Computer Networks 56 (2012) 3723–3743

Contents lists available at SciVerse ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Required extra capacity: A comparative estimation of overprovisioning needed for a classless IP backbone ☆

M. Yuksel [a,*], K.K. Ramakrishnan [b], S. Kalyanaraman [c,1], J.D. Houle [d], R. Sadhvani [e]

[a] University of Nevada – Reno, Reno, NV 89557, USA
[b] AT&T Labs Research, Florham Park, NJ 07932, USA
[c] IBM Research – India, Bangalore 560 071, India
[d] AT&T, Middletown, NJ 07748, USA
[e] Verizon Wireless, Basking Ridge, NJ 07920, USA

ARTICLE INFO

Article history:
Received 30 November 2011
Received in revised form 4 July 2012
Accepted 13 August 2012
Available online 21 August 2012

Keywords:
Net neutrality
Class-of-service
Economics
Performance
Dimensioning

ABSTRACT

The benefit of Class-of-Service (CoS) is an important topic in the "Network Neutrality" debate. As part of the debate, it has been suggested that over-provisioning is a viable strategy to meet performance targets of future applications, and that there is no need for to worry about provisioning differentiated services in an IP backbone for a small fraction of users needing better-than-best-effort service. In this paper, we quantify the extra capacity requirement for an over-provisioned classless (i.e., best-effort) network compared to a CoS network providing the same delay or loss performance for premium traffic. We first develop a link model that quantifies the *required extra capacity* (*REC*). To illustrate key parameters involved in *analytically* quantifying REC, we start with simple traffic distributions. Then, for more bursty traffic distributions (e.g., long-range dependent), we find the REC using ns-2 simulations of CoS and classless links. We, then, use these link models to quantify the REC for network topologies (obtained from Rocketfuel) under various scenarios including situations with "closed loop" traffic generated by many TCP sources that adapt to the available capacity. We also study the REC under link and node failures. We show that REC can still be significant even when the proportion of premium traffic requiring performance assurances is small, a situation often considered benign for the over-provisioning alternative. We also show that the impact of CoS on best-effort (BE) traffic is relatively small while still providing the desired performance for premium traffic.

1. Introduction

The Internet has been a tremendous success with large numbers of day-to-day communication applications migrating to it, as network connectivity and web-usage have become nearly ubiquitous. With the massive deployment of residential broadband access, user expectation of Internet services has moved from best-effort (BE) connectivity to having reasonable performance and capacity for all types of applications. Entertainment is also viewed by many as a major application area taking advantage of IP networks. Consumers would prefer to use a converged single "pipe" for all their communication and entertainment needs, if possible. Performance-sensitive applications like video [3], games, and voice-over-IP (VoIP) are offered over such a converged end-to-end IP network [4]. To satisfy requirements posed by such applications, Internet Service

☆ Preliminary versions of this work, appeared in [1] as a short paper, and in [2].

* Corresponding author. Tel.: +1 775 327 2246.
E-mail addresses: yuksem@cse.unr.edu (M. Yuksel), kkrama@research.att.com (K.K. Ramakrishnan), shivkumar-k@in.ibm.com (S. Kalyanaraman), jdhoule@att.com (J.D. Houle), rita.sadhvani@verizonwireless.com (R. Sadhvani).

[1] Dr. Kalyanaraman was with the ECSE, Department of Rensselaer Polytechnic Institute, Troy, NY during the initial part of this work.

1389-1286/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.comnet.2012.08.007

Providers (ISPs) need to provision their networks to meet the service level agreements (SLAs) of their business customers, despite high and variable amounts of background traffic from all customers. Such provisioning must also be resilient to changes in customer demand, changes in application mix and network failures. Customer experience needs to be protected and predictable despite failures causing traffic re-routing. ISPs indeed provision their networks conservatively based upon predicted loads to provide good service throughout, which makes most of over-provisioning a practice of protection capacity.

There has been a wide ranging debate on the issue of "Network Neutrality" which involves both economic and technical aspects [5,6]. One key technical aspect of the debate is whether best-effort application traffic should be carried along with other (so-called "premium") traffic for which SLA commitments have been made (or are expected, either explicitly or implicitly) without differentiation. At one end of the opinion spectrum in the debate, it is suggested that there should be no differentiation of traffic and all performance requirements should be met only by over-provisioning the network. The question, then, is whether this can be done with a small amount of additional capacity or is there a need to significantly over-provision the network? Our study focuses on this specific question. We compare a classless network which is over-provisioned against an engineered network using per-class queuing to offer Class-of-Service (CoS) (i.e., differentiated service) and meet user expectations and SLAs. We recognize that Quality-of-Service (QoS) has been extensively studied for many years. However, quantifying the extent of over-provisioning needed in a network to match the SLAs achieved by a corresponding differentiated network still needs to be addressed.

The hypothesis of this paper is that an over-provisioned single-queue network service for meeting the SLAs of performance-sensitive traffic and regular best-effort traffic is inefficient (from a capacity viewpoint) compared to an engineered network offering *simple* 2-queue CoS differentiation. Though this basic fact is known in the network performance evaluation community, our paper refines it to identify parametric regions where this inefficiency exists and is pronounced. We show that *this inefficiency is significant even for moderate utilizations and becomes particularly pronounced when premium traffic is a smaller fraction of the overall traffic mix*.

We model the basic SLA requirements that applications may need, in terms of delay or loss. We then estimate the *required extra capacity* (*REC*) for a classless link to match the performance (in delay or loss) provided by its CoS-based correspondent. We generalize this single link model to an ISP network taking into account the network topology, traffic matrices (based on a gravity model), and shortest path routing.

We recognize that a dominant amount of traffic on the Internet is from TCP traffic generated by applications that are adaptive and not as sensitive to delay as application traffic carried over UDP. The TCP source also adapts to the available capacity, thereby reducing loss to the extent possible, when operating in either a CoS or classless environment. We examine the REC in these cases, where both the best-effort (BE) and premium class traffic use TCP as well as the case when only the BE class traffic uses TCP. Because TCP is mainly sensitive to loss, the emphasis of our study of the REC with TCP traffic is primarily on achieving the same packet loss probability with a classless network as with a CoS network.

The rest of the paper is organized as follows: In the next section, we first describe our modeling framework. Then, to quantify REC, we detail link models for Poisson, MMPP, and LRD traffic cases. By considering both delay and loss as the performance targets, we provide a detailed discussion of the link models' implications in Section 4. In Section 5, we extend the link models to a network model using sample ISP topologies. For performance ranges pertaining to some legacy applications, we present the REC results for these ISP topologies, followed by conclusions.

1.1. Contributions and Findings

Understanding the potential benefits of CoS provisioning and how best to dimension a network to accommodate future growth in demands are topics explored extensively by researchers. In this paper, we provide a different perspective on *the value of providing CoS as the fraction of premium traffic changes*. Building on this basic perspective, we provide estimations on how much extra capacity would be required (i.e., required extra capacity (REC)) at the link as well as network levels, if the fraction of premium traffic varies.

Our modeling effort develops an analytical and simulation-based link model of REC, and then extends the link model to a network model to provide estimates of "network REC". Based on the quantitative modeling of several ISP topologies obtained from Rocketfuel [7], we extensively study REC within a *network* setting. The primary component of our modeling effort that is novel is the network level model.

Our major contributions and findings are as follows:

- *Analytic framework for quantifying REC*. We provide a simple analytic framework for quantifying the REC as a function of the proportion of the traffic requiring premium service, the utilization, and the performance target.
- *Quantifying REC in a network setting*. We outline a systematic way of extending the link level REC estimation to a network setting.
- *REC is higher when the proportion of traffic using the premium class is smaller*. Our network model analysis provides insight into the *significant REC needed even when the proportion of premium traffic is small*. This result invalidates the common knowledge that significant over-provisioning is only needed when the premium traffic is a large proportion of the overall network traffic. Higher utilization and traffic burstiness make REC even higher.
- *Effect of long-range dependent (LRD) traffic*. We show that traffic patterns with more burstiness, as in LRD traffic, cause REC to increase by orders of magnitude in comparison to short-range dependent traffic like a Markov-Modulated Poisson Process (MMPP) [8].

LRD models are known to approximate the Internet traffic at various time-scales [9–11]. Our results give a glimpse of the increased capacity requirement as burstiness increases.

- *Effect of closed-loop (i.e., TCP) traffic.* We found that when the sources are adaptive, using TCP for both the BE and premium class traffic, the REC for the classless network is still significant. We ensure that an appropriate share of the resource is available for the BE class (rather than being starved out by a pure-priority service) using a reasonable deficit-round-robin [12] scheduling algorithm. We also show that when the premium class traffic is non-adaptive and the BE traffic is TCP, the extent of REC grows even more.
- *Impact on performance of the BE class*. Previous work has suggested [13] that a lower priority class may be undesirably affected by a higher priority class. We quantitatively show this impact, and observe that, with the differentiation that CoS support provides, the loss and delay impact on BE traffic is relatively small while still providing the desired performance for premium traffic.
- *Failure analysis*. Our network-level analysis includes quantifying the REC when a link or node fails. We show that substantial additional capacity may be needed across a significant number of links to support the traffic being re-routed due to failures. And, REC due to the incremental effect of failures over-and-above the regular burstiness of traffic is significant.

1.2. Related work

There is a vast body of network QoS literature studying different queueing, scheduling and buffer management mechanisms to allocate finite capacity and delay (given an average utilization) amongst flows at a statistically multiplexed resource [14,15]. Recent work by Ciucu et al. [16] proposes a provisioning strategy based upon statistical service curve characterization and argues that scheduling has little value added above such provisioning, when traffic is shaped. In our work, a key difference is that we do not have admission control or shaping/policing of input traffic. But, the network must still honor premium-traffic SLAs. In this context, we show that simple CoS scheduling is still valuable. CoS classes tend to be reprovisioned, but over longer time-scales (min/h) in response to aggregate demand pattern changes. The flow-aware networking approach [17,18] suggests the use of implicit differentiation by using per-flow queuing and per-flow admission control. In contrast, our work focusses on a simple 2-class vs. 1-class model at the aggregate level without admission control.

