INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2004; **17**:193–215 (DOI: 10.1002/dac.652)

Modelling multi-dimensional QoS: some fundamental constraints

Nelson X. Liu*,† and John S. Baras‡

Institute for Systems Research, University of Maryland, College Park, MD 20742, U.S.A.

SUMMARY

In this paper, we model multi-dimensional QoS in a unified framework, and study some fundamental constraints from the network and the traffic on realizing multiple QoS goals. Multi-dimensional QoS requirements are quantitatively represented using a QoS region. Based on the theory of effective bandwidths, the framework connects the throughput, the delay, and the loss rate in a uniform formula. Important traffic and network factors, namely, the burst size and the link speed, are involved. With this framework, it is found that the burst size sets hard limit on the QoS region that can be achieved, and that the matching between the link speed and the node processing power can greatly improve the limit. It is also made clear that while pure load imbalance among links does not affect the QoS region, the heterogeneities of burst size or link speed may severely degrade the QoS performance. Applying the theory to real-time services in differentiated services architecture, we show it provides a useful tool for QoS prediction and network dimensioning. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: quality of service; multi-dimensional QoS; network dimensioning; effective bandwidth; QoS region; burst size; link speed; network heterogeneity

1. INTRODUCTION

The Internet is accommodating more and more services to support different applications and fulfil different needs of users. An important issue with the multiple service networks is to provide multi-dimensional quality of service (QoS) for different services in the same infrastructure. Because of the conflicts between QoS goals and the complexity of the service setting, this is a challenging problem.

There has been continuing effort on QoS research for Internet. Huge literature exists with regard to QoS mechanisms, such as the packet scheduling in the router [1], the admission control at network edge [2], and the rate adaptation in the end system [3]. Although great achievement has been made, much of the work either focuses on a single QoS dimension [4–9] like the delay, the throughput, or the loss rate, or is dedicated to design of architectures or algorithms

*Correspondence to: Nelson Liu, Institute for Systems Research, University of Maryland, College Park, MD 20742, U.S.A.
†E-mail: nliu@isr.umd.edu
‡E-mail: baras@isr.umd.edu

Contract/grant sponsor: DARPA; contract/grant number: N66001-00-C-8063

[2, 3, 10, 11]. Not much effort has been put on considering multiple dimensions as a whole, examining inherent relations between them, discussing the nature of their conflicts, and evaluating the conditions of the network and the traffic to realize multiple QoS goals altogether. With increasing demands on multi-dimensional QoS, it is necessary to address these issues in a unified framework.

Paper [12] demonstrates some examples that multi-dimensional QoS requirements may not be satisfied when we expect they are. An extreme case for real-time services in the differentiated services (DS) networks is given in Reference [13], in which simultaneous arrivals of bursts severely delay some packets in a node and the delay accumulates exponentially with the number of hops. This case may be tolerated in practice with statistical QoS. However, we need find out what we can gain by sacrificing the loss rate, and in what conditions the real-time deadline can be fulfilled with an acceptable loss rate and a reasonable throughput. To do this, a model formally connecting multiple QoS dimensions seems to be indispensable.

In this paper, we make an attempt to formulate such a framework. We use a QoS region to quantitatively represent multi-dimensional QoS goals. Relations between different QoS dimensions are established through the theory of effective bandwidths. With this framework we explore some fundamental conditions imposed by the traffic and the network on realizing multi-dimensional QoS. These are basic constraints in the sense that they set the limit for the performance that any packet scheduling algorithm or traffic shaping algorithm can achieve, and that they should count in even preliminary network dimensioning to supply QoS. How the traffic and the network factors affect multiple QoS dimensions at the same time are displayed as well.

The rest of the paper is organized as follows. In Section 2 the QoS region is defined. In Section 3 the establishment of relations between different QoS dimensions is presented. Section 4 to Section 7 are dedicated to conditions and effects of the traffic and the network in supporting multi-dimensional QoS. Among them Section 4 is for the burst size, Section 5 is for the link speed, and Section 6 is for traffic and link heterogeneities. Section 7 gives an example applying the theory in DS networks. Section 8 concludes the paper.

2. THE MULTI-DIMENSIONAL QOS REGION

We use multi-dimensional QoS region to represent the multiple QoS requirements quantitatively. Assume there are N QoS indices $R_1, R_2, \ldots, R_N$, which may represent the throughput, the loss rate, and the delay, etc. An index R_i has a value scope of $I_i = [\mathrm{MIN}_i, \mathrm{MAX}_i]$. All possible values of these indices form an N-dimensional space $S(R_1, R_2, \ldots, R_N)$. A particular QoS requirement r_i is specified in the form $r_i = [\min_i, \max_i]$, where $[\min_i, \max_i] \subseteq I_i$ is an valid sub-section of I_i. For example, we can require that the throughput should not be less than 80% by specifying $r_1 = [0.8, 1.0]$. A QoS region $Q(r_1, r_2, \ldots, r_N)$ is a N-dimensional sub-area in $S(R_1, R_2, \ldots, R_N)$ where all QoS requirements $r_1, r_2, \ldots, r_N$ are satisfied at the same time. Since multiple QoS requirements compete limited network resources to get satisfied, R_i's satisfaction means less resource available for R_j $(j \neq i)$. So it holds that $Q(r_1, r_2, \ldots, r_N) \subset S(R_1, R_2, \ldots, R_N)$ for practical networks.

In this paper we focus on three most important QoS dimensions, the throughput ρ, the loss rate φ and the delay π. In particular, we are most interested in the real-time services, which have strict requirements on all three dimensions. The principles in this paper generally apply to the

best-effort and the elastic services as well. For the real-time services, the most critical QoS requirement is the delay. If the end-to-end delay of a packet goes beyond a deadline, the packet becomes useless. On the other hand, it is not necessary, though preferred, to take much effort to reduce the delay further when the deadline is satisfied. Because of this, we can first establish a basic condition to fulfil the delay requirement, and then examine factors that affect the throughput and the loss rate—it is always meaningful to improve the later two. This is the philosophy behind the approach we use to establish the conditions for multi-dimensional QoS in following sections.

One QoS region of common interest is $Q_c = Q([\rho_c, 1.0], [0, \varphi_c], [0, \pi_c])$, meaning the throughput is not less than ρ_c, the loss rate is not greater than φ_c, and the delay is not beyond π_c. It is simply denoted as $Q_c(\rho_c, \varphi_c, \pi_c)$ provided there is no confusion. If only one point within a QoS region can be realized, the region is said to be reachable; otherwise, unreachable. When the delay requirement π_c is configured as default, as for the real-time services, the notation of $Q_c(\rho_c, \varphi_c, \pi_c)$ is further simplified as $Q_c(\rho_c, \varphi_c)$. The QoS region $Q_c(\rho_M, \varphi_c)$ where $\rho_M = \max\{\rho \mid \varphi \leqslant \varphi_c\}$ is called the premium QoS region at loss rate φ_c, meaning it has the biggest throughput when the loss rate is not greater than φ_c. For comparison purposes, it is convenient to define QoS regions bounded with the loss rate $\varphi_c = e^{-k}$, $k = 0, 1, 2, \ldots$. We call $Q_c(0, e^{-k})$ the kth QoS region, and denote it as Q_k. The throughput $\rho_k = \max\{\rho \mid \varphi \leqslant e^{-k}\}$ can be used to measure the scope of the region and is called the size of Q_k.

3. RELATING MULTIPLE QOS DIMENSIONS

As we have suggested in Section 2, the three dimensions of QoS goals conflict with one another. We need to find out relations between them and identify key factors to fulfil them altogether. This is through the theory of effective bandwidths. Before starting to establish the relations, we first introduce the traffic model assumed in the analysis.

3.1. The bursty traffic model

The traffic model we have in mind is a very general one. We view the traffic as a series of bursts, separated by idle periods. For simplicity, assume the burst arrivals are of a Poisson process with average rate v, and the burst size is exponentially distributed with mean b. However, the principle in this paper can be easily extended to any bursty traffic model with explicit effective bandwidth (see Section 3.2). When passing through a link, the burst series appears to be a Markovian on–off process [14]. The sum of means of the on and the off periods is $T = 1/v$. Assume the link speed is h. Then the means of the on and the off periods are the following:

$$\frac{1}{\mu} = \frac{b}{h} \tag{1}$$

$$\frac{1}{\lambda} = T - \frac{1}{\mu} \tag{2}$$

A feature of this model is that it involves the burst size and the link speed. This turns out to be important in finding out the traffic and network constraints on multi-dimensional QoS, as we will show later.

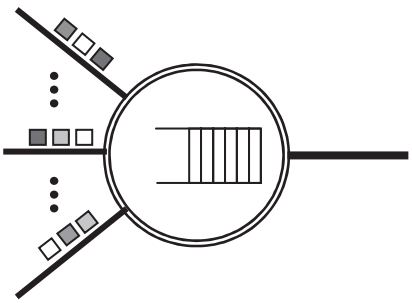

Figure 1. A network node.

