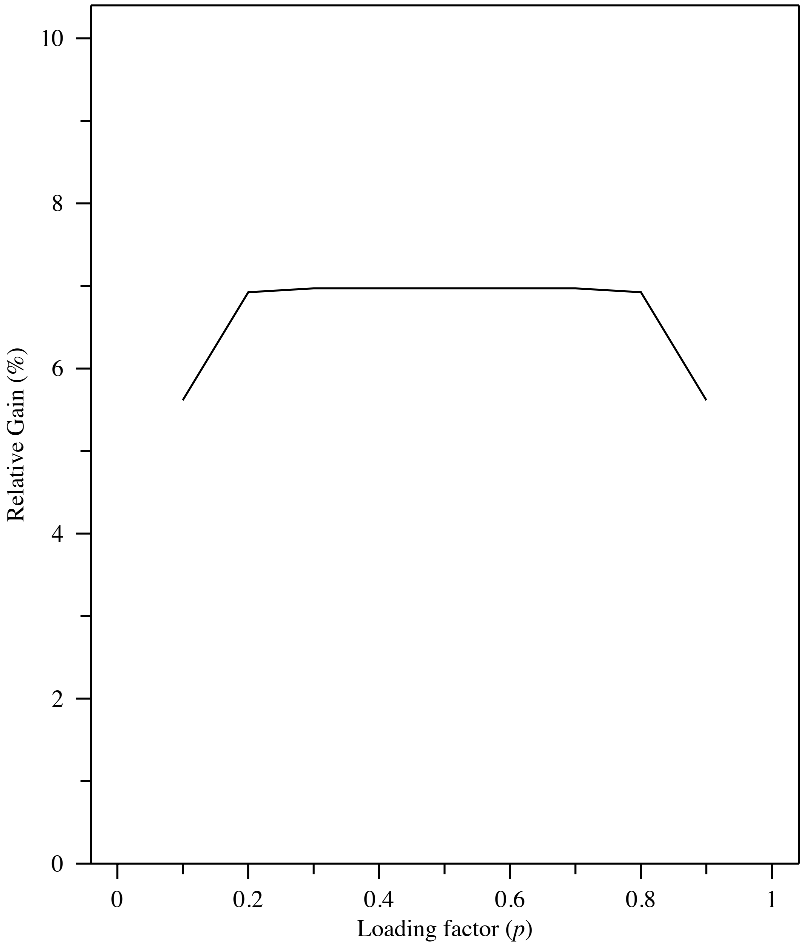

Figure 18: M2 Source Model: Relative performance of the OPT and EQ policies for the alternate four hop network model with cross traffic

briefly digress to consider one other QOS metric, the average packet delay. It will be seen that this QOS metric leads to a very contrasting performance of the EQ and OPT policies.

Appendix A outlines the computation of the RGR for the average delay QOS metric and the M1 source model. The appendix also analyzes the performance of the EQ and OPT allocation policies for the network models of Section 5 and 6. For the network model of Section 5, we find that the OPT policy outperforms the EQ policy significantly when the *average delay QOS values* are *small* and there is an imbalance in the nodal bandwidths. One may recall, in the case of the loss QOS metric, that 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 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.

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 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.
- The relative performance of QOS allocation policies is of interest only when there are resource or load imbalances in the network.
- 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.,.

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 (prelimnary work [Yat] 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. Then an optimal allocation of the QOS among the nodes with high RGR values and a simple strategy for the low RGR value nodes could be adopted.

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 apriori. 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 partion 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 vaue 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.

A second issue of interest is to consider alternate approximations for the performance of the voice sources multiplexer. One of particular interest to the authors is the fluid-flow approximation [AMS82]. Third, alternate source models such as the (σ, ρ) characterization of Cruz [Cru91] could also be considered. 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

[A$^+$67] Milton Abramowitz et al. *Handbook of Mathematical Functions with Formaulas, Graphs and Mathematical Tables.* National Bureau of Standards, 1967.

[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.

[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.

[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.

[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.

[HLP52] G. Hardy, J.E. Littlewood, and G. Polya. *Inequalities.* Cambridge University Press, 1952.