Consideration of the cost-effectiveness of over-provisioning as an architectural paradigm was done earlier, e.g., [19]. Recent work [20–22] examined the benefit of over-provisioning to overcome traffic and revenue uncertainty and to accommodate scaling up of the network, while we examine the relative benefit of CoS support in terms of capacity savings. An analysis similar to ours was done by Sahu et al. [23] in comparing loss performance of forwarding behaviors (i.e., discard eligibility vs. priority) of the DiffServ architecture. Instead of services specific to the DiffServ architecture, our work compares the classless service to the class-of-service environment in general. We provide the quantitative comparison at the edge-to-edge (g2g) level with full consideration of network-specific issues such as the topology and the traffic matrix. Rather than the end-to-end (e2e) performance characteristics, we focus on the g2g performance requested of a backbone ISP within the context of classless vs. CoS provisioning. To answer the question of "Would there be need for a large amount of extra capacity in order to provide service to all traffic with a performance similar to the premium class quality in a CoS network?" from an ISPs point of view, g2g is a reasonable granularity to work with as typical SLAs are made over g2g performance measures.

Kelly [24] argues that queueing delays may become small in comparison to propagation delays at higher link speeds and suggests that differentiation between traffic classes may become redundant. His study assumes that the dominant end-to-end protocols are TCP-like adaptive protocols, which are primarily sensitive to loss. However, we believe that there will be aggregation of a large number flows at the core of IP backbones and the overall performance will be dependent more on the aggregate behavior rather than on individual flow behavior. Further, recent trends clearly show that there is a significant and growing amount of RTP/UDP-based non-adaptive audio, video, and gaming traffic on IP backbones [4].

In a similar vein, Gibbens et al. [13] conclude that differentiated forwarding is unlikely to provide significant performance distinction unless the higher class traffic damages the lower class. Their focus is primarily on loss, assuming that TCP is the dominant protocol. However, we examine the impact on delay in addition to loss. We show quantitatively that the increase in loss for BE in a CoS network is negligibly small compared to a classless environment running at the same utilization. We believe this to be the case even if TCP is the dominant protocol, especially at higher utilizations. Further, we show that premium services could be provided without necessarily hurting the lower class performance significantly, even for medium utilizations with the proportion of premium class traffic being small, which is the current operating region for the Internet.

2. Model Framework

Our model framework enables a comparison of the capacity required for a classless service vs. a Class-of-Service (CoS) network with two classes for various ISP topologies. We start with a simple comparative model of the two services at the link level, which poses the question: "How much extra capacity needs to be provisioned for the classless service to meet the same *performance* (e.g., in terms of delay or loss) as the premium class traffic achieves in a CoS link with the same *aggregate (including both premium and best-effort (BE)) traffic load*?" To substantiate this link model, we develop the relationship between the required extra capacity (REC) and the following

parameters: premium class performance (delay or loss) and aggregate traffic load. We then extend this link-level model to a network model where edge-to-edge (g2g) premium class performance goals are built upon link-level performance goals through g2g paths. This enables us to use the link-level REC model for calculating the needed extra capacity for each link in the network.

A key difference in our model from the existing work in the literature is that we let the performance target move rather than fixing it to a particular value. Our model takes the performance achieved by the premium class of the CoS as the basis and searches to answer the question: How much more overprovisioning would we have to do with a classless service if we were to provide the same performance to all traffic? As it will become more clear later, this model captures all "possible" cases in terms of the performance target. By allowing a moving performance target, we automatically eliminate the infeasible performance values (i.e., tighter than what can be achieved by a premium class). Thus, our model is generic in quantifying the REC, and finding the REC values for a particular performance target is only a lookup from our model. Further, this allowing a moving performance target is a more fair way of comparing the CoS and classless cases in terms of quantifying the inefficiency of classless service, which is the main goal of this paper.

2.1. Preliminaries

We start by considering two traffic classes on a CoS link: *premium class* and *best-effort class*. We set a performance target of delay or loss for the premium traffic on the CoS link, and then seek to find the required extra capacity (REC) for a classless link (which treats both traffic classes equally) to achieve the same performance target for both the traffic classes. Fig. 1 illustrates the comparison of the two cases at the link level. Let the aggregate traffic rate be λ_D to be served by a CoS link with a capacity of μ_D. Also let a fraction of this aggregate traffic be premium class traffic with a rate of $\lambda_{Prem} = g\lambda_D$ while the remaining is best-effort (BE) class traffic with a rate of $\lambda_{BE} = (1-g)\lambda_D$. For the premium class traffic, we define a performance target ζ, in terms of delay or loss.

Given the parameters as illustrated in Fig. 1, we formulate the necessary classless link capacity μ_N to achieve the same performance target ζ for the aggregate traffic λ_D. From this, we can calculate REC for the classless link in terms of *rate* as $\mu_N - \mu_D$ (or as a *percentage* $100(\mu_N/\mu_D - 1)$). In this model, we use average delay t_{target} or average loss probability p_{target} as the performance target. With loss, an additional parameter is the buffer size, which we express as K for each of the traffic classes in the *CoS link* and $2K$ for the aggregate traffic in the *classless link*. Notice that, for a fair comparison, we use the same total buffer of $2K$ in both the CoS and classless cases.

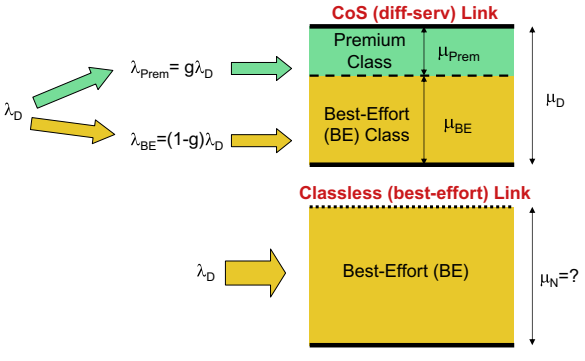


Fig. 1. Link-level comparison of two service types: CoS vs. classless.

Because *non-preemptive priority queuing* is a simple, analytically tractable packet scheduling policy for CoS support, we base our analysis on it. We note that our estimates of REC will be conservative when we assume that the aggregate traffic (i.e., premium + BE) exhibits the same relationship for the first two moments of the traffic (i.e., the relationship between the mean and the variance) in each class. That is, if we are modeling the premium class traffic with a Poisson process of rate $\lambda_{Prem} = g\lambda_D$ and the BE class traffic with a Poisson process of rate $\lambda_{BE} = (1-g)\lambda_D$, then we assume that the aggregate traffic for the classless service is also a Poisson process with rate λ_D.

One issue of concern here is that the superposition of two independent exponential streams yields a more bursty traffic than the burstiness of individual traffic streams [25,26]. Later in Section 3.2.6, we try relaxing our 'superposition of the two Poisson flows' assumption for the CoS case, and show that REC becomes higher when the premium traffic has a less bursty (e.g., deterministic) distribution than the BE traffic as well as the aggregate traffic. So, we use the same burstiness behavior for both the BE and premium classes to be on the safe side and stay conservative in our REC estimates. On another note, it is also possible that multiplexing of many streams may compose a less bursty aggregate stream if individual streams are non-independent and appropriately scaled [27]. Such reduction in burstiness will require aggregation of many streams to attain limiting effects and is not applicable to our model as our aggregation involves only two flows.

Different interpretations of "burstiness" exist in the traffic modeling literature. Generally speaking, burstiness refers to the bi-modal behavior of traffic where an on–off structure exists in the arrival process. Burstiness from an on–off arrival process can be captured by its counting process, i.e., the time series showing the number of arrivals per unit time. One can produce more bursty (in terms of the packet count) traffic by controlling the parameters of the on–off process or the number of on–off processes as was shown in [25,26]. Since our work is centered around the queueing behavior and performance, we use the traffic streams counting process when quantifying the performance metrics such as delay or loss. So, as in [28], our usage of the term "bursty" refers to the amount of variance in the traffic streams counting process.

3. Link model

We develop our REC link model based on three different traffic models: Poisson traffic (to provide us an initial

analytically closed-form understanding), a Markov-Modulated Poisson Process (MMPP) [8], and long-range dependent (LRD) traffic. When estimating REC values for a network, later in Section 5, we use the MMPP-based link model for estimating the REC values on individual links. This is a "conservative" choice since a more variant/bursty traffic yields larger REC values, as also verified by our experiments in this paper. The counting process for a short-range dependent (SRD) traffic model such as MMPP has smaller variance in comparison to an LRD traffic model. Literature showed that the Internet traffic exhibits long-range-dependency and thus its counting process has more variance than an SRD model like MMPP or Poisson [28,29].

3.1. Analytical model: Poisson traffic

To illustrate the functional relationship between REC and the other parameters, we *analytically* derive equations for REC by assuming (for simplicity of analysis) Exponential service time (i.e., packet size) distribution. The Poisson traffic and MMPP traffic models allow us to develop initial analytical understanding of REC. In comparison to Poisson, the MMPP traffic is still short-range dependent but allows us to model more bursty scenarios. Since the MMPP traffic is analytically tractable and can be made more bursty than Poisson traffic, it also allows us to clearly observe the trends in REC behavior, as burstiness increases.

We model the case when traffic is assumed to be Poisson and the performance target is queueing delay (i.e., t_{target}) or packet loss probability (i.e., p_{target}). Let μ_N be the required capacity for the classless link to be able to match the premium class performance with CoS.