We consider the network node in Figure 1, in which traffic is fed from n links to a single-server infinite FIFO queue. Denoting the load on link i is r_i, we represent the throughput of the node, ρ, as the system utility

$$\rho = \frac{\sum_{i=1}^{n} r_i}{C} \tag{3}$$

where C is the queue service rate. It is also referred to as the network utility sometime in this paper. If the traffic on all input links is homogeneous, i.e. it has the same b and T for every link, then the throughput is

$$\rho = \frac{nr}{C} = \frac{nb}{CT} \tag{4}$$

Given throughput ρ and burst size b, T can be represented as

$$T = \frac{nb}{C\rho} \tag{5}$$

So (2) becomes

$$\frac{1}{\lambda} = T - \frac{1}{\mu} = \frac{nh - C\rho}{C\rho h} b \tag{6}$$

In (6) the traffic parameter λ is expressed with link speed h, throughput ρ and burst size b. This indicates the traffic and the network are correlated in one aspect.

3.2. *Effective bandwidth*

In this part we give some background information about the theory of effective bandwidths [15–19, 24]. Let $W(t)$ be the amount of traffic during the period $[0, t]$. The effective bandwidth is defined as the following:

$$A(s,t) = \frac{1}{st} \log E[e^{sW(t)}] \quad 0 < s, \;\; t < \infty \tag{7}$$

It turns out that for any fixed t, $A(s,t)$ is between the average traffic rate and the peak rate. Intuitively, if the traffic is furnished with the bandwidth equivalent to the peak rate, there would be a bandwidth waste. However, if the bandwidth just equals the average rate, the performance may be very bad because of the burstiness of the traffic. The effective bandwidth provides a good reference point for the bandwidth needed to realize certain performance level. The effective

bandwidth has some good properties. One of them is the additivity. If n flows of traffic are independent, and the effective bandwidth of the ith low is $\alpha_i(s,t)$, then the effective bandwidth of the aggregate traffic is

$$A(s,t) = \sum_{i=1}^{n} \alpha_i(s,t) \tag{8}$$

Extensive discussions and examples about the effective bandwidth can be found in Reference [18].

3.3. Establishing relations between QoS dimensions

In this part we establish relations between three most important QoS dimensions, the throughput, the delay and the loss rate. Assume a network node dedicates bandwidth C to real-time services, and the nodal deadline for packets of the services is $\pi_c = D$. We get a critical queue length

$$q_D = C \cdot D \tag{9}$$

All packets beyond this point in the queue violate their deadlines and can be dropped. In the following part of the paper, the violation rate and the loss rate are thought of as being equivalent. According to the theory of effective bandwidths, the probability that the queue length is greater than q_D for Markovian traffic can be estimated as

$$\varphi = P[q > q_D] \approx e^{-q_D\delta} \tag{10}$$

where δ is a constant satisfying certain condition which we will explain shortly. Equation (10) actually establishes the relation between deadline D and loss rate φ. Increase of D can make φ decrease exponentially.

We need to bring the throughput ρ into the scenario. It turns out that ρ is implied in δ, and (10) is a unified formula involving three QoS dimensions. We will show this in the rest part of this section. We are interested in the steady packet loss behaviour when $t \to \infty$. According to the effective bandwidth theory, δ satisfies the following condition when $t \to \infty$:

$$\delta = \max\{s : A(s) \leqslant C\} \tag{11}$$

where $A(s)$ is

$$A(s) = \lim_{t\to\infty} A(s,t) \tag{12}$$

We call the inequality within the curly bracket on the right-hand side of (11) the effective bandwidth condition. It is a fundamental constraint that must be enforced to get statistical QoS.

It has been shown that for a Markovian traffic source with two states ON and OFF, the effective bandwidth is [18]

$$\alpha(s) = \frac{1}{2s}\left(hs - \mu - \lambda + \sqrt{(hs - \mu + \lambda)^2 + 4\lambda\mu}\right) \tag{13}$$

where μ and λ are the transition rates from ON state to OFF state and from OFF state to ON state, respectively, and h is the traffic rate in the ON state. From the probability theory, this is equal to say that the means of the on and the off periods are $1/\mu$ and $1/\lambda$ if they are exponentially distributed. So (13) applies to the fluid bursty traffic model in Section 3.1 as well. h is just the speed of the link where the traffic is passing. With the additive property of the effective

bandwidth, the aggregate of homogeneous traffic input to the node has the following effective bandwidth:

$$A(s) = n\alpha(s) \tag{14}$$

where $\alpha(s)$ is the effective bandwidth of the traffic on one link.

Now we will find out how δ is related with ρ. From (11), δ is the maximum value of s satisfying the effective bandwidth condition. From formula (11), (13) and (14), we get

$$\frac{n}{2s}\left(hs - \mu - \lambda + \sqrt{(hs - \mu + \lambda)^2 + 4\lambda\mu}\right) \leqslant C \tag{15}$$

It is

$$\sqrt{(hs - \mu + \lambda)^2 + 4\lambda\mu} \leqslant \frac{2C}{n}s - hs + \mu + \lambda \tag{16}$$

Squaring both sides in the above equation, we get

$$(hs - \mu + \lambda)^2 + 4\lambda\mu \leqslant \left(\frac{2C}{n}s - hs + \mu + \lambda\right)^2 \tag{17}$$

Developing the square on the right-hand side, and moving the terms to the left-hand side, we get

$$\left[\frac{C}{n}h - \left(\frac{C}{n}\right)^2\right]s^2 - \left[\frac{C}{n}(\mu + \lambda) - \lambda h\right]s \leqslant 0 \tag{18}$$

We know $s > 0$. So

$$\left[\frac{C}{n}h - \left(\frac{C}{n}\right)^2\right]s \leqslant \frac{C}{n}(\mu + \lambda) - \lambda h \tag{19}$$

Assume $h \geqslant C/n$ (Then $h < C/n$ is trivial, in which no queue exists at all), the left-hand side of the equation is non-negative. So

$$s \leqslant \frac{\mu + \lambda}{h - (C/n)} - \frac{\lambda h}{(C/n)(h - (C/n))} \tag{20}$$

Rewriting it as

$$s \leqslant \frac{\mu + \lambda}{h - (C/n)}\left(1 - \frac{1}{(C/n)}\frac{(h/\mu)}{(1/\lambda) + (1/\mu)}\right) \tag{21}$$

From formula (1), (2), and (5) we know

$$\frac{1}{\lambda} + \frac{1}{\mu} = T = \frac{nb}{C\rho} \tag{22}$$

From formula (1) we also know

$$\frac{h}{\mu} = b \tag{23}$$

Substituting (22) and (23) into (21), we get

$$s \leqslant \frac{\mu + \lambda}{h - (C/n)}(1 - \rho) \tag{24}$$

With formula (1) and (6) we have

$$\mu + \lambda = \frac{nh^2}{(nh - C\rho)b} \tag{25}$$

Using it into (24), we finally get

$$s \leqslant \frac{h^2}{\big(h - (C/n)\big)\big(h - (C/n)\rho\big)} \frac{1-\rho}{b}$$

Therefore,

$$\delta = \frac{h^2}{\big(h - (C/n)\big)\big(h - (C/n)\rho\big)} \frac{1-\rho}{b} \tag{26}$$

Thus we get the relation between δ and ρ.

Equations (10) and (26) are key formulas in this paper. They establish the relation between three QoS dimensions, the loss rate φ, the throughput ρ, and the delay deadline D. This quantitative formulation clarifies the intuitive relations among conflicting QoS dimensions. It shows that while the increase of the deadline allows the loss rate to decrease exponentially, the relation between the loss rate and the throughput is more complex. As we see, the equation includes network parameter h and traffic parameter b in addition to the QoS indices. This indicates that the relations between QoS dimensions depends on network and traffic factors, and enables us to find out critical conditions for realizing multi-dimensional QoS goals, as we will do in next several sections.

4. EFFECTS OF THE BURST SIZE

It has been well known that the burst size of the traffic affects the network performance significantly. In this section we will analyse it from the prospective of multi-dimensional QoS and see what limitations it impose on the QoS behaviour. Assume link speed $h \to \infty$, thus every burst arrives immediately once it is generated. This is the worst case that produces the higher bound of the packet loss rate.

4.1. Burst size constraints

Theorem 4.1
In a high-speed network ($h \to \infty$), for given burst size b, the loss rate for the real-time services with deadline D cannot be lower than the following limit:

$$\varphi \approx \mathrm{e}^{-\frac{q_D}{b}} \tag{27}$$

Proof
When $h \to \infty$, (26) becomes

$$\delta = \frac{1-\rho}{b} \tag{28}$$

From (10) the packet loss rate is

$$\varphi \approx e^{-\frac{q_D(1-\rho)}{b}} \tag{29}$$

Obviously,

$$\varphi > e^{-\frac{q_D}{b}}, \quad \forall \rho > 0$$ □

Theorem 4.1 indicates b sets a lower bound for the loss rate; if b is big enough, almost all packets will get lost ($\varphi \to 1$ when $b \to \infty$). However, when $b \to 0$, $\varphi \to 0$. This means if b is small enough, the packet loss rate can be arbitrarily low for any throughput. So b has critical affect on the QoS behaviour of a node.