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

[Kur92] James F. 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 James F. 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, James F. 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.

[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.

[Pap84] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw Hill, 1984.

[PG92] Abhay Parekh and Robert Gallager. A generalized processor sharing approach to flow control in integrated services networks - the single node case. In *INFOCOM'92*, pages 915–924, 1992.

[PZ82] William R. Parzinsky and Philip W. Zipse. *Introduction to Mathematical Analysis.* McGraw-Hill Book Company, 1982.

[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.

[Yat] David Yates. Computer Science Dept., University of Massachusetts at Amherst, Work in progress, November 1992.

Appendix A: Optimal average delay QOS allocation

In this appendix, we briefly discuss the performance of allocation policies when the average delay is the QOS metric of interest.

Consider M1 sources and infinite queueing capacity at a node. This leads to a $M/M/1$ queueing model and the nodal performance is specified as:

$$q = G(N) = \frac{1}{\mu - N\lambda_s} \tag{48}$$

where μ is the nodal bandwidth. The RGR value is then easily computed as:

$$\Phi(\mu, q) = \frac{1}{\mu q - 1}. \tag{49}$$

Figure 19 shows the RGR as a function of the nodal bandwidth and delay QOS requirement. It can be seen that the RGR value is large when the nodal bandwidth and the delay QOS requirements are small. For example, consider a node with $\mu = 1000$ units of bandwidth and increase the delay QOS value from $q = 2 \times 10^{-03}$ to $q + \Delta q = 4 \times 10^{-03}$ units, i.e., a unit increase in the QOS allocation at that node. The RGR value for this node is now unity, i.e., the upper bound on the gain in the supportable load is unity, allowing for large gains in traffic load. Indeed as $q \to 1/\mu$ for a fixed value of the nodal bandwidth μ or as $\mu \to 1/q$ for a fixed value of the QOS requirement q, we have $RGR \to \infty$. Hence, it is in this regime of *small delays* or *small nodal capacities* that one can expect *significant improvements over a naive EQ Policy.*

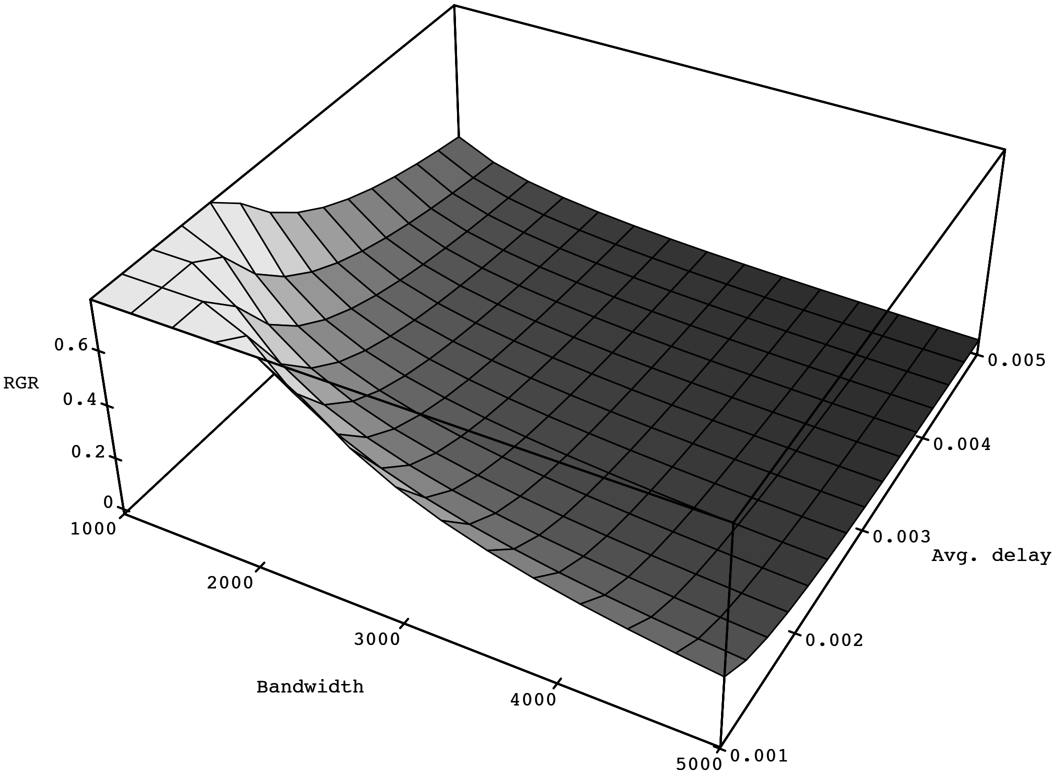

Figure 19: Relative Gain Ratio for the average delay QOS metric and the $M/M/1$ queue