By using the M/M/1 and M/M/1/K relationships for non-preemptive priority queueing, we first derive (details of this derivation are in Appendix A) the performance (t_{target} or p_{target}) for the premium class achieved by the CoS link:

$$\begin{aligned} t_{target} &= \frac{1+(1-g)\rho}{1-g\rho} \\ p_{target} &= \frac{1-g\rho}{1-(g\rho)^{K+1}}(g\rho)^K \end{aligned} \tag{1}$$

where $\rho = \lambda_D/\mu_D$ is the aggregate traffic load at the CoS link, and K is the buffer size available to each of the two CoS classes separately.[2] Notice that we express the delay t_{target} in terms of "packets". For the rest of the paper, we will use this notion of delay, which is helpful especially for deriving conclusions on REC independent of the CoS link capacity μ_D and the average packet size. Thus, once we set values for g and ρ (and K for the case of loss), we also set the performance target, which is the performance achieved by the premium class.

Following the performance target, we formulate REC (in percentage) to match the performance in (1) with the classless service:

$$\begin{aligned} REC_{delay} &= 100\left[\frac{(1-g\rho)}{1+(1-g)\rho}+\rho-1\right] \\ REC_{loss} &= 100\left[\frac{\rho}{\sqrt{g\rho}}\sqrt[2K]{\frac{1-(g\rho)^{K+1}}{1-g\rho}}-1\right] \end{aligned} \tag{2}$$

Similar to the insensitivity results in queueing theory [30,31], observe that both the REC (Eq. (2)) and the performance targets (Eq. (1)) do not depend on the distribution governing the service times, and are driven by only two parameters: fraction of premium traffic g and link utilization ρ. This insensitivity of REC to the service times holds as long as the arrivals are Poisson and allows dimensioning of the REC results for a desired performance target easily. We will use this relationship to guide our observations and plots later. Specifically, we will display the performance targets (i.e., t_{target} and p_{target}) as shades of color/gray on graphs plotting REC vs. g and ρ.

3.2. Simulation-based model: MMPP and LRD traffic

To approximate a scenario similar to IP backbone links, we use LRD arrivals [28,29] and deterministic service time distributions [32]. To calculate REC under LRD traffic, we use a careful simulation-based method (detailed in Section 3.2.1) to *empirically* obtain the link-level estimates of REC. We apply the same method to calculate the MMPP-based REC link models. We parameterize the MMPP traffic generator so that it yields a conservative (less bursty than LRD). We later use the LRD and MMPP link model results to obtain estimates of REC in a network setting in Section 5.

To observe how REC behaves as the traffic becomes more bursty, we first examine the case when the traffic is characterized as an MMPP stream [8]. With an exponential service time distribution, the MMPP/M/1 model still allows for analytical derivation of REC to obtain insights into its behavior. The simplest MMPP traffic model can be developed by means of two states ($i = 1,2$) each corresponding to a particular sending rate λ_i of a Poisson process. Let the average sending rate of the overall MMPP traffic stream be λ_t and the sending rate of the first state be a fraction $0 < a < 1$ of the average rate (i.e., $\lambda_1 = a\lambda_t$), and the ratio of the traffic rates of the two MMPP states is $r = \lambda_2/\lambda_1$. Specifically, we set the two traffic rates as

$$\begin{aligned} \lambda_1 &= a\lambda_t \\ \lambda_2 &= ar\lambda_t \end{aligned}$$

where $0 < a < 1$ and $r > 1/a$. In order to compose a traffic stream with an average rate λ_t by using these MMPP state sending rates, we have to set the state probabilities, π_1 and π_2, as follows:

$$\begin{aligned} \pi_1 &= (ar-1)/(ar-a) \\ \pi_2 &= (1-a)/(ar-a) \end{aligned}$$

Note that the product $a\ r$ is a measure of burstiness, since both a and r increases the variance of the MMPP traffic [33]. Thus, the burstiness of the stream can be tuned via the parameters a and r.

Due to the exponential inter-arrival times, MMPP generates a short-range dependent traffic stream which is

[2] We choose equal buffer sizes for each class to make sure that we only quantify the effect of differentiation of forwarding.

conservative in terms burstiness in comparison to IP traffic [28,29]. We use MMPP traffic to make a conservative estimate of REC. Further, we used LRD traffic sources since the literature suggests that the Internet traffic (beyond timescales of RTTs) can be modeled as LRD with the Hurst parameter ranging roughly from 0.75 to 0.9 [10] (higher values imply more burstiness). Also, we use deterministic packet service time distributions for simplicity (considered reasonable based on IP traffic [32]). We used the LRD traffic generator from [34], which used aggregation of many MMPP streams to establish long-range dependence in the traffic [25]. The next subsections present the details of the simulation-based calculation of REC and the results for MMPP/M/1, MMPP/M/1/K, LRD/D/1, and LRD/D/1/K link models.

3.2.1. Link model simulation and validation

In order to obtain an accurate link model for the LRD and MMPP cases that are then used in the network-level analysis, we used ns-2 simulations to calculate the REC. We simulated both the CoS link and the classless link for various ρ and g values, and matched the empirical performance of the premium class in the CoS link to the empirical performance of the aggregate traffic on the classless link. To simulate the CoS link, we used non-preemptive priority queuing of the flows from the two classes being served by the link. For the classless link simulation, we used a FIFO queue for the aggregate flow, which is the superposition of the two flows of the CoS case.

In order to find the REC values by simulation, we matched the performance (i.e., delay or loss probability) experienced by the premium class flow in the CoS link with the one experienced by the aggregate flow over the classless link, within a 1% margin of error. We first simulated the CoS link for a given capacity (e.g., $\mu_D = 10$ Mb/s), utilization $\rho = \lambda_D/\mu_D$, fraction of premium traffic $g = \lambda_{Prem}/\lambda_D$, and buffer size K (if loss probability is the performance target). This empirically gave us the performance goal, i.e., t_{Prem} or p_{Prem}. We then matched this performance goal in the classless link simulations, i.e., $t_{target} = t_{Prem} \pm 1\%$. To find the classless link capacity μ_N required to match to the premium class performance, we iteratively updated μ_N and observed whether the classless service performance matches that of the premium class traffic in the CoS case. Fig. 2 shows the detailed flowchart of this procedure of searching the μ_N value that matches the performance of the premium class. This search procedure involves two phases. First, we increase the μ_N value to find a maximum bound for it. Then, in the second phased, we apply a binary search for a μ_N value that matches the performance of the premium class in the CoS case.

To gain confidence, we repeated the classless and CoS link simulations (shown as sharp rectangles in Fig. 2) and used the average observed performance across repetitions. Specifically, we repeated simulations 6 times for MMPP traffic and 72 times for LRD traffic. We kept the simulation length at 5000 s for MMPP traffic and 20,000 s for LRD traffic. Fig. 4a and b show the 95% confidence interval of the target delay values for MMPP/M/1 and LRD/D/1 link models respectively. Confidence intervals for the other link model simulation were similar to the ones in Fig. 4.

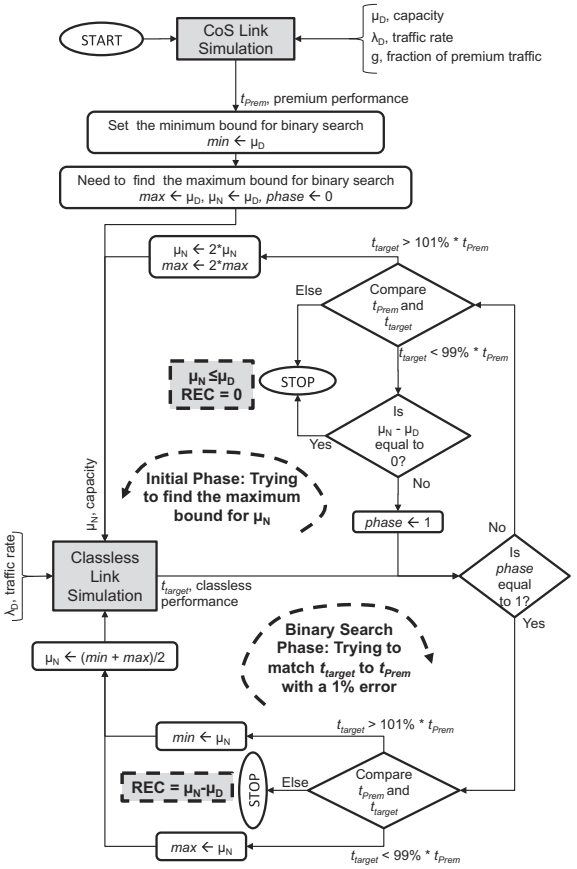
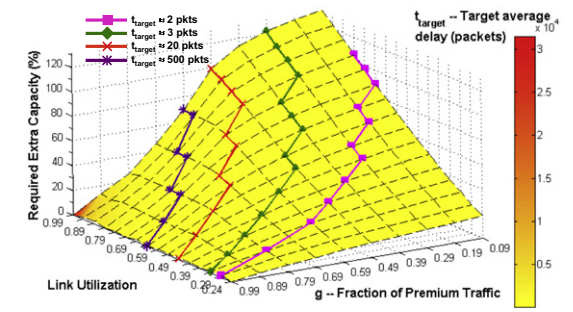


Fig. 2. Simulation-based link model: the performance of premium class is empirically matched with the performance of a classless service. The capacity of the classless link (i.e., μ_N) is increased or decreased based on a binary search algorithm to find the classless link capacity that gives a performance within 1% of what was achieved by the premium class in the CoS link. The gray boxes with dashed lines show the two possible endings for the algorithm.