Let us represent b in terms of the size of q_D and let

$$w = \frac{q_D}{b} \tag{30}$$

Figure 2 shows the relation between φ and ρ for different w. The curves mark the lower limits of φ in different ρ. The intersection points of the curves with the y-axis indicate the lower bounds of loss rates that can never been overcome. For example, if $b \geqslant q_D/2$ (or $w \leqslant 2$), then there always be $\varphi \geqslant e^{-2} \approx 13.5\%$. It is disappointing that φ increases exponentially with the increase of ρ. This means the cost will be high if we want to improve throughput by allowing some packet loss rate.

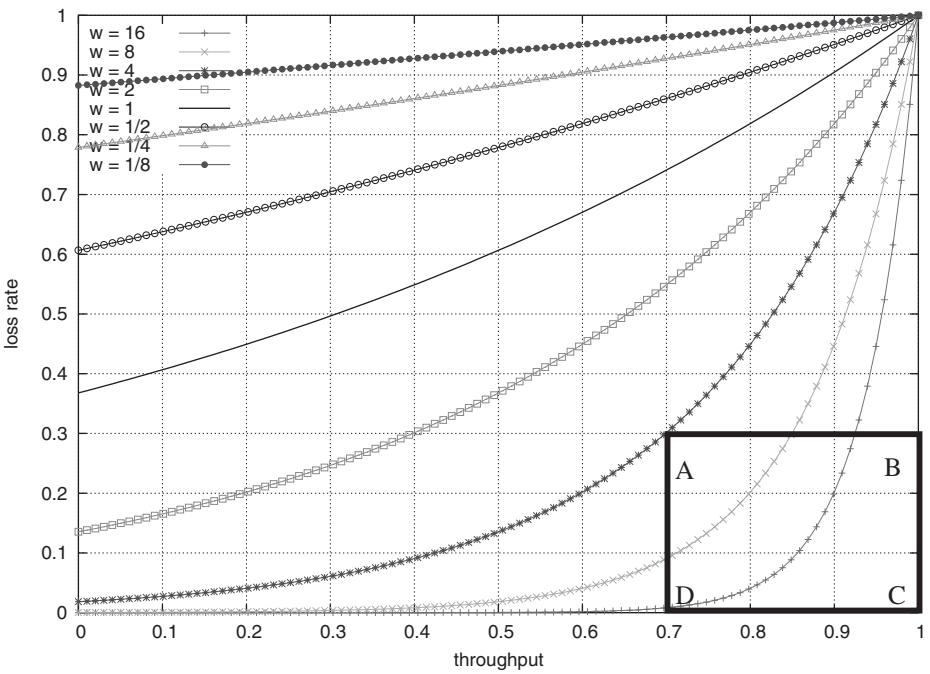

Figure 2. Relations between loss rate and throughput for different burst sizes.

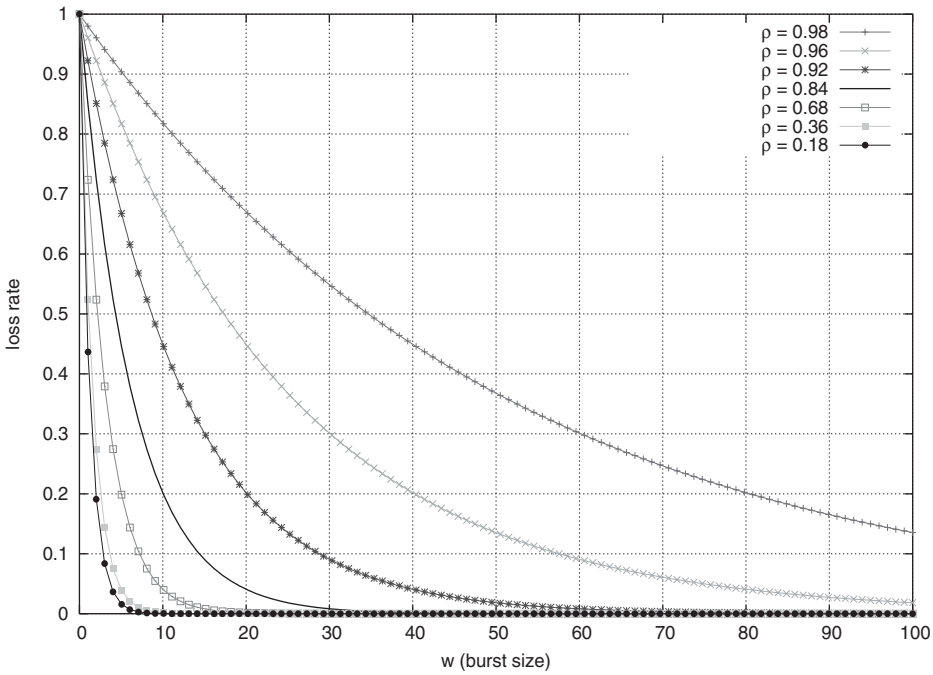

Figure 3. Relations between loss rate and burst size for different throughputs.

The change of φ with b is directly shown in Figure 3. We can see that the smaller the b is (or the bigger the w is), the lower the φ is. In another word, decreasing b can make φ decrease to an acceptable level. In general φ decreases quickly with the decrease of b (or the increase of w in the figure). Halving the burst size can significantly improve the QoS performance.

Theorem 4.2
In a high-speed network ($h \to \infty$), for the real-time services with deadline D, a QoS region $Q_c(\rho_c, \varphi_c)$ is reachable only if the burst size satisfies the following condition:

$$b \leqslant -\frac{1-\rho_c}{\ln \varphi_c} q_D \tag{31}$$

Proof
This is a direct result from (29). □

Though simple, this burst size condition is physically important. For example, in Figure 2 the area ABCD indicates a QoS region $Q_c(0.7, 0.3)$. This goal cannot be realized if only $b > q_D/4$, namely, $w < 4$. The condition sheds light on the design of the traffic shaper.

4.2. Scaling QoS regions

The curve l in Figure 4 shows the relation between φ and ρ for certain burst size. The kth QoS region Q_k is an area in the up-left part above l. For example, the whole area ABlCA is Q_0, the

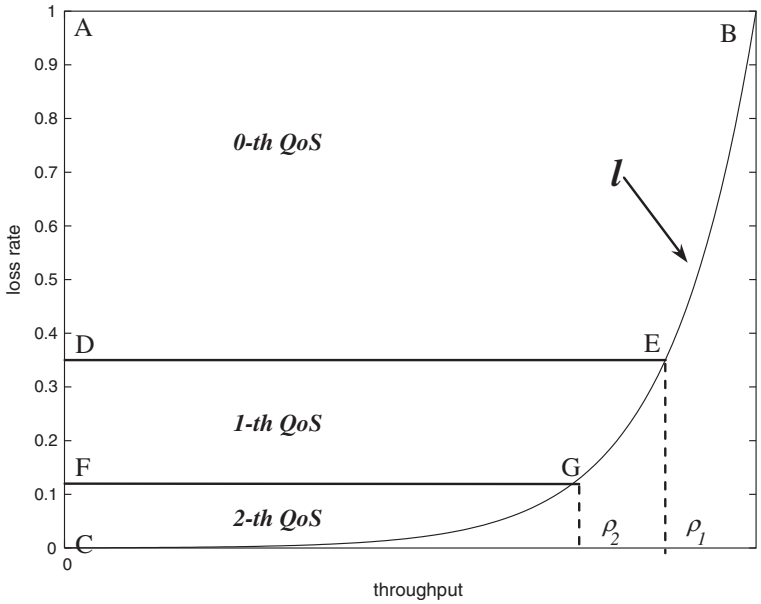

Figure 4. QoS regions.

area DElCD is Q_1, and the area FGlCF is Q_2. Q_1 is also called the basic QoS region. As we mentioned in Section 2, the size of Q_k can be measured with maximum throughput of the region, ρ_k.