Extensive numerical investigations in the context of the network models of section 5 and 6 have confirmed the general trends predicted by the RGR computation above. We present here one sample computation. Consider the tandem network model of Section 5 with five hops, i.e., $h = 5$. The link bandwidths at the nodes are taken to be $1000, 1100, 1200, 1300$ and 1400 units respectively. Figure 20 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 small gains are available. Results for the network model of Section 6 with cross traffic indicate that the delay QOS allocation policy performance is relatively sensitive to the loading factor on the two paths. This is again in stark contrast to the results for the loss QOS metric.

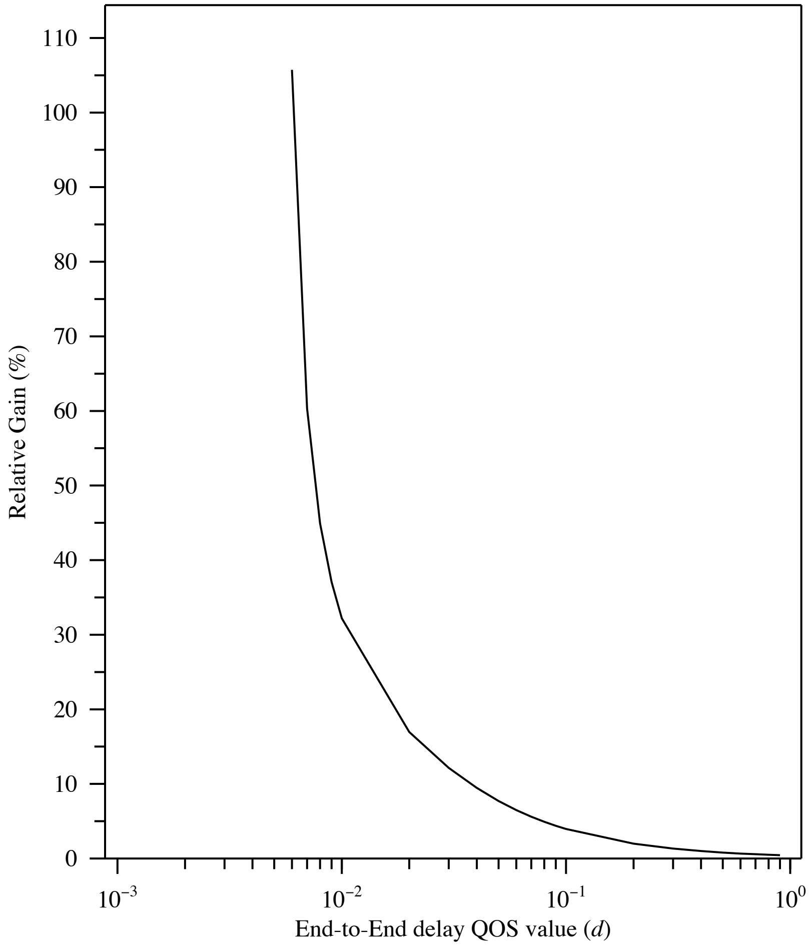

Figure 20: Relative performance of EQ and OPT delay QOS allocation policies as a function of delay requirement - Five hop model with M1 sources

Appendix B: Optimal loss QOS allocation policy

In this appendix, we present the solution for the maximum number of connections supportable in the tandem network model of Section 5 with the loss probability as the QOS metric (see Section 5.2). First, we ignore the effect of upstream losses on downstream nodes. Subsequently, we will account for upstream losses. We will consider alternatively the M1 and M2 source traffic models.