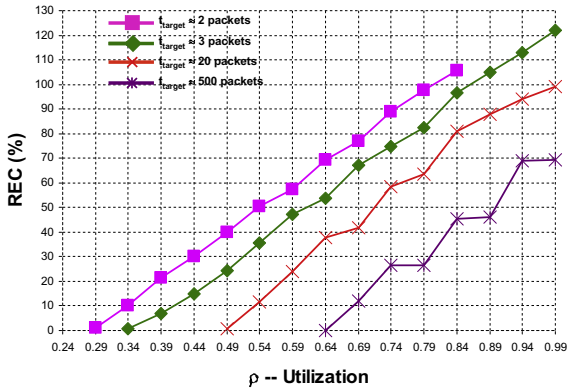
3.2.2. Achieving a delay target: MMPP/M/1

By using simulation-based estimation of REC as described in Section 3.2.1, we obtained REC values when the average delay is the performance goal. For selected burstiness cases (i.e., a and r), the graphs in Fig. 3a and Fig. 5 plot REC as a function of link utilization ρ and premium traffic fraction g. The darkness of the REC surface shows the target delay (i.e., t_{target} also a function of ρ and g) in terms of the number of packet service times. As we go from Figs. 3a–5, the increased burstiness of the MMPP traffic (i.e., the product ar) also causes the REC to become significantly larger. To make the relationship clearer, Fig. 3b and c plot REC in two dimensions against ρ and g respectively, while keeping the performance target fixed. So, the 3D graphs exhibit the REC trends as ρ and g varies. To see the quantitative nature of REC for a given performance target, we look at the "fixed-performance target" lines on the 3D surface.

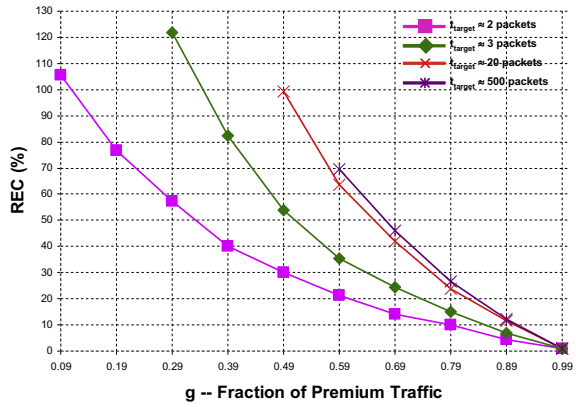
As we see in Fig. 3b and c, the REC grows as the link utilization becomes higher, but more so when the fraction of premium traffic g is smaller. On the other hand, when the traffic is predominantly of the premium class, there is less



(a) REC vs. ρ and g (in 3D)



(b) REC vs. ρ for fixed t_{target} (in 2D)



(c) REC vs. g for fixed t_{target} (in 2D)

Fig. 3. DELAY – MMPP/M/1 link model with $a = 0.5$, $r = 4$: The darkness (color) of the surface in (a) shows the target average delay, normalized in units of packets. For example, "1000 packets of delay" equals to 819.2 ms and 8.192 ms delay respectively for 10 Mbps and 1 Gbps links carrying packets of size 1 KB. The colored lines, which are also shown in (b) and (c), on the surface roughly show the points where the target delay is of a certain number of packets, irrespective of link speeds. For example, the purple line shows 1000 packets of delay. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

benefit from the differentiation. When the proportion of premium class traffic is small, an arrival of that class at a classless queue would have to be serviced quickly. Thus the classless service would require a higher service rate than the CoS-based service which would treat the premium-class arrival with priority, keeping the delay experienced by that arrival small. As a result, for the same performance target, *the REC is higher for smaller g*, which is especially clear from Fig. 3c. We believe this is important as we anticipate that current networks will likely see only a slowly increasing amount of premium class traffic. We would like to note that even for large g values (e.g., $g \approx 0.6$) under a well-behaved traffic like MMPP, the REC can be quite significant (e.g., 70%) at very relaxed performance targets (e.g., $t_{target} \approx 500$ packets) (see Fig. 3c).

3.2.3. Achieving a Loss Target: MMPP/M/1/K

We now look at the MMPP traffic case when the performance goal is in terms of average packet loss probability, i.e., p_{target}. Note that the CoS link provides an equal buffer size of K to each of the traffic classes, and the classless link uses all the buffer (i.e., total of $2K$) for the aggregate traffic. Figs. 6 and 7 show the percentage REC for a range of a and r combinations and buffer sizes. We used three buffer sizes $K = 10$ ms, 25 ms, 100 ms to be independent of the link speed. Further, to keep the delay sufficiently low for legacy application requirements, these buffer sizes are reasonable [35]. The darkness of the REC surface (scale shown in the vertical bar on the right) shows the target average loss probability (i.e., p_{target}) percentage. As the burstiness of the MMPP traffic increases, the REC also increases accordingly. Our simulations treated loss probabilities less than 10^{-5} (i.e., 0.001%) as zero (marked with an "Insufficient precision" label in Fig. 7) and the loss probability achieved by the premium class is less than 10^{-5} in these flat regions. Therefore, REC values for small g and ρ values will likely to be higher than what is shown in Fig. 7.

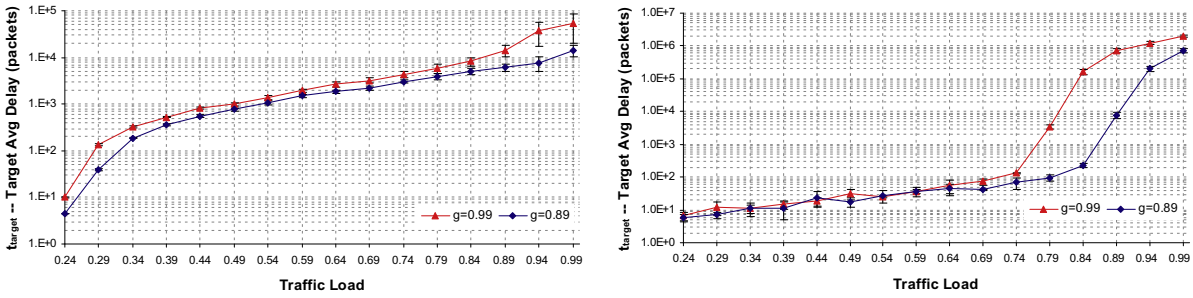
From Figs. 6 and 7, we observe once again that the REC grows with utilization, particularly when g is small. Also, as one would expect, as we increase the amount of buffering K from 10 ms to 100 ms, the REC reduces, and the range of utilization where there is little or no REC required also slowly increases. As the utilization increases, the loss probability goes up (increasing darkness). If the acceptable packet loss target is small, the REC also has to be higher.

3.2.4. Achieving a delay target: LRD/D/1

Fig. 8a shows the REC under LRD traffic when the performance target is the average delay, t_{target}. We use a Hurst parameter value of 0.75. The traffic is therefore much more bursty than the MMPP case. Notice that the vertical axis is logarithmic, and the REC is much larger than in the case where the arrival is MMPP. The higher burstiness in the traffic results in a much higher capacity requirement for the classless link to clear the backlog. Again, we observe for a fixed t_{target} that the REC grows as g decreases. An important observation to make is that the REC can be quite significant even when the proportion g is relatively modest. Even at low utilizations, with $g = 0.2$, the REC can be over 100%.

3.2.5. Achieving a loss target: LRD/D/1/K

Fig. 8b shows the REC for LRD traffic when the target is again the average packet loss probability, with the same assumptions about the buffer size, $K = 100$ ms, for the CoS and classless cases as before. Again, we use a Hurst



(a) MMPP with a=0.5, r=8

(b) LRD with H=0.75

Fig. 4. Simulation confidence for link models: 95% confidence intervals of the target delay performance for the two largest g values. Confidence intervals for other cases are smaller.

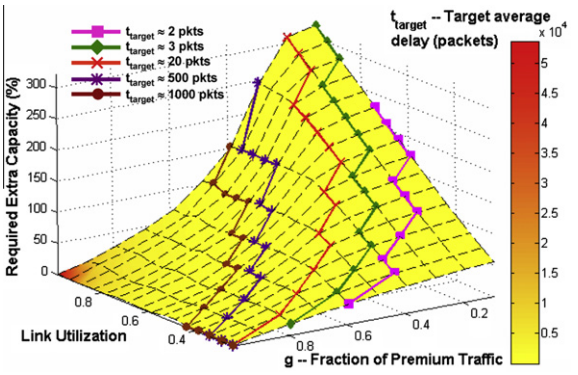


Fig. 5. DELAY – MMPP/M/1 link model with a = 0.5, r = 8: more bursty traffic increases REC.

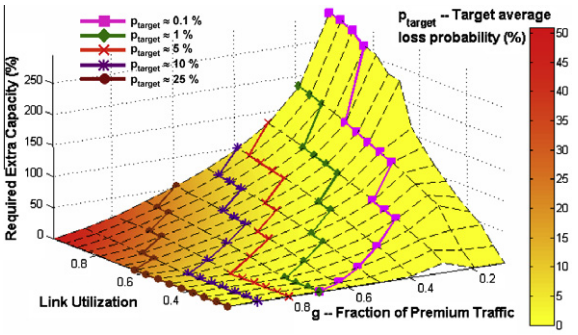
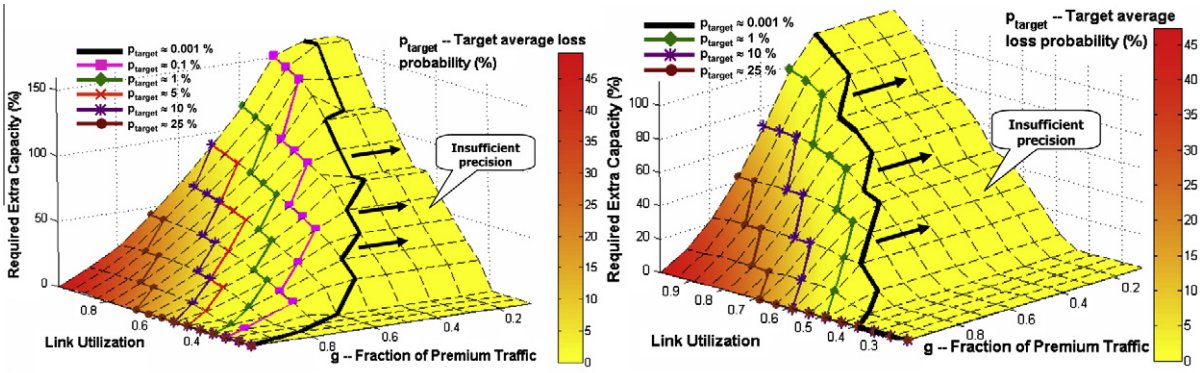