For the kth QoS region, from formula (29) we get

$$-\frac{q_D(1-\rho_k)}{b} = -k \tag{32}$$

So the size of Q_k is

$$\rho_k = 1 - \frac{kb}{q_D} = 1 - \frac{k}{w} \tag{33}$$

As an example, for a given w, the basic QoS region cannot be larger than

$$\rho_1 = 1 - \frac{1}{w} \tag{34}$$

Formula (33) shows clearly that QoS regions shrink linearly with the increase of b, as illustrated in Figure 5. Assume the kth QoS region changes from ρ_k to ρ_k' when the burst size changes from b to b', from (32) we can get

$$\frac{1-\rho_k'}{1-\rho_k} = \frac{b'}{b} \tag{35}$$

Formula (33) also suggests another form of the burst size condition. Given an arbitrary ρ_k as a pre-defined kth region, we can get a higher bound of b to make it reachable:

$$b_k = q_D(1-\rho_k)/k \tag{36}$$

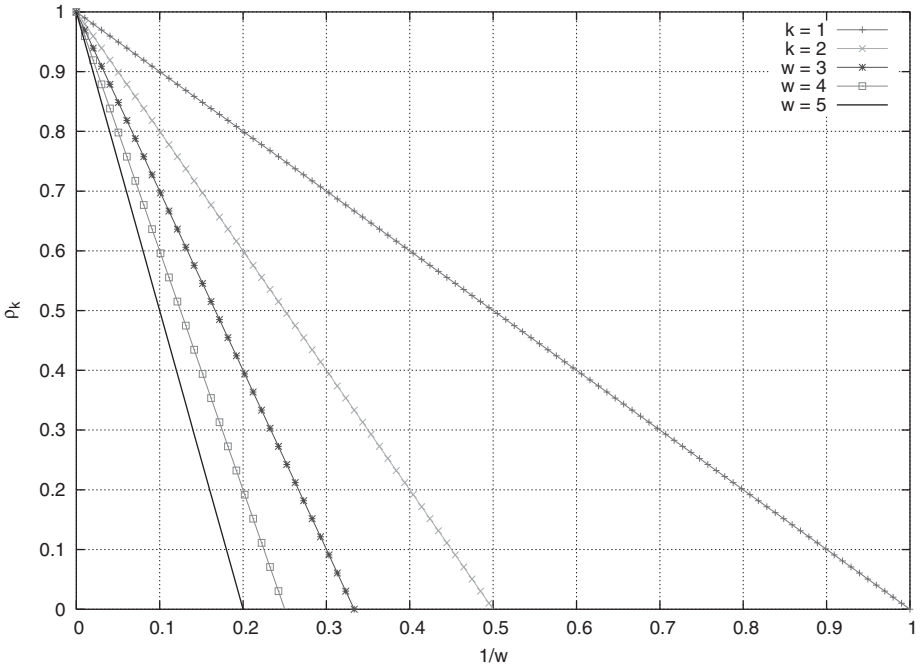

Figure 5. Change of kth QoS region with burst size.

4.3. *Marginal cost of statistical QoS*

We define the marginal cost of statistical QoS as the maximum increase of φ caused by unit increase of ρ in the kth QoS region. From formula (29) we can get

$$\frac{\mathrm{d}\varphi}{\mathrm{d}\rho} \approx \frac{q_D}{b} \mathrm{e}^{-\frac{q_D(1-\rho)}{b}} = \frac{q_D}{b}\varphi \tag{37}$$

When $\rho = \rho_k$ we have

$$\left.\frac{\mathrm{d}\varphi}{\mathrm{d}\rho}\right|_{\rho=\rho_k} \approx \frac{q_D}{b}\varphi|_{\rho=\rho_k} = \frac{w}{\mathrm{e}^k} \tag{38}$$

We see the cost becomes higher with the decrease of b, though the decrease enlarges the kth QoS region.

For a predefined size of Q_k, ρ_k, with formula (33) we have

$$\left.\frac{\mathrm{d}\varphi}{\mathrm{d}\rho}\right|_{\rho=\rho_k} \approx \frac{k}{(1-\rho_k)\mathrm{e}^k} \tag{39}$$

As an example, for Q_2 with $\varphi = \mathrm{e}^{-2} \approx 0.135$, we have

$$\left.\frac{\mathrm{d}\varphi}{\mathrm{d}\rho}\right|_{\rho=\rho_2} > 1 \quad \text{if } \rho_2 > 0.729 \tag{40}$$

This means the cost of statistical QoS is generally very high for a reasonable QoS region.

Above analyses in this section are based on the assumption of $h \to \infty$. The change of h, however, also has impact on the QoS region. When $h < \infty$ a better performance can be expected, as we will analyse next.

5. EFFECT OF THE LINK SPEED

Figure 6 illustrates the relation between φ and ρ for different h given other parameters. (The parameters are set as $D = 4.8$ ms, $C = 10$ Mbps, $b = 600$ bytes, and $n = 10$.) We see that φ is much higher for large h than for small h. When h is comparable to C/n, φ is low even for a high throughput. This indicates that just lowering the input link speed to C/n may improve the QoS performance significantly.

Theorem 5.1
Given deadline D and burst size b, when $h \to \infty$, the premium QoS region at loss rate $e^{-k}, k = 0, 1, 2, \ldots$, is $Q_c(1 - kb/q_D, e^{-k})$; when $h = C/n$, it is $Q_c(1, e^{-k})$, i.e. any valid QoS region is reachable.

Proof
As we mentioned in Section 4, when $h \to \infty$, (29) holds. Let $\varphi = e^{-k}$ in it, we get $\rho = 1 - kb/q_D$, which is also given in (33). It is easy to see that this value is the biggest one that ρ can expect

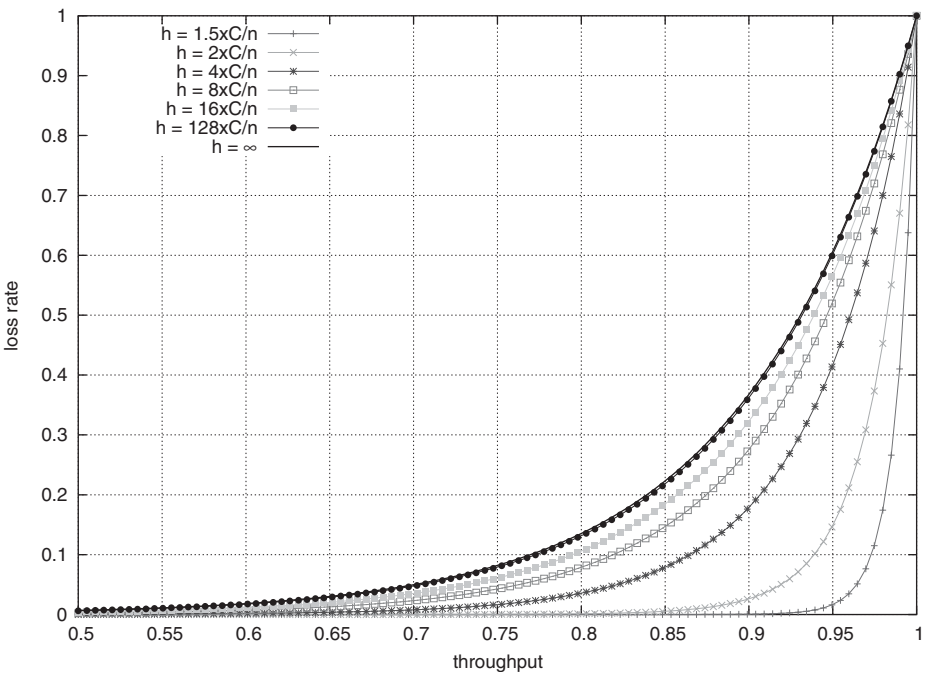

Figure 6. Relations between loss rate and throughput for different link speeds.

when $\varphi \leqslant e^{-k}$, because from (29) φ is a monotonically increasing function of ρ. So $Q_c(1 - kb/q_D, e^{-k})$ is the premium QoS region at $\varphi = e^{-k}$.

When $h = C/n$, from (26) $\delta = 0$, then $\varphi = 0$, $\forall \rho \leqslant 1$. So the premium QoS region at any $\varphi > 0$ is $Q_c(1, \varphi)$. □

Theorem 5.1 indicates that if the link speed is as low as C/n, the throughput can be arbitrarily high, and the loss rate can be arbitrarily low. The change of φ with h is shown directly in Figure 7. We see the increase of input link speed from C/n quickly damages the QoS performance. The figure also shows that higher the ρ is, the more sensitive the φ is to the change of h.

The reason why the link speed has critical effect on the node's QoS behaviour is this: the link can be viewed as a traffic shaper. If the link speed is too fast, it has no shaping effect at all, and the bursty traffic has a potential to 'stuff' the node. By contrast, if the link is slow, it can smooth the burst traffic before it is fed into the node. When h is as low as C/n, all bursts are completely smoothed. If h is even lower than C/n, however, the system efficiency decreases because the node processing power is wasted. In summary, the matching between the link speed and the node processing power produces the biggest premium QoS region and highest system efficiency.