Consider the M1 traffic model and tandem $M/M/1/k$ queues. In Section 5.2, the maximum number of connections that could be admitted under the optimal loss QOS allocation policy was the largest value of ρ_1 that satisfied

$$f(\rho_1) = 1 - \prod_{i=1}^{h}(1 - G_i(\rho_i, k_i)) - b \leq 0 \tag{50}$$

where

$$G_i(\rho_i, k_i) = \frac{(1-\rho_i)\rho_i^{k_i}}{1-\rho_i^{k_i+1}}. \tag{51}$$

and

$$\begin{aligned} \rho_i &= \frac{\mu_1}{\mu_i}\rho_1 \\ \rho_1 &= \frac{N\lambda_s}{\mu_1}. \end{aligned} \tag{52}$$

We will show in the following that $f(\rho_1)$ is an increasing function of ρ_1 for $\rho_1 \in I = (0, \infty)$. The solution to $f(\rho_1) = 0$ is, hence, unique and is the optimal solution.

Taking derivatives,

$$\frac{df(\rho_1)}{d\rho_1} = \sum_{j=1}^{h} \frac{dG_j(\cdot)}{d\rho_1} \prod_{i=1,\, i\neq j}^{h} (1 - G_i(\cdot)) \tag{53}$$

Since $0 < G_i(\cdot) < 1.0$ and $dG_i(\cdot)/d\rho_1 > 0$ (note that $dG_i(\cdot)/d\rho_1$ is indeterminate for $\mu_1\rho_1/\mu_i = 1.0$ where we define it to be $(k/2(k+1))(\mu_1/\mu_i) > 0$, the limiting value at that point) on I, we have that $df(\rho_1)/d\rho_1 > 0$ and hence $f(\rho_1)$ is an increasing function on I [PZ82, Theorem 5.7].

Next we account for the upstream losses in determining the uniqueness of the optimal policy. From Section 5.2, we have the following relation among the nodal traffic intensities under the optimal policy:

$$\rho_i = \frac{\mu_1}{\mu_i}\rho_1 \prod_{j=1}^{i-1}(1 - G_j(\rho_j)) \quad i = 2, 3, \cdots, h. \tag{54}$$

We will show in the following that ρ_i, $i = 1, 2, \cdots, h$ is an increasing function of ρ_1. Since $G_i(\cdot)$ is an increasing function of ρ_i, we will have demonstrated that $G_i(\cdot)$ is an increasing function of ρ_1.

The following lemma will prove useful in the following proof.

Lemma 1 *Let $f(\rho) = \rho(1-q)$ where $q = (1-\rho)\rho^k/(1-\rho^{k+1})$ and k is an arbitrary positive integer. Then $f(\rho)$ is an increasing function of ρ for $\rho \in I$.*

Proof:

With minor algebraic manipulation, we have

$$\begin{aligned} f(\rho) &= \frac{\rho(1+\rho+\rho^2+\cdots+\rho^{k-1})}{1+\rho+\rho^2+\cdots+\rho^k} \\ &= \frac{g(\rho)}{1+g(\rho)}, \end{aligned} \tag{55}$$

where $g(\rho) = \rho + \rho^2 + \cdots + \rho^k$. Taking derivatives with respect to ρ, we have:

$$f^{'}(\rho) = \frac{1}{(1+g_\rho)^2} g^{'}(\rho) \tag{56}$$

Now $g^{'}(\rho) = 1 + 2\rho + \cdots + k\rho^{k-1} > 0$ for $\rho \in I$. Hence $f^{'}(\rho) > 0$ for $\rho \in I$ and by Theorem 5.7 [PZ82] our proof is complete. ■

We are now ready to show that ρ_i is an increasing function of ρ_1.

Theorem 1 *$\rho_i = (\mu_1/\mu_i)\rho_1 \prod_{j=1}^{i-1}(1 - G_j(\rho_j))$ $i = 1, 2, \cdots, h$ is an increasing function of ρ_1 for $\rho_1 \in I$.*

Proof:

The proof is by induction.

Base Case ($i = 1$):

Clearly, ρ_1 is itself an increasing function of ρ_1.