Fig. 6. LOSS – MMPP/M/1/K link model: the surface darkness shows the target loss probability. The buffer size is K = 10 ms, and MMPP parameters are a = 0.5 and r = 4. The colored lines on the surface roughly show the points where the target loss probability is of a certain value. For example, the blue line shows the points for 0.1% of average loss. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

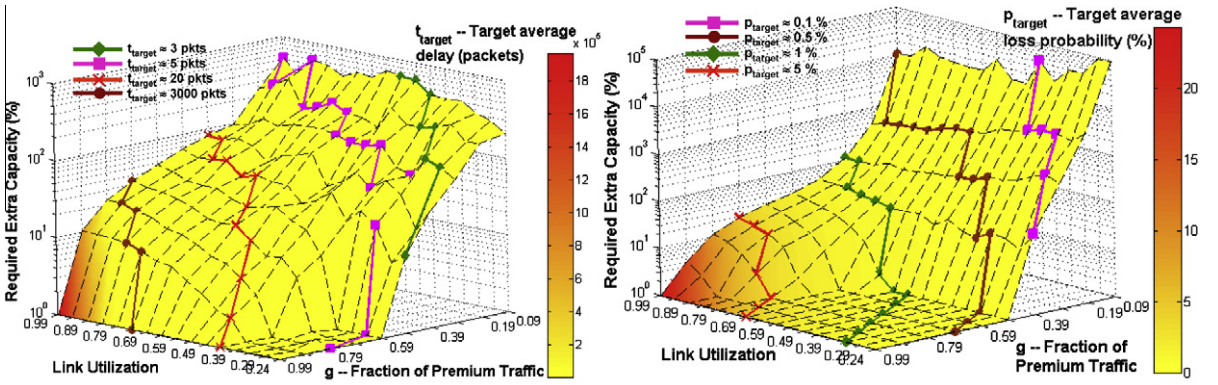
parameter value of 0.75. We observe that the REC is much higher now than the MMPP traffic case, confirming our intuition that REC increases with more burstiness in the traffic. It is also interesting that even with small g and low utilization ρ, REC can be quite high, which we could not observe in the MMPP/M/1/K model (Fig. 7) due to the insufficient precision in the simulations. In the case of LRD traffic, the simulation precision is enough to uncover this interesting behavior. The reason behind this behavior is that, at small g, the loss target is low (due to the small proportion of premium traffic, the CoS network achieves a very low loss probability for the premium traffic), and the classless link has to have a significantly higher capacity to match the stringent loss probability achieved by the



(a) $a = 0.5, r = 4, K = 25ms$

(b) $a = 0.5, r = 4, K = 100ms$

Fig. 7. LOSS – MMPP/M/1/K link models for various buffer sizes K = 25 ms and K = 100 ms, roughly corresponding to 1500 and 6000 packets (since the total buffer is 2K by our model) for a 1 Gb/s link carrying 1 KB packets.



(a) DELAY – LRD/D/1: Hurst=0.75 (b) LOSS – LRD/D/1/K: K=100ms, Hurst=0.75

Fig. 8. LRD-based link models: the surface darkness shows the performance target (delay or loss). The vertical axis is on logarithmic scale and the actual REC values are much larger than the ones under MMPP traffic.

premium class. As g gets larger, the loss target is higher, and correspondingly the REC is lower. It is only when the utilization gets close to 1 that the REC is high, even for a large g.

3.2.6. Mix of distributions for premium and best-effort traffic

Realistic traffic for the premium class is likely to be less bursty than the BE class traffic, since premium class typically serves multimedia applications with non-adaptive smooth traffic behavior. We consider the effect of this less bursty Premium traffic on our quantification of REC. Instead of applying the same burstiness to both Premium and BE class traffic, we used deterministic (i.e., constant-bit rate) traffic for the Premium class and LRD with Hurst parameter = 0.75 for the BE class in our simulations. Fig. 9 shows the REC under this scenario. Interestingly, the performance that Premium class achieves is a lot better (always under 2 packets of delay) in comparison to the case when premium class is assumed to have the same burstiness as the BE class, as shown in Fig. 8a. The performance of the premium class gets worse only because of the non-preemptive queueing, but that limits the degradation to about 0.5 packet delay. When the premium class is less bursty, REC is roughly the same (or slightly higher) when the fraction of premium class g is small. When g is larger, REC is significantly more in comparison to Fig. 8a. The observation is that *a less bursty premium class traffic will only make the REC estimates higher*. Thus, for the rest of the paper, in order to not over-estimate REC, we choose to make the assumption that premium class and BE class traffic have the same level of burstiness.

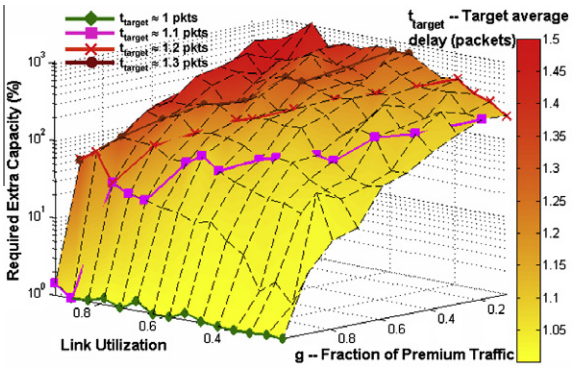


Fig. 9. DELAY – LRD/D/1 link model when the premium class traffic is deterministic. Hurst parameter is 0.75.

3.3. Protocol effects: open- vs. closed-loop traffic

TCP is the dominant transport protocol in the Internet. It adapts to the amount of capacity available for each individual transport level connection by increasing or decreasing the amount of load it puts into the network. It is important to examine the effect on REC in the context of such adaptive protocols. TCP is primarily sensitive to packet loss and its adaptive nature depends on the feedback received in terms of the packet loss probability p. The insights from [36] clearly showed that long-lived TCP flows performance is mainly driven by loss. Thus, we use packet loss probability experienced by long-lived TCP flows as the key metric to estimate REC under closed-loop traffic.

Since short-lived TCP flows typically operate with small window sizes and often do not exit the Slow Start phase, their performance is mainly dominated by timeouts in response to losses [37]. Such timeouts happen due to TCPs inability to employ Fast Recovery before it transmits a minimum number of segments [37]. This high sensitivity of short-lived TCP against losses has been tackled in various ways such as queuing disciplines favoring short-lived flows [38,39], changes to initial behavior of TCPs congestion control algorithm [40], and explicit congestion notifications (ECNs) from the network [41,42]. For REC estimates involving short-lived TCP flows, additional research beyond what we address below may be appropriate.

We examine the amount of REC by simulating a large number of TCP-SACK connections sharing a bottleneck link. First, we examine the case when both the best-effort and premium class traffic use TCP. If we used a simple priority queueing structure, the best-effort TCP traffic may be starved. As such, the comparison of the two types of

service (CoS vs. classless) would not be fair, and the REC arrived at will be excessive. To avoid this, we use a Deficit Round-Robin (DRR) queueing discipline [12] with equal weights to premium and BE classes. We used g, the fraction of premium traffic, to divide the TCP flows into premium and BE classes, e.g., the premium class has 200 out of a total of 1000 TCP flows when $g = 0.2$. The buffer size for each class was $K = 122$ packets (and $2K$ for the classless case), compared to a bandwidth-delay product of 9.8 packets. This works out to a buffer with a maximum queueing delay of 100 ms for each class. As an example scenario, the end-to-end RTT in our simulation setup was 8 ms. The bottleneck link capacity was 10 Mbps and the packet size was 1 KB. We purposefully set the buffer size to a large value (i.e., more than 10-fold difference) in comparison to the bandwidth-delay product, since we experiment with at least 100 TCP flows. We sought to operate in a range where the packet loss rates were small enough, and each TCP flow operates in congestion avoidance mode for long enough to achieve steady state.

Fig. 11a and b show the REC as the total number of TCP flows and fraction of premium flows vary. A critical observation to make here is the fact that achieved loss rates are much higher in comparison to the LRD traffic case in Fig. 8b. Thus, the REC values shown in Fig. 11a and b are significantly higher for the same loss performance target. Further, Fig. 11b clearly shows that a lower p_{target} results in a higher REC, and to retain p_{target} one can only support smaller number of TCP flows with the classless service.

Thus, the results of our TCP experiments clearly show that REC is higher for TCP traffic than for LRD traffic, with everything else staying the same. So, again, we choose to base our network-level REC estimates on MMPP and LRD traffic for the sake of staying conservative in our REC estimates.

4. Impact of CoS on best-effort

In this section, we examine the utility of even the simple non-preemptive priority based CoS service in retaining the performance of best-effort traffic relative to the performance achieved in the classless network. Fig. 10 shows the operating regions of interest based upon the two key parameters: the utilization, ρ, and the fraction of premium traffic, g. Regions B (moderate ρ, low-to-medium proportion of premium traffic g) and A (with a higher ρ) are of interest, where CoS may help achieve premium traffic SLAs. Note that at very low utilizations (regions C, F in Fig. 10), there is little impact on best effort traffic from the presence of CoS or premium traffic. Regions E and D are of less interest here, since they involve high values of g and high utilization, i.e., unlikely operating regimes in the near-to-medium term.