The traffic shaping effect of the link has impacts on the effect of b. Keeping $\rho = 0.98$, Figure 8 illustrates that how φ increases with h for different b. In the figure, b is given in the form of w, as defined in formula (30). We can see that the smaller the b is, the less the h affects the φ. When b is small enough ($w \sim 100$), even big h does not worsen φ much. From the point of view of traffic shaping, a traffic shaper has a shaping region. It only smoothes the traffic that has a

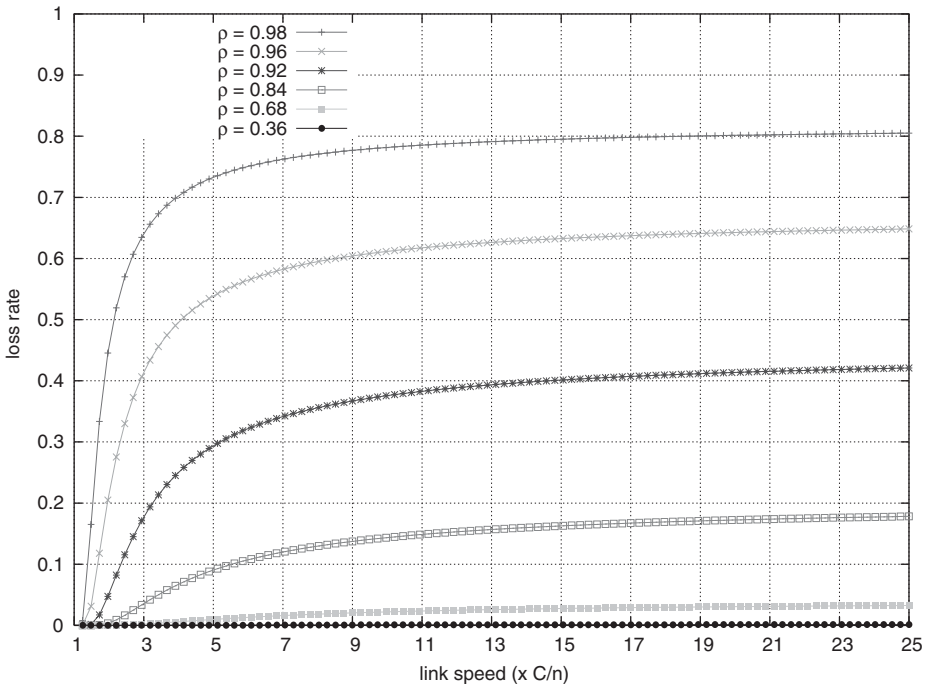

Figure 7. Relations between loss rate and link speed for different throughputs.

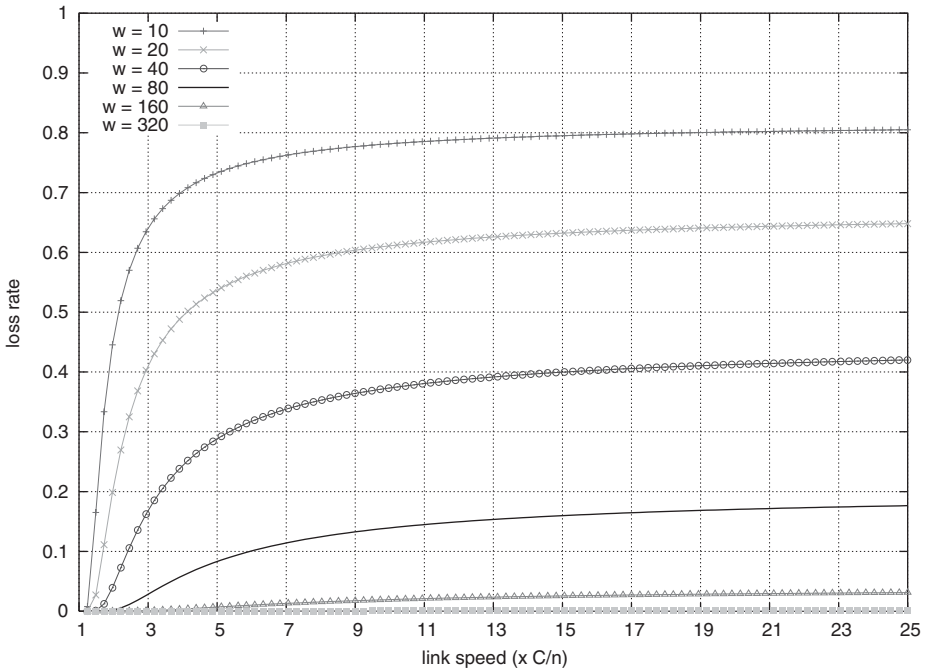

Figure 8. Relations between loss rate and link speed for different burst sizes.

bursty level above a threshold. When b is small, the traffic is already smooth. So the shaping of the link does not make much sense. This conforms to the result in Section 4.

The shaping effect of the link also helps enlarge the QoS regions. Denote

$$\eta = \frac{C/n}{h} \tag{41}$$

Obviously,

$$\eta \leqslant 1 \tag{42}$$

We call η the link utilization. When C/n is fixed, η can be used as an indicator of the link speed. The faster the link is, the lower η is. Rewrite formula (26) as

$$\delta = \frac{1}{(1-\eta)(1-\eta\rho)} \frac{1-\rho}{b} \tag{43}$$

Denote

$$\gamma = \frac{1}{(1-\eta)(1-\eta\rho)} \tag{44}$$

We have

$$\begin{aligned} \gamma &\to \infty, \quad \text{when } \eta = 1 \\ 1 < \gamma &< \infty, \quad \text{when } 0 < \eta < 1 \\ \gamma &= 1, \quad \text{when } \eta = 0 \end{aligned} \tag{45}$$

We call γ the shaping factor. It is a good indicator of the shaping effect of a link. $\gamma \to \infty$ means the traffic is completely shaped; $\gamma = 1$ indicates there is no shaping effect. From (10), (43), and (45), we get

$$\varphi \approx e^{-\frac{\gamma}{b} q_D(1-\rho)} \tag{46}$$

Assume the kth QoS region is ρ'_k when the shaping factor is γ. So we get

$$-\frac{\gamma q_D(1-\rho'_k)}{b} = -k \tag{47}$$

Comparing (47) with (32) we get

$$\frac{1-\rho'_k}{1-\rho_k} = \frac{1}{\gamma} \tag{48}$$

With this formula we can calculate the enlarged kth QoS region. This formula applies to any order of QoS regions. As an example, when $\eta = 0.2$, if the old kth QoS region has a size $\rho_k = 0.8$ (whatever k is), then the new kth region is $\rho'_k \approx 0.87$. Note that γ is ρ related when applying the formula.

Above we have shown the critical effect of the link speed on the QoS behaviour. However, it should be pointed that in real networks the contribution of the link as a traffic shaper may be very limited. In fact, the shaping factor γ decreases very quickly with the increase of the link speed. Figure 9 shows the change of γ with h. We can see that for $\rho = 0.84$ γ decreases from ∞ to only 3.45 as the link speed doubles from C/n (or η decreases by half from 1.0). It is common

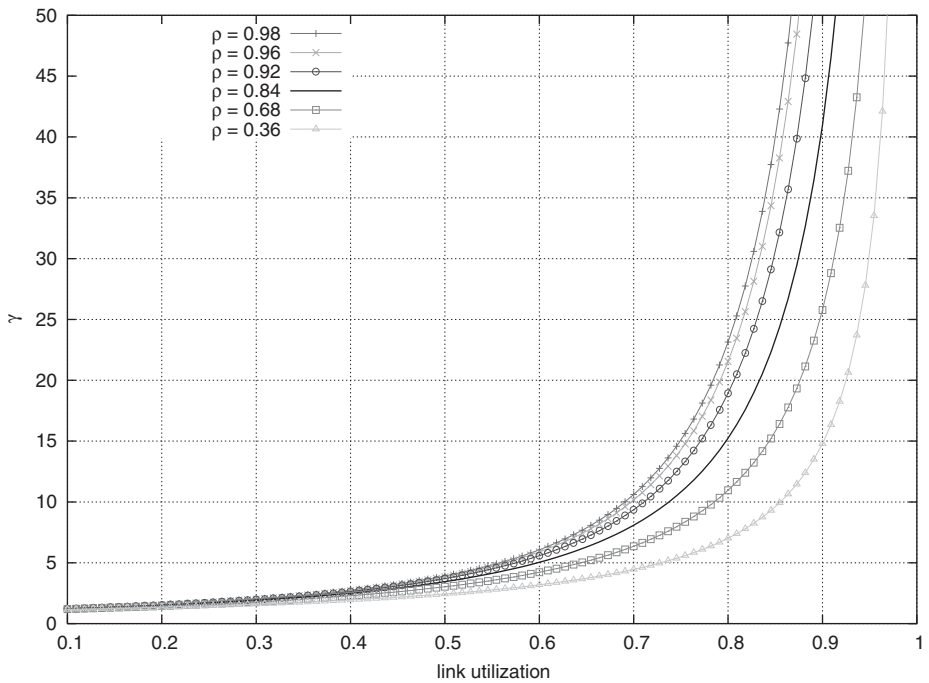

Figure 9. Changes of shaping factor with link speed for different network utilities.

that in core networks η is less than 0.5. So the shaping effect of the link is not always visible, and the results for $h \to \infty$ in Section 5 are generally good approximations in real networks.