Induction Step:

Let ρ_{i-1} be an increasing function of ρ_1. We now wish to show that ρ_i is an increasing function of ρ_1. We can rewrite ρ_i as

$$\rho_i = \frac{\mu_{i-1}}{\mu_i}\rho_{i-1}(1 - G_{i-1}(\rho_{i-1})) \tag{57}$$

By Lemma 1, we have that ρ_i is an increasing function of ρ_{i-1}. But ρ_{i-1} is an increasing function of ρ_1 by the induction hypothesis. Hence ρ_i is an increasing function of ρ_1. This completes the proof. ■

Finally, we detail the computation of the upper bound for the relative gain of the OPT policy over the EQ policy for the loss QOS metric considered in Section 5.2 for the M1 source model and tandem $M/M/1/k$ queues. The bound for the relative gain was computed as

$$\frac{N - N_{eq}}{N_{eq}} = \Phi(k, 1 - (1-b)^{1/h})(h-1) \tag{58}$$

where k is the buffer size at the bottleneck node and $\Phi(\cdot)$ is the RGR value for the $M/M/1/k$ queue. In order to establish the above inequality, we only need to show that

$$\frac{b-1+(1-b)^{1/h}}{1-(1-b)^{1/h}} < h-1. \tag{59}$$

The above inequality is easily derived using [HLP52, Equation 2.15.4] which states that $(1-b)^{1/h} < 1-b/h$. Hence

$$\begin{aligned} \frac{b-1+(1-b)^{1/h}}{1-(1-b)^{1/h}} &< \frac{b-1+(1-b)^{1/h}}{b/h} \\ &< h(1-1/b)+\frac{(1-b)^{1/h}}{b/h} \\ &< h(1-1/b)+h/b-1 \\ &< h-1. \end{aligned} \tag{60}$$

Next we consider the M2 source model and the voice multiplexer. The optimal number of connections in Section 5.2 was taken to be the largest value of N that satisfied

$$F(N)=1-\prod_{i=1}^{h}(1-G_i(N))-b\leq 0 \tag{61}$$

where

$$G_i(N)=e^{-(ln(2\pi)+(\frac{C_i-Nm}{\sqrt{N}\sigma})^2)/2}. \tag{62}$$

We will now show that $F(N)$ is an increasing function of N and hence there exists a unique value of N that satisfies $F(N)=0$ and is the optimal solution.

Taking derivatives we have

$$F^{'}(N)=\sum_{j=1}^{h}\frac{dG_j(\cdot)}{dN}\prod_{i=1,\, i\neq j}^{h}(1-G_i(\cdot)) \tag{63}$$

Hence, it only needs to be shown that $dG_j(\cdot)/dN > 0$. It can be easily shown that

$$\frac{dG_j(\cdot)}{dN}=\frac{G_j(\cdot)(C_j^2-N^2m^2)}{2N^2\sigma^2} \tag{64}$$

Hence, we have $dG_j(\cdot)/dN > 0$ provided $C_j > Nm$, i.e., the utilization is less than unity.

Appendix C: Note on the Gaussian bit rate approximation

In this appendix, we first examine the appropriateness of modeling the aggregate bit rate of a superposition of on-off sources as a Gaussian random variable (RV). We then examine how this approximation can be used to determine the number of sources that can be multiplexed onto a link of given capacity while satisfying a specified loss probability for the sources. The problem can be seen to be equivalent to the problem of inverting the CDF of a Gaussian random variable. Next, we address the problem of determining the loss probability given a fixed number of sources being multiplexed onto a link of known capacity.

Consider the case of N identical and independent on-off sources with exponentially distributed on and off periods. Define

A**:** Random variable representing the number of active sources, i.e., sources in the active state, at any arbitrary point in time.

p**:** Probability of a source being in the active state at any arbitrary point in time.