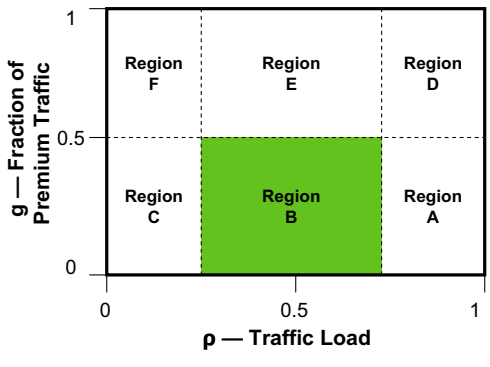


Fig. 10. CoS regions of interest.

The performance experienced by the best effort traffic in a classless (CL) service is approximated by the BE performance values when $g \approx 0$ (i.e., when there is little premium traffic). This is shown by the solid-line elliptical regions marked as "Region CL" in Fig. 12. With the classless service, premium class traffic would also suffer the same delay and losses as shown in Region CL for any level of the fraction g (and hence potentially not meet desired SLAs).

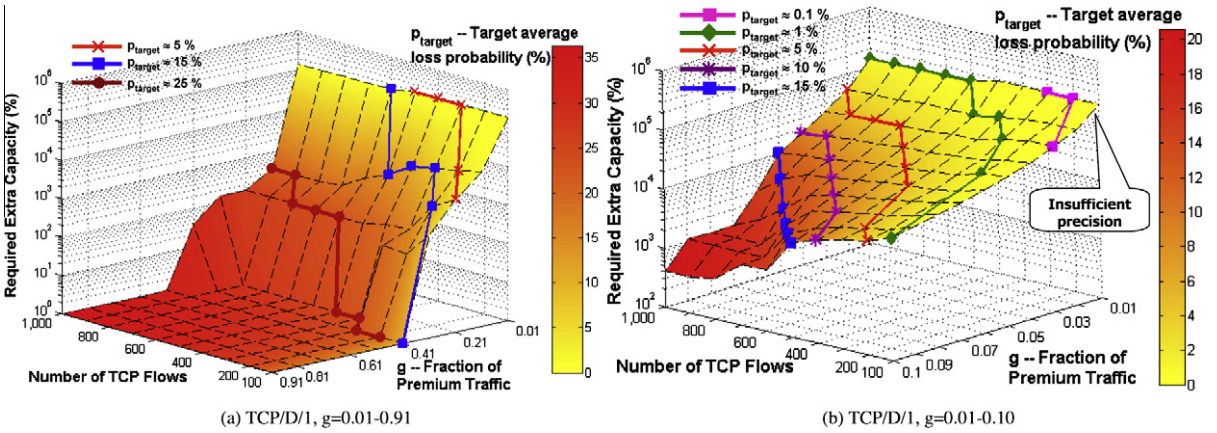
Fig. 12 shows loss and delay for best effort (BE) for the MMPP/M/1/K link model. Regions A, B and CL are overlaid on these graphs. In regions A and B of Fig. 12a and b, we see that BE performance with CoS (compared to the performance achieved with an equivalent classless scheme shown by region CL in the same figures) is *not noticeably degraded* both in loss and delay. More specifically, best effort traffic does not suffer even when the utilization is high (region A), for a premium traffic fraction (g) of up to 0.5–0.6, a parameter that can be managed in practice by network engineering. At the same time, in these scenarios (especially in region A) CoS serves an essential purpose of allowing premium traffic to meet performance requirements, despite burstiness in traffic and high utilization. Observe also that this range of g is the same operating regime we saw earlier that *requires greater excess capacity* in the equivalent over-provisioned classless network. We recognize that BE traffic is likely to be loss sensitive rather than being sensitive to increased delay.

As the fraction g of premium traffic increases close to 1, and the link utilization (ρ) is also high, the best effort traffic observes increased delay as shown in Fig. 12b. This is because the effective per-packet service time for BE in those regimes is higher due to CoS prioritization of premium traffic. However, loss for the BE traffic does not increase substantially in the CoS case compared to the classless case, even in these stressful regions as seen in Fig. 12a.

To summarize the link-level results, we observe that REC can be quite large. The REC was ~100% even at a reasonable average link utilization of 40%, for a relatively small proportion (e.g., $g < 0.2$) of premium class traffic with conservative assumptions on the burstiness of the traffic (MMPP). Traffic burstiness increases the REC. Under similar conditions, the REC under LRD is an order of magnitude higher, i.e., ~1000%.

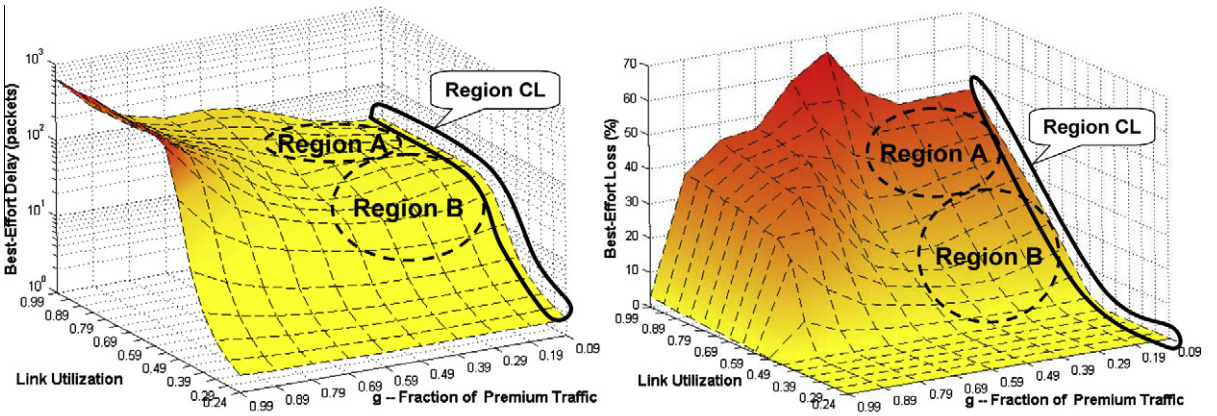
5. Network model

We extend our model framework by generalizing the single link model to a network model. We focus on developing our network model to reflect a typical ISP's backbone network. Crucial components of a network model include



(a) TCP/D/1, g=0.01-0.91 (b) TCP/D/1, g=0.01-0.10

Fig. 11. LOSS – protocol effects on REC: both premium and BE traffic are TCP in (a) and (b).



(a) BE loss (b) BE delay

Fig. 12. Best-Effort (BE) class performance for the MMPP/M/1/K link model when MMPP's burstiness is defined by $a = 0.5$ and $r = 4$.

(i) a *topology* (i.e., adjacency matrix, link weights, link propagation delays, link capacities) and (ii) a *traffic matrix*. Given the topology and the traffic information, we then to calculate REC for the complete network. We call this required extra capacity for the complete network as "*network REC*"(*NREC*).

We first calculate a routing matrix R for the ISP network from the link weight information. With a traffic matrix T, we then calculate the traffic load pertaining to individual links by taking the product of T and R. For each of these link traffic loads, we apply the link model described earlier and find the RECs for each individual link. Finally, we calculate the network REC (NREC) by averaging the individual link RECs across all links of the network. For NREC, there are two possible ways of averaging the link RECs: (i) $NREC_A$, to calculate the ratio of the sum of the RECs of *all* links of the entire network, and (ii) $NREC_I$, to calculate the average of the ratio of the REC of each *individual* link. Mathematically speaking, for a network with set of links L, if REC_l and c_l represent the REC and the capacity of link l respectively, then $NREC_A$ and $NREC_I$ are calculated as:

$$NREC_A = \frac{\sum_{l\in L} c_l REC_l}{\sum_{l\in L} c_l}$$

$$NREC_I = \frac{\sum_{l\in L} REC_l}{|L|}$$

In brief, $NREC_A$ expresses the *total extra capacity needed* to make the network as a whole meet the g2g performance goals and $NREC_I$ expresses the *average extra capacity needed on each link* of the network to meet the g2g performance goals. For example, $NREC_I$ would likely reflect better the situation when a link with a small capacity requires a significant additional capacity (e.g., if it is the bottleneck).

Notice that NREC calculation is not just calculating expectations over distributions of the link-level RECs. For example, although the network topologies are available through public sources, the traffic matrix and link capacities are generally not publicly available, and end-to-end capacity estimation of individual links is not yet reliable. Further, after estimating the product of the T and R matrices, some links will be infeasible as their estimated capacities may be lower than the estimated traffic load on them.

5.1. Methodology

The goal of the network model is to determine the *percentage additional capacity needed for a classless network over a CoS network* on an edge-to-edge (g2g) basis. For a network with N nodes, L links, and $F = N(N-1)$ flows, let $T_{N\times N}$ be the traffic matrix. If there exists a positive flow from ith node to jth node, then $T_{i\times j}$ is the traffic rate in Mb/s from ith node to jth node. If not, then $T_{i\times j}$ is 0. Let $\lambda_{F\times 1}$ be the traffic vector, which is the vectorized version of $T_{N\times N}$ such that $\lambda_{(i-1)N+j} = T_{i\times j}$ where $i, j = 1,\ldots,N$ and $F = 1,\ldots,N(N-1)$. Let $R_{F\times L}$ be the routing matrix, where $R_{i\times j}$ is 1 if the ith flow traverses the jth link. If not, $R_{i\times j}$ is 0. Also, let $A_{N\times N}$ be the adjacency matrix, $W_{N\times N}$ be the link weight matrix, $S_{N\times N}$ be the link propagation delay matrix, and $C_{N\times N}$ be the link capacity matrix. The network model requires the following inputs:

- *The traffic matrix:* $T_{N\times N}$.
- *Topology information:* Adjacency matrix $A_{N\times N}$, link weight matrix $W_{N\times N}$, link propagation delay matrix $S_{N\times N}$, link capacity matrix $C_{N\times N}$.
- *The link model:* The link model formulates the REC ($\mu_N - \mu_D$) for a given traffic load (i.e., $\lambda_{F\times 1}$) and performance goal.
- *Premium class performance:* t_{target} or p_{target}, the performance goal to be achieved.