6. EFFECTS OF TRAFFIC AND LINK HETEROGENEITIES

In this section we will see the QoS behaviour for heterogeneous traffic and heterogeneous links.

6.1. Load imbalance

Theorem 6.1
If all input links have the same speed and their traffic has the same burst size, then load imbalance among the links does not change the QoS region of the node.

Proof
We consider the effect of load imbalance on QoS region by comparing it with the case of load balance. Denote the total traffic load as R. In the case of load balance, the load on each link is

$$r = \frac{R}{n} = \frac{b}{T} \tag{49}$$

Thus

$$T = \frac{nb}{R} \tag{50}$$

We know that the arrival of bursts on each link is a Poisson process. With the superposition property of the Poisson process [20], the superposition of n independent Poisson processes is still a Poisson process. The average interval between successive bursts in the overall traffic is

$$T_{\text{all}} = \frac{T}{n} = \frac{b}{R} \tag{51}$$

In the case of load imbalance, we assume n_1 links out of n each has a traffic load of r_1, and the other $n_2 = n - n_1$ links each has a traffic load of r_2, where $r_1 \neq r_2$. But the total load is the same, namely,

$$R = n_1 r_1 + n_2 r_2 = n_1 \frac{b}{T_1} + n_2 \frac{b}{T_2} \tag{52}$$

Again, from the superposition property of Poisson process the overall traffic is a Poisson process. The inter-arrival time of the resulting traffic, T'_{all}, satisfies the following relation:

$$\frac{1}{T'_{\text{all}}} = \frac{n_1}{T_1} + \frac{n_2}{T_2} \tag{53}$$

But from (52) we know

$$\frac{n_1}{T_1} + \frac{n_2}{T_2} = \frac{R}{b} \tag{54}$$

So

$$T'_{\text{all}} = \frac{b}{R} = T_{\text{all}} \tag{55}$$

This means that we cannot really distinguish the overall input traffic in the load imbalance scenario from that in the load balance scenario. They are statistically identical (the traffic is fully characterized by b and T). Therefore, load imbalance between input links does not affect the system's QoS region. □

6.2. Burst size heterogeneity

Assume the traffic on n_1 links has a burst size of b_1 and the traffic on the other $n_2 = n - n_1$ links has a burst size of b_2. Without losing generality, suppose $b_1 > b_2$. We will analyse whether this burst size heterogeneity affects the QoS behaviour. Assume all links have equal speed h and equal traffic load r. Denote

$$\beta = \frac{n_1}{n} \tag{56}$$

So

$$n_1 = n\beta \tag{57}$$

$$n_2 = n - n_1 = n(1 - \beta) \tag{58}$$

Obviously, when $\beta = 0$ or 1, traffic on all links is homogeneous with burst size b_2 or b_1. It is easy to see that their kth QoS regions have the following relation:

$$Q_k|_{\beta=0} = Q_k|_{b=b_2} \supset Q_k|_{\beta=1} = Q_k|_{b=b_1} \tag{59}$$

More specifically, with (35), the Q_k's sizes have the following relation

$$\frac{1 - \rho_k|_{\beta=0}}{1 - \rho_k|_{\beta=1}} = \frac{b_2}{b_1} \tag{60}$$

For the same load, from (10) and (26) the loss rates differ in this way

$$\frac{\ln \varphi|_{\beta=0}}{\ln \varphi|_{\beta=1}} = \frac{b_1}{b_2} \tag{61}$$

When β increases from 0 to 1, the overall traffic is a mixture of bursts of size b_1 and b_2. The packet loss rate φ changes from $\varphi|_{\beta=0}$ to $\varphi|_{\beta=1}$. It turns out that the relation between φ and β when $0 < \beta < 1$ is a very complex non-linear one, which we will address elsewhere. In most practical cases, φ falls between $\varphi|_{\beta=0}$ and $\varphi|_{\beta=1}$. Hence we can roughly evaluate the effect of burst size heterogeneity from $\varphi|_{\beta=0}$ and $\varphi|_{\beta=1}$. From (60) and (61) we can see that if $b_2 \ll b_1$, the difference between the two QoS regions are big. So the degree of heterogeneity has significant affect on the QoS behaviour. If, however, b_2 and b_1 are both very small and comparable, the QoS behaviour would be rather insensitive to the change of β.

6.3. Link heterogeneity

Suppose n_1 links each has a speed of h_1, and the other $n_2 = n - n_1$ links h_2. Without losing generality, let $h_1 > h_2$. Assume the traffic is homogeneous on all links. Again, denote $\beta = n_1/n$. Similarly, we get

$$Q_k|_{\beta=0} = Q_k|_{h=h_2} \supset Q_k|_{\beta=1} = Q_k|_{h=h_1} \tag{62}$$

For the same load, from (10) and (26) the relations between the loss rates are not linear, but the following holds:

$$\varphi|_{\beta=0} < \varphi|_{\beta=1} \tag{63}$$

When β increases from 0 to 1, φ changes from $\varphi|_{\beta=0}$ to $\varphi|_{\beta=1}$ in a very complex non-linear way. But for practical cases, we can still view $\varphi|_{\beta=0}$ and $\varphi|_{\beta=1}$ as the bounds of φ. From Figure 7 we can see that if h_2 is small compared with C/n but h_1 is very big, the difference between from $\varphi|_{\beta=1}$ and $\varphi|_{\beta=0}$ is large. Then the QoS behaviour is sensitive to the change of β. If, however, both h_1 and h_2 are big, the heterogeneity does not matter much. The effect of link heterogeneity also depends on the throughput. It is important only when the traffic load is high. The burst size b modulates the effect of link heterogeneity, too. As we know from Section 3.1, if b is small enough, the loss rate φ is very low even for $h \to \infty$. In that case the link heterogeneity is not important.

7. AN EXAMPLE

Figure 10 shows a part of a backbone network where optical links of speed 2.5 Gbps are connected to 100 Mbps fast Ethernet networks through router A, B, and a link of 155 Mbps. Assume there are ten 2.5 Gbps links connected to router A. In general, router A is a communication bottleneck. Assume the network is DS-capable and router A implements two EF PHBs [21], one for voice service and one for video service. As defined in Reference [21], an EF PHB is a router mechanism in the DS network to support real-time services. It ensures that the EF packets are serviced at a given output interface with a rate not less than their arrival rate. In this sample, the bandwidth shares of the voice and the video services are 10 and 50 Mbps, respectively. Each input optical link can collect up to 1 Mbps voice traffic and 5 Mpbs video traffic. We now analyse the QoS behaviours of router A for these services with the theory in this paper, and compare them with the simulation results with the simulator NS-2 [22].

It is reasonable to assign a nodal deadline $D = 3$ ms for the voice service, for normally the end-to-end deadline is in the range of 10–40 ms [13]. Set $b = 60$ bytes. From the network configuration we know $C = 10$ Mbps, $h = 2.5$ Gbps, and $n = 10$. From formula (9), the maximum allowable queue size is $q_D \approx 3840$ bytes. With (10) and (26), we get the relation

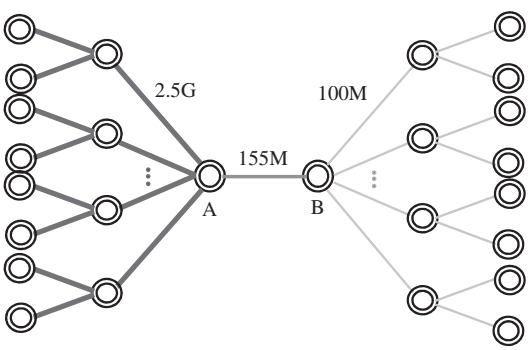

Figure 10. Sample network with EF PHBs for voice and video services.