Then

$$P(A=k)=\left(\begin{array}{c} N \\ k \end{array}\right) p^k(1-p)^{N-k}. \tag{65}$$

Now, by the DeMoivre-Laplace theorem [Pap84] we have

$$\lim_{N\to\infty} P(A=k)=\frac{1}{\sqrt{2\pi Np(1-p)}}e^{-(k-Np)^2/2Np(1-p)} \tag{66}$$

Note that the right-hand side of the above equality is the pdf of a Gaussian RV with appropriate mean and variance. The above suggests that the aggregate bit rate may indeed be approximated by a Gaussian RV provided the number of sources is reasonably large. We test this hypothesis for the case of 15 M2 sources. Figure 21 shows the exact and normal pmf for the number of active sources. The error value in the figure is simply the sum of the squares of the differences between the exact and Gaussian pmfs. The approximation thus appears to be reasonable.

Next we consider the inversion of the cdf of a Gaussian random variable. Let $B \sim N(m,\sigma)$. Let α be some unkown variable to be determined such that $P(B > m+\alpha\sigma) \leq q$ and we desire the smallest such value of α. Since B is a Gaussian RV, $Y=(B-m)/\sigma \sim N(0,1)$ and we have [A$^+$67, eqn. 26.2.25]

$$P(Y\leq\alpha)\geq 1-\frac{1}{\alpha}(2\pi)^{-1/2}e^{-\alpha^2/2}. \tag{67}$$

Note that Abramowitz [A$^+$67] recommends that the bound be used only for $\alpha > 2.2$. However, it can be clearly seen [A$^+$67, Fig. 26.1] that the bound holds for all values for α with somewhat lesser tightness. Hence

$$P(Y>\alpha)\leq\frac{1}{\alpha}(2\pi)^{-1/2}e^{-\alpha^2/2} \tag{68}$$

Now, we desire to find the value of α such that $P(Y>\alpha)\leq q$. To do this we set the right-hand side of the above inequality to q, i.e.,