The network model takes the following steps to calculate the network REC (NREC):

- *Step 1:* Construct the *routing matrix* $R_{F\times L}$ based on shortest path first (Djiskstra's) algorithm using the topology information $A_{N\times N}$ and $W_{N\times N}$.
- *Step 2:* Form the *traffic vector* $\lambda_{F\times 1}$ from $T_{N\times N}$.
- *Step 3:* Calculate the *traffic load on each link* by performing the matrix operation $Q = R^T\lambda$, where $Q_{L\times 1}$ is the link load vector (in Mb/s).
- *Step 4:* Check the *feasibility* of the traffic load and routing. If any link's capacity is less than the load onto that link, then we fix the infeasibility by increasing the capacity of that link.
- *Step 5:* Calculate the *per-link* REC by using Q_i as the total traffic rate for the ith link and the performance goal t_{target} or p_{target} for that link i.
- *Step 6:* Calculate the *network* REC (NREC) by averaging the per-link RECs from Step 5.

5.2. Topology

To obtain some of the topology information, we used the Rocketfuel [7] data repository which provides router-level topology data for six ISPs: Abovenet, Ebone, Exodus, Sprintlink, Telstra, and Tiscali. Specifically, it provides A, W, and S for the six ISPs, but an estimation of C is not provided. Table 1 shows a summary of the topology information for the six Rocketfuel topologies. We updated the original Rocketfuel topologies such that all nodes within a PoP (assuming that a city is a PoP) are connected with each other by adding links to construct at least a ring among routers in the same PoP.

5.3. BFS-based link capacity model

In order to assign estimated capacity values for individual links of the Rocketfuel's topologies, we use a technique based on the Breadth-First Search (BFS) algorithm. We, first, select the maximum-degree router in the topology as the center node for BFS to start from. After running BFS from the max-degree router, each router is assigned a *BFS distance* value with respect to the center node. The center node's distance value is 0.

Given these BFS distances, we apply a very simple strategy to assign link capacities: Let the BFS distances for routers i and j be d_i and d_j respectively. For the links (i,j) and (j,i) between the routers i and j, the estimated capacity $C_{i,j} = C_{j,i} = \kappa[\max(d_i,d_j)]$ where κ is a decreasing vector of conventional link capacities. In this paper, we used: $\kappa[1]$ = 40 Gb/s, $\kappa[2]$ = 10 Gb/s, $\kappa[3]$ = 2.5 Gb/s, $\kappa[4]$ = 620 Mb/s, $\kappa[5]$ = 155 Mb/s, $\kappa[6]$ = 45 Mb/s, and $\kappa[7]$ = 10 Mb/s. So, for example, a link between the center router and a router with BFS distance 5 will be assigned 155 Mb/s as its estimated link capacity. Similarly, a link between routers with distances 1 and 3 will be assigned with a capacity of 2.5 Gb/s.

The intuition behind this BFS-based method is that an ISPs network would have higher capacity links towards the center of its topology. It has been generally found that the routers in the center of an ISP topology are more likely to have a higher degree. This general intuition may be supported by the study in [43] showing that router technology has been producing higher degree-capacity combinations for the core routers internal to the backbone, which is particularly meaningful for our study since we consider the backbone of an ISP rather than the customer-facing interfaces at the edge of ISP topology. However, when

Table 1
Rocketfuel-based router-level ISP topologies.

ISP	# of Routers	# of Links	Degree (avg/max)	BFS distance (avg/max)	Min degree	Min BFS distance	# of Edge routers	# of g2g flows
Abovenet	141	922	6.6/20	2.3/4	9	3	108	11,556
Ebone	87	404	4.7/11	3.3/7	6	4	66	4290
Exodus	79	352	4.5/12	3.0/5	6	4	60	3540
Sprintlink	315	2334	7.4/45	2.7/7	9	5	254	64,262
Telstra	108	370	3.8/19	3.5/6	5	4	84	6972
Tiscali	161	876	5.6/31	2.6/5	8	4	125	15,500

considering the large number of peering relationships that edge routers have, the evidence for or against this intuition is limited at this point. Notice that these link capacity estimates are *initial* values, and will possibly change when the model attempts to fix unfeasibility due to mismatch between these estimated link capacities and link traffic loads, caused by the combination of the traffic load and the shortest path routing.

5.4. Traffic model

A crucial piece in modeling a network is the workload model, in particular the traffic matrix. Each flow in the network model must reflect the traffic from one edge router to another edge router. Thus, there are two important steps in constructing a reasonable traffic matrix. First, we identify the edge routers as a subset of routers in the Rocketfuel topologies, which is detailed in Appendix B. Once edge routers are identified in a topology, we use the gravity model [44,45] to construct a feasible traffic matrix composed of edge-to-edge (g2g) flows, and the specifics of our method to establish the gravity-based traffic matrix is detailed in Appendix C.

5.5. Edge-to-edge (g2g) performance goal

To evolve from the link model of Section 3 to the network model, we split the g2g performance goals for the individual links of the g2g path. Determination of g2g performance goals is driven by real-time applications such as VoIP. Specifically, the maximum one-way delay, acceptable for most interactive voice usage is about 150 ms. Based on the analysis in [46], the delay budget for queueing in the backbone network is approximately 10 ms after taking into account propagation, coder, silence suppression, de-jitter buffer and access network delays. Typically, the requirements for packet loss for encoded speech are 1% or less. Thus, we use these ranges to set the g2g performance goals in our network model analysis. As an example, a toll-quality IP Telephony service typically imposes performance requirements. These include: (i) Low end-to-end packet delay so that it does not interfere with normal voice conversations, and (ii) Low packet loss: so as to not perceptibly impact either voice quality or the performance of other equipments that use it as the underlying communication medium, e.g. legacy fax.

While loss concealment algorithms can be used to reproduce intelligible speech even with higher loss rates, the resulting performance may often be considered to be inadequate. However, in addition to such QoS needs under typical conditions, premium application traffic also expect to have their services protected under transient failure conditions as well. While the above may be a worst-case situation, it is important to note that interactive real-time applications impose non-trivial constraints of loss and delay on the network.

5.5.1. Apportioning delay

In order to split g2g delay target t_{target} on individual links, we simply divide the delay requirement equally on each link of the path assuming that t_{target} is only the delay in queueing and insertion into the links.[3] After splitting the delay on individual links for all g2g flows, we collect the tightest (i.e., minimum) delay requirement on each individual link among the delay requirements imposed by each g2g flow traversing the link.

Let $\mathbf{F}$ be the set of all g2g flows and $\mathbf{L}$ be the set of all links. Given the g2g delay requirement t_{target}, we calculate the delay requirement of flow f on the link i as:

$$t_{f,i} = \frac{t_{target}}{l(f)}, \quad f \in \mathbf{F}, \quad i \in \mathbf{L} \tag{3}$$

where $l(f)$ is the number of links the flow f traverses. Then, for each link $k \in \mathbf{L}$, the delay requirement is:

$$\hat{t}_k = \min_{f \in \mathbf{F}} t_{f,k}.$$

When $\hat{t}_k$ turns out to be unrealistically small, we set $\hat{t}_k$ to be the minimum possible, which is the packet service time (based on the packet size and the link's capacity.)

Though this method of calculating individual links' delay requirements from the g2g delay requirement t_{target} is plausible, better methods are possible. For example, after determining that $\hat{t}_k$ is tighter than a flow f's delay requirement $t_{f,k}$ on the link k, it is possible to distribute the extra delay margin $t_{f,k} - \hat{t}_k$ for flow f to the links on f's path other than the link k. This would relax the delay requirement on these other links and possibly affect the NREC results. Although one can sketch such seemingly better ways of distributing g2g delay requirements on individual links, its affect on the end result will not be significant. Thus, in this paper, we used this simple method of equally apportioning the g2g delay requirement on individual links.

5.5.2. Apportioning loss

We apply a similar procedure to apportion the g2g loss probability target p_{target} on individual links. Specifically, for a flow f traversing $l(f)$ links, we assign the survival probability to each link as the $l(f)$ th root of the overall path's survival probability $1 - p_{target}$. Thus, the loss probability requirement of flow f on the link i is:

$$p_{f,i} = 1 - \sqrt[l(f)]{1 - p_{target}}, \quad f \in \mathbf{F}, \quad i \in \mathbf{L}. \tag{4}$$

We then collect the tightest loss probability requirement on each individual link among the loss probability requirements imposed by each g2g flow traversing the link. That is, for each link $k \in \mathbf{L}$, the loss probability requirement is:

$$\hat{p}_k = \min_{f \in \mathbf{F}} p_{f,k}.$$

5.6. Network model results

We now present the quantitative results on the *network REC* (NREC) for the various Rocketfuel topologies. In order to generate the NREC results, our network model uses link

[3] The actual g2g delay will also include the g2g path's propagation delay (which is available from the Rocketfuel data).

model results, presented in Section 3, for a given utilization and performance target. We perform a lookup of the link model simulation result and use linear interpolation on the link model surfaces (e.g., Fig. 3a) with the available datapoints. Sometimes, the link model surface might not have the datapoint corresponding to the performance target and utilization combination. For example, the LRD/D/1/K model in Fig. 8b does not have the datapoints for $p_{target} = 0.01\%$ when the link utilization ρ is greater than 0.5. In such cases, we *conservatively* assume that the link's REC is equivalent to the closest point on the surface, even though the real REC value would be higher.

We calculate NREC both for MMPP traffic and LRD traffic. The MMPP traffic allows us to observe NREC values under smooth, well-behaved, and short-range-dependent traffic with very conservative burstiness behavior. The LRD traffic allows us to see how much larger REC would become under more bursty traffic. In both cases, we use *conservative* parameters for the traffic burstiness, i.e., $a = 0.5$, $r = 4$ for MMPP, and Hurst = 0.75 for LRD. The IP traffic is often more bursty than this [9]. Also, when loss probability is the performance goal, we use a buffer size of $K = 100$ ms, which is conservative in comparison to conventional buffer sizes on IP backbone links. Also, if we are interested in keeping the g2g delay small, this is often achieved by keeping the buffer size small [35].