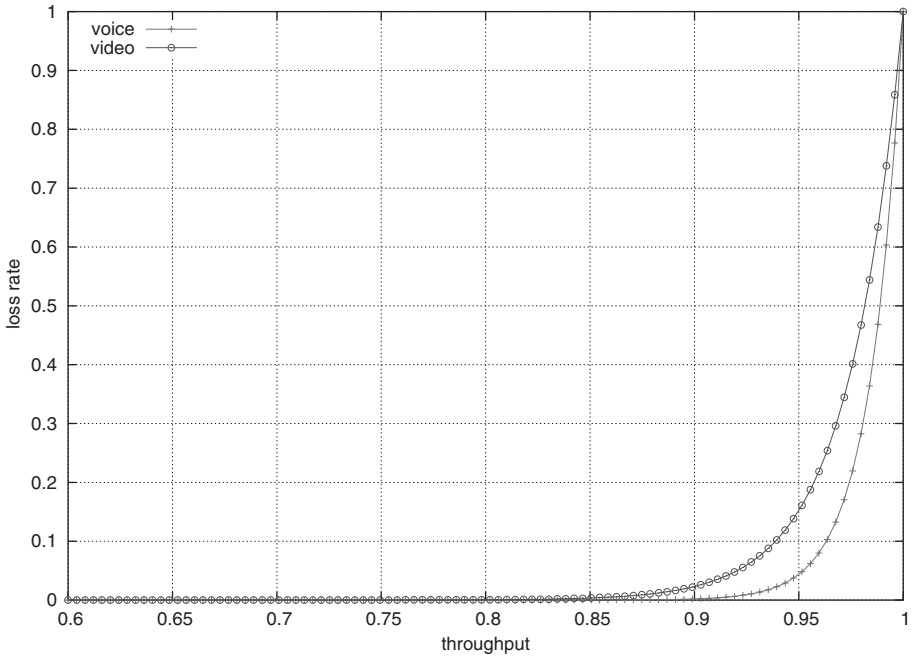

Figure 11. QoS behaviours for voice and video services.

between φ and ρ for the voice service:

$$\varphi \approx \mathrm{e}^{-\frac{62.525(1-\rho)}{1-0.0004\rho}} \tag{64}$$

It is illustrated in Figure 11. From (9) and (30) we can get $w = 62.5$. With formula (33) we know the size of the kth QoS region is

$$\rho_k = 1 - \frac{k}{62.5} \tag{65}$$

For example, $\rho_1 = 0.984$, $\rho_2 = 0.968$, $\rho_3 = 0.952$, and $\rho_4 = 0.936$. In particular, we notice that the premium QoS region at loss rate $\mathrm{e}^{-4} \approx 1.83\%$ is $Q_c(93.6\%, 1.83\%)$. So this system can provide very good QoS for the voice service.

Viewing a packet as a burst, with formula (35) we can see how different packet sizes affect the QoS.

$$\rho'_k = 1 - \frac{1-\rho_k}{60} b' = 1 - \frac{k}{62.5 \times 60} b' \tag{66}$$

Figure 12 shows the change of ρ_k with the burst size. We see that when the burst size is 300 bytes, ρ_4 decreases to 68%. This suggests that the maximum packet size should not be above 300 bytes to achieve reasonable QoS.

Now for the video service we choose the nodal deadline as $D = 6$ ms, and the average burst size $b = 1$ kB. The bandwidth share is $C = 50$ Mbps. In a similar way, we can get the QoS

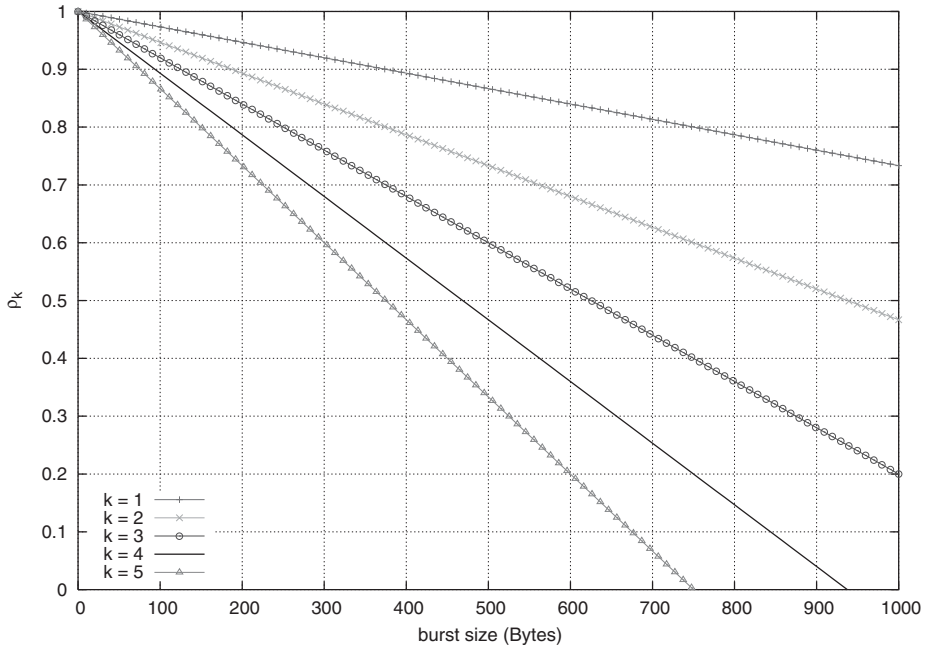

Figure 12. Changes of QoS regions with burst size for voice service.

behaviour for this service:

$$\varphi \approx \mathrm{e}^{-\frac{37.575(1-\rho)}{1-0.002\rho}} \tag{67}$$

It is also illustrated in Figure 11. The size of Q_k is

$$\rho_k = 1 - \frac{k}{37.5} \tag{68}$$

So we get $\rho_1 = 0.973$, $\rho_2 = 0.947$, $\rho_3 = 0.92$, and $\rho_4 = 0.893$. The premium QoS region at loss rate 1.83% is Q_c (89.3%, 1.83%). Though smaller than that of the voice service, it is still satisfactory. The change of ρ_k with the burst size is

$$\rho_k' = 1 - \frac{1-\rho_k}{1000} b' = 1 - \frac{k}{37.5 \times 1000} b' \tag{69}$$

It is illustrated in Figure 13. We see ρ_4 can increase to 98.7% if the packet size decreases to 500 bytes, which is excellent.

To validate above analyses, we compare the QoS behaviours with the simulation results. In the simulations, we use a class-based WFQ to share bandwidth among different services. There are totally three classes: the voice and the video services are two classes, and the rest traffic is viewed as the background traffic and holds another class. The weights among the classes are 2:10:19. Figure 14 gives the results for the voice service. In this simulation, the throughputs of the video service and the background traffic are kept as 90% and 95, respectively, while we change that of the voice. Figure 15 is for the video service. The voice and the background traffic throughputs are kept as 95 and 94% in this simulation. From these results we see that above

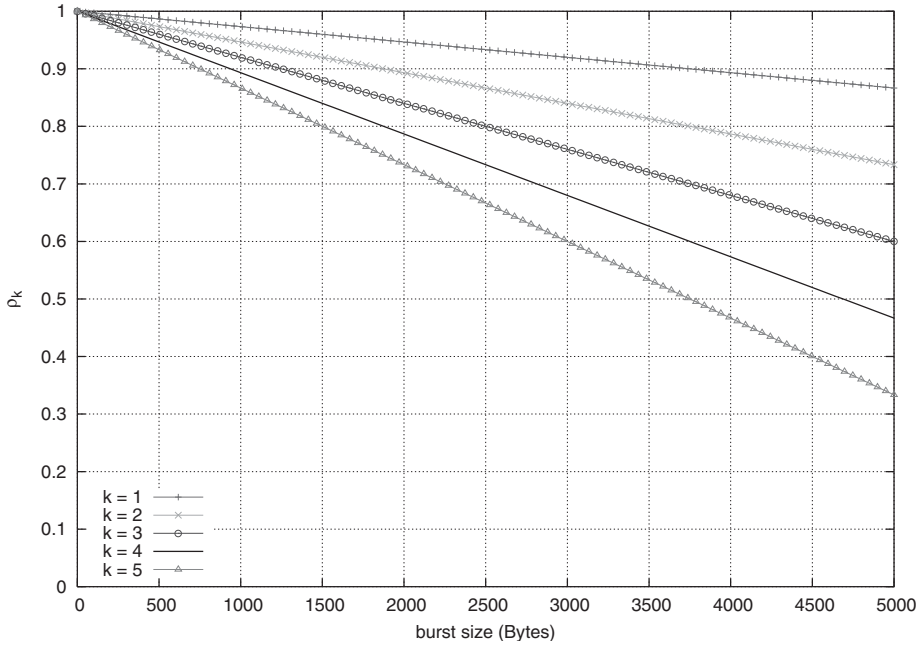

Figure 13. Changes of QoS regions with burst size for video services.

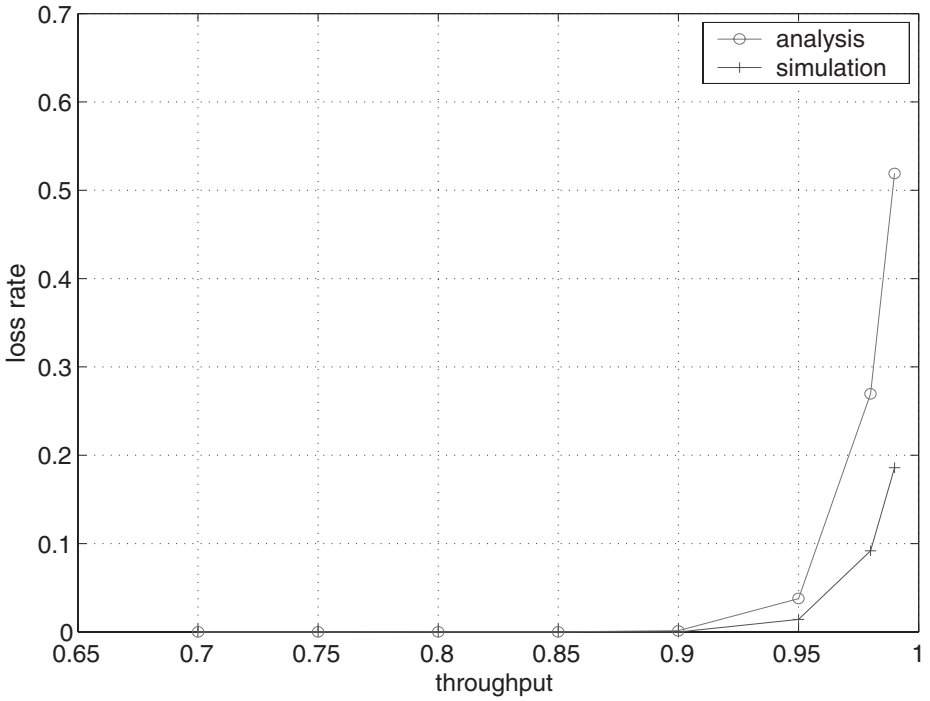

Figure 14. Comparisons of analytical and simulated QoS behaviours for voice service.