$$\frac{1}{2\pi\alpha^2}e^{-\alpha^2}=q^2 \tag{69}$$

Now, assume that $\alpha\geq 1$ which implies that $1/\alpha^2\leq 1$. Hence if

$$e^{-\alpha^2}=2\pi q^2, \tag{70}$$

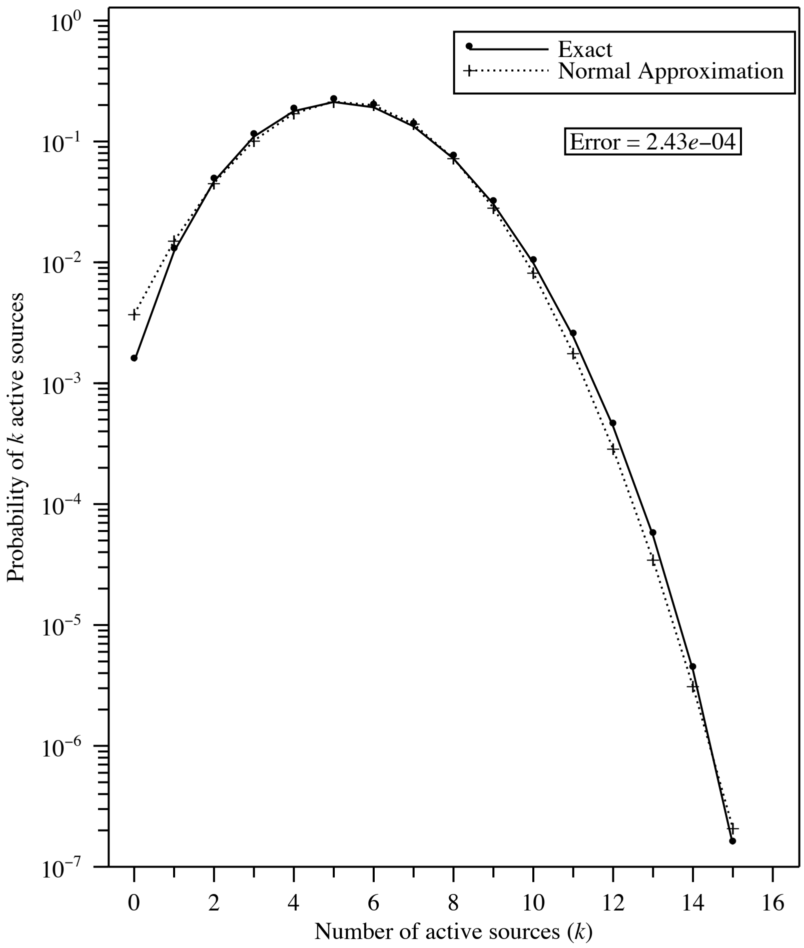

Figure 21: Gaussian approximation for the number of active M2 sources

we will have $P(Y > \alpha) \leq q$. Solving the above equation for α yields

$$\alpha = \sqrt{-2ln(q) - ln(2\pi)} \tag{71}$$

However, by the earlier assumption we require $\alpha \geq 1$. This requires that $q \leq 0.2419$. Hence for $q \leq 0.2419$, choosing $\alpha = \sqrt{-2ln(q) - ln(2\pi)}$ results in $P(B > m + \alpha\sigma) \leq q$ as desired. As a point of interest, note that with $q \leq 0.035$, $\alpha \geq 2.2$ and the bound on the tail will be tight.

We now derive an expression for the loss probability given a fixed number of sources and a link of known capacity. As previously, let $B \sim N(m, \sigma)$ be a Gaussian RV representing the aggregate bit rate of the superposition. The loss probability is then taken to be bounded by [GG92]

$$q \leq P(B > C) \tag{72}$$

where C is the link capacity. Since $P(B > C) = P(Y > (C - Nm)/(\sqrt{N}\sigma))$, where $Y \sim N(0,1)$, we have

$$q \leq \frac{1}{\sqrt{2\pi}\alpha} e^{-\alpha^2/2} \leq \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2} \tag{73}$$

provided $\alpha = (C - Nm)/(\sqrt{N}\sigma) \geq 1$.

Appendix D: Convexity of the loss probability in the voice multiplexer

In this appendix, we seek to determine if the loss probability in the voice multiplexer is a convex funtion of the number of sources being superposed at the input.

From Section 4, we have the loss in the voice multiplexer as

$$q = G(N, C) = e^{-(ln(2\pi) + (\frac{C-Nm}{\sqrt{N}\sigma})^2)/2}. \tag{74}$$

Taking derivatives with respect to N, we have

$$G''(N) = \frac{G(N)C}{N^3\sigma^2}(\frac{(C^2 - N^2m^2)^2}{4N\sigma^2C^2} - 1) \tag{75}$$

Let

$$\frac{C - Nm}{2\sqrt{N}\sigma} \geq 1. \tag{76}$$

It can be easily shown that this implies that $G''(N) > 0.0$. After algebraic manipulations it can be shown that the above inequality implies that

$$N \leq L(C, m, \sigma) = (\frac{-\sigma}{m} + \sqrt{(\frac{\sigma}{m})^2 + \frac{C}{m}})^2 \tag{77}$$

As a point of interest, we compute the value of $L(\cdot)$ for some typical parameter values. Let $C = 1$ Mbps and m and σ take the values for the voice source in Section 3. We then have $L(\cdot) = 66.75$. This value of $L(\cdot)$ corresponds to a utilization of approximately 75% on that link.

Appendix E: An Optimization Problem

In this appendix, we solve the following optimization problem

$$\begin{aligned} &\text{Maximize} \quad N \\ &\text{Subject to} \quad F_i(N) \leq 0 \;\; i = 1, 2, \cdots, m. \end{aligned} \tag{78}$$

We assume that the $F_i(N)$, $\forall i$ are increasing functions of N. We first conjecture on the optimal solution for N, say N^*, and establish that it is indeed the optimal solution.

First, define N_i as:

$$N_i = \{N : F_i(N) = 0\} \;\; i = 1, 2, \cdots, m. \tag{79}$$

Without loss of generality we assume that $N_1 \leq N_j \;\; \forall j$. We claim that $N^* = N_1$. Since $F_i(\cdot)$ is an increasing function of N, it must be clear that $F_i(N^*) \leq 0$. Hence, N^* is certainly a candidate for the optimal solution. In fact, we will show trivially that it is indeed the optimal solution. Let $K > N^*$ be the optimal solution. Then since $F_1(\cdot)$ is an increasing function of N we have $F_1(K) > 0$ which violates the QOS constraint and hence cannot be the optimal solution. Hence, N^* is indeed the optimal solution.