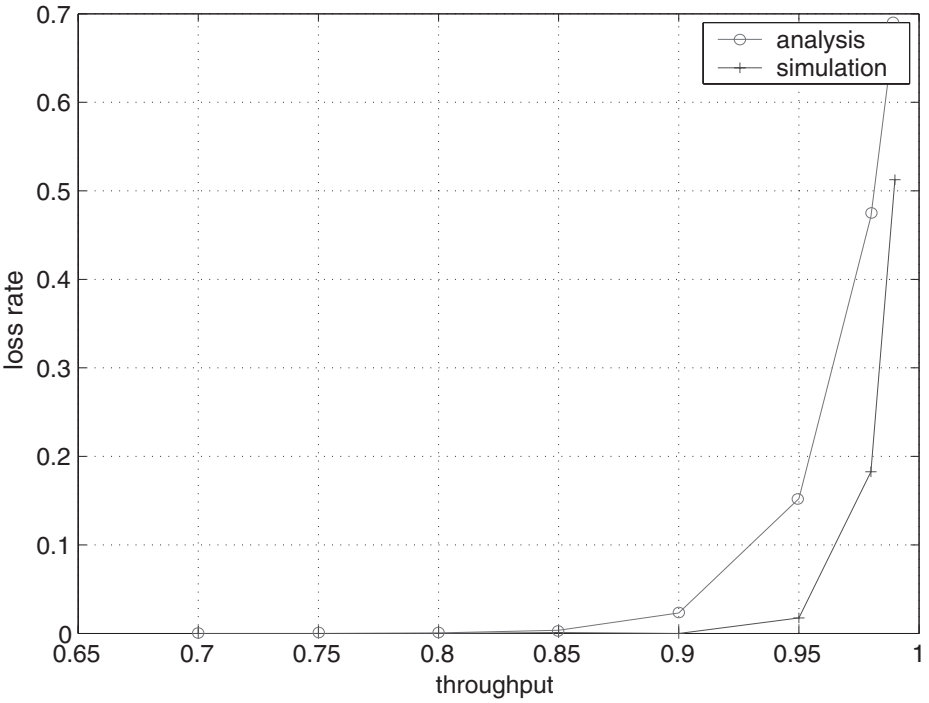

Figure 15. Comparisons of analytical and simulated QoS behaviours for video service.

QoS region analyses generally provide good higher bounds for the loss rates at a wide range of throughputs. The differences between the analyses and the simulation results at high loads are due to the processor-sharing gain of the WFQ [1]: the voice service in Figure 14 and the video service in Figure 15 get excess bandwidths from the rest services because the latter cannot use up their shares. So the performance bound given by the analysis can be surely guaranteed in practice. This suggests that our theory gives a reliable tool for network dimensioning to provide multi-dimensional QoS.

As for this particular example, above analyses indicate that in a practical network setting the rate configuration of the EF PHB defined in Reference [21] is generally sufficient for supplying good QoS for real-time services if the burst is well controlled. The extreme case mentioned in Section 1, which aroused much controversy and led to the redefinition of the EF PHB [23], can be tolerated by dropping the packets that violate their deadlines without affecting the multi-dimensional QoS satisfaction in general.

8. CONCLUSIONS

In this paper, we study multiple QoS dimensions altogether, and formulate a theoretical framework to explore relations between different dimensions. The QoS region is used to quantify multi-dimensional QoS requirements. Based on the theory of effective bandwidths, we reach a uniform formula to connect the throughput, the delay, and the loss rate for Markovian traffic. Important traffic and network factors, i.e. the burst size and the link speed, are involved. With this

framework, it is found that the burst size sets hard limit on the QoS region that can be achieved, and the matching between the link speed and the node processing power can greatly improve the limit. It is also made clear that while pure load imbalance among links does not affect the QoS region, the heterogeneities of burst size or link speed may severely degrade the multi-dimensional QoS performance. Applying the theory to real-time services in the DS architecture, we show that the analysis provides a useful tool for QoS prediction and network and traffic planning.

ACKNOWLEDGEMENTS

This work was partially supported by DARPA under contract No. N66001-00-C-8063. Part of the work was done when the first author was with the University of Cambridge, UK.

REFERENCES

1. Parekh A, Gallager R. A generalized processor sharing approach to flow control—the single node case. *INFOCOM'92*, vol. 2, Florence, Italy, May 1992.
2. Knightly EW, Shroft NB. Admission control for statistical QoS: theory and practice. *IEEE Network* 1999; **17**(2):20–29.
3. Floyd S, Handley M, Padhye J, Widmer J. Equation-based congestion control for unicast applications. *SIGCOMM*, August 2000.
4. Cruz R. A calculus for network delay, part I: networks elements in isolation. *IEEE Transactions on Information Theory* 1991; **37**(1):114–121.
5. Le Boudec J. Delay jitter bounds and packet scale rate guarantee for expedited forwarding. *INFOCOM 2001*.
6. Andrews M. Probabilistic end-to-end delay bounds for earliest deadline first scheduling. *INFOCOM'00*, Tel-Aviv, Israel, March 2001.
7. Reisslein M, Ross KW, Rajagopal S. Guaranteeing statistical QoS to regulated traffic: the single node case. *IEEE Infocom 1999*, New York, USA, March 1999.
8. Sivaraman V, Chiussi FM. Providing end-to-end statistical delay guarantees with earliest deadline first scheduling and per-hop traffic shaping. *INFOCOM 2000*, Tel Aviv, Israel, 2000.
9. Sivaraman V, Chiussi FM. Statistical analysis of delay bound violations at an earliest deadline first (EDF) scheduler. *Performance Evaluation* 1999; **36**(1):457–470.
10. Blake S, Black D *et al*. An architecture for differentiated services. *RFC 2475*, December 1998.
11. Dovrolis C, Stiliadis D, Ramanathan D. Proportional differentiated services: delay differentiation and packet scheduling. *ACM SIGCOMM '99*, Cambridge, Massachusetts, USA, 1999.
12. Alvarez J, Hajek B. Observations on using marks for pricing in multiclass packet networks to provide multidimensional QoS. *Conference on Information Sciences and Systems*, Princeton, NJ, USA, March 2000.
13. Charny A. Delay bounds in a network with aggregate scheduling. *Draft Version*, ftp://ftpeng.cisco.com/ftp/acharny/aggregate_delay_v4.ps, February 2000.
14. Anick D, Mitra D, Sondhi MM. Stochastic theory of data-handling system with multiple resources. *The Bell System Technical Journal* 1962; **61**(8):1871–1894.
15. Chang C, Thomas JA. Effective bandwidth in high-speed digital networks. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):1091–1100.
16. Courcoubetis C, Siris VA, Stamoulis GA. Application and evaluation of large deviation techniques for traffic engineering in broadband networks. *ACM SIGMETRICS '98/PERFORMANCE '98*, Madison, Wisconsin, June 1998.
17. Wischik D. The output of a switch, or, effective bandwidths for networks. *Queueing Systems* 1999; **32**:383–396.
18. Kelly FP. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications*, Kelly FP, Zachary S, Ziedins IB (eds). Oxford University Press: Oxford, 1996; 141–168.
19. Weiss A. An introduction to large deviations for communication networks. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):938–952.
20. Woodward ME. *Communication and Computer Networks: Modeling with Discrete-Time Queues*. Pentech Press Limited: London, 1993.
21. Jacobson V, Nichols K, Poduri K. An expedited forwarding PHB. *RFC 2598*, June 1999.
22. http://www.isi.edu/nsnam/ns/
23. Davie B, Charny A *et al*. An expedited forwarding PHB (Per-Hop Behaviour). RFC 3246, March 2002.
24. De Veciana G, Kesidis G, Walrand J. Resource management in wide-area ATM networks using effective bandwidths. *IEEE Journal on Selected Areas in Communications* 1995; **13**(6):1081–1090.