Figs. 13 and 14 show the NREC values for the two ISP topologies Exodus and Sprintlink, where the top row shows the results for LRD traffic and the bottom shows MMPP traffic. We do not include the results for the other topologies as they are similar in behavior and also the actual values. The graphs in the figure show $NREC_I$ with solid lines and $NREC_A$ with dashed lines. We show NRECs for five levels of traffic load, which translates to different average network utilizations for each ISP topology. For example, the maximum traffic load we could carry with the Exodus topology resulted in 80% avg. utilization while the Sprintlink topology had 68% avg. utilization.

Fig. 13 shows NRECs when the performance target is the g2g queueing delay. It is clear that both $NREC_I$ and $NREC_A$ increase as the average link utilization increases, especially when the target average g2g queueing delay is smaller. Also, as expected, LRD traffic results in an order of magnitude larger NRECs in comparison to the case with MMPP traffic. For example, for a g2g queueing delay target of 5 ms and a 40% utilized Sprintlink network, NREC under MMPP traffic is about 20% while it would be about 100% with LRD traffic. This difference becomes more evident when the target g2g queueing delay is smaller.

Fig. 14 shows NRECs when the performance target is the g2g loss probability. Again, it is clear that both $NREC_I$ and $NREC_A$ increase as the average link utilization increases. For the LRD traffic, there is a flat region for the NREC values when the target g2g loss probability is below 0.1%. This is mainly due to the fact that our LRD/D/1/K link model cannot capture very low loss probability targets with high enough precision. If there was enough precision, we would have observed that the NREC behavior consistently drops with increasing target g2g loss probabilities, like in the MMPP traffic case.

$NREC_A$ and $NREC_I$ are closer to each other for Sprintlink than for the Exodus topology. This can be explained by the fact that Sprintlink topology is more "meshed" with more evenly distributed load across its links, yielding a situation with fewer bottlenecks. Intuitively, in such a case, the required increase in capacity is focused on those few links, which results in an imbalance between the two NREC measures.

When the performance target is delay, $NREC_A$ becomes larger than $NREC_I$. The reason behind this is the apportionment of the g2g performance target on individual links. As can be seen from Eqs. (3) and (4), for a given overall perfor-

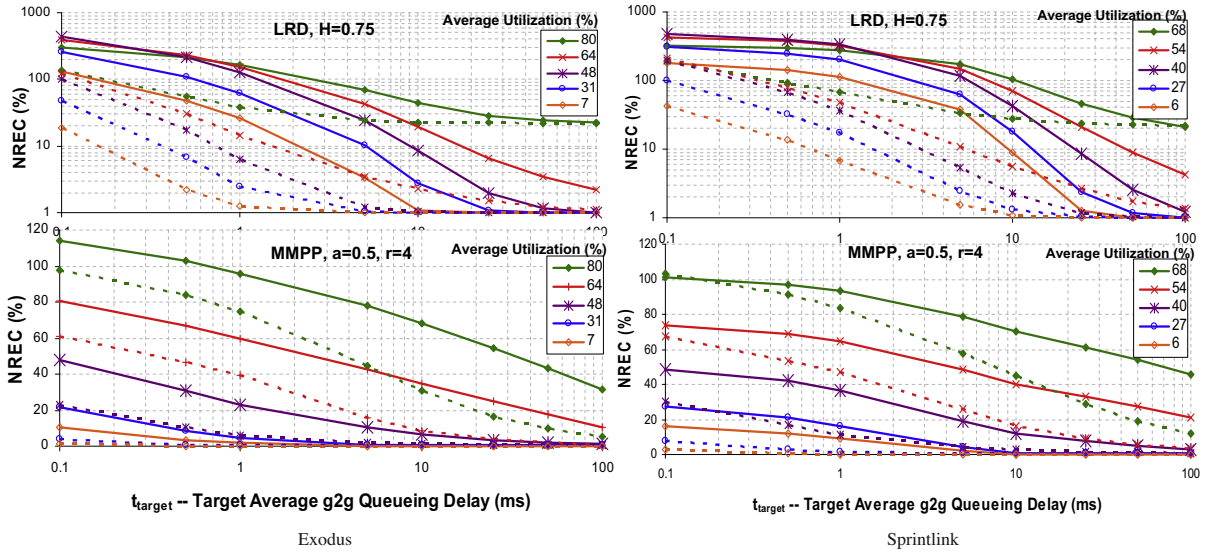


Fig. 13. G2G DELAY – NREC against the target average g2g queueing delay under two kinds of traffic: MMPP vs. LRD. The solid lines show $NREC_I$ while dashed lines show $NREC_A$.

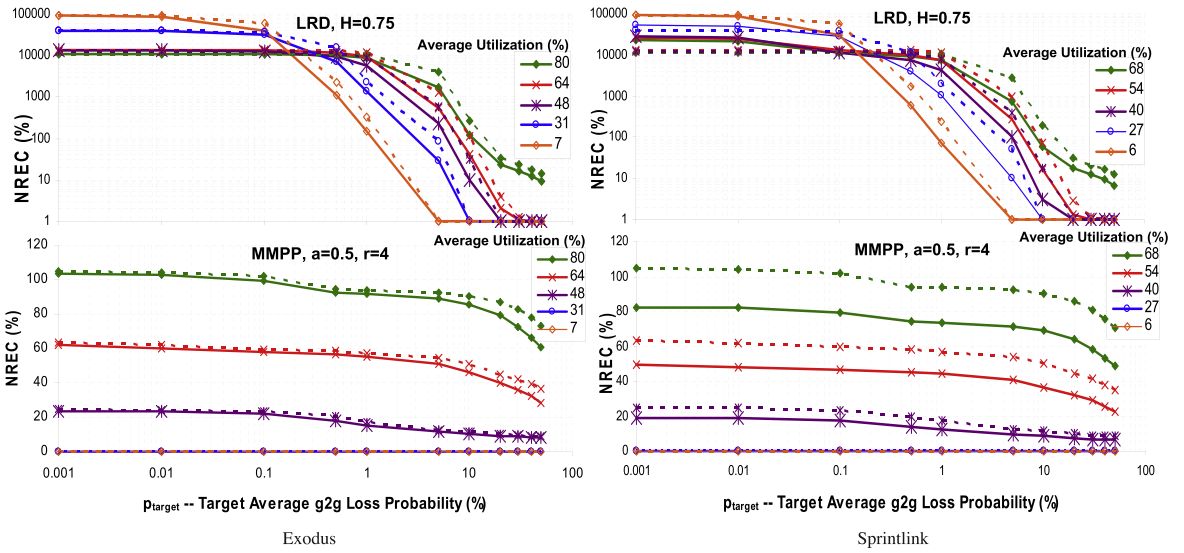


Fig. 14. G2G LOSS – NREC against the average g2g loss probability under two kinds of traffic: MMPP vs. LRD. The buffer size at each link of the network is $K = 100$ ms. The solid lines show $NREC_I$ while dashed lines show $NREC_A$.

mance target, the target queueing delay budget for an individual link *reduces* as the number of links l on the g2g path increases. That is, the longer the g2g path, the more stringent the delay requirement on one of its links. Intuitively, most of the g2g paths in the network topologies are short, while only a few are long distance paths. Thus, the delay apportionment naturally causes the links at the center of the topology to have a tighter delay requirement.

6. NREC under failures

A crucial factor in an ISP's network design is the tolerance to failures. We analyzed the behavior of NREC under failures with the Rocketfuel ISP topologies. To see the effects of link failures, we recalculated the routing matrix of the ISP's new topology after a link failure, and then recalculated the loads on individual links after rerouting on alternate shortest paths. Based on the new link traffic loads, we recalculated the NREC. For node failures, we assumed that a node generating traffic (i.e., edge node) does not fail, to assure that all the traffic before failure still gets carried by the network after failure. This is appropriate to make an fair comparison between the failure-free and the failure cases of the same topology. Further, we focus only on the g2g loss probability as the performance target when considering the failure cases. To stay conservative in our NREC estimates, we used the MMPP/M/1/K link model with $K = 100$ ms, $a = 0.5$, and $r = 4$ (see Fig. 7b). Using the LRD/D/1/K link model would certainly result in higher NREC under failures.

One aspect of NREC analysis under failures is the need to tackle link 'utilizations' larger than 100%, since some of the rerouted flows cause some links to get overloaded. Our intent here is to estimate what would be the extra capacity required to carry the increased offered load, hence the need to estimate REC at higher 'utilizations'. We used a conservative linear extrapolation of the link model results in order to perform the lookup for each link REC to calculate the network REC after failures. To gain confidence in the conservativeness of our extrapolation, we simulated the MMPP/M/1/K link model up to 399% utilization which explicitly showed a linear (sometimes super-linear) increase in REC.

As shown in Fig. 15, we have observed that the increase in $NREC_A$ and $NREC_I$ can be quite different when a link failure occurs. For example, for 66% utilized Exodus topology with a 0.1% g2g loss probability performance target, the increase in $NREC_A$ due to a link failure is 0.4% on average and 5.1% at maximum, while the increase in $NREC_I$ is 58.5% on average and 1328.5% at maximum for the same scenario. This means that the aggregate capacity of the entire network needs to be increased up to an additional 5.1% *over-and-above the NREC pertaining to the topology without a failure* (examples of which are shown in Fig. 14). Although this increase amount is small in percentage, it corresponds to a considerable increase in the total network capacity as this is the needed additional aggregate capacity due to a single link failure. Also, high values of the needed increase, $NREC_I$, show that the impact on specific links can be quite significant and that the larger REC is required across several links. The increase in NREC due to node failures showed a very similar behavior (i.e., low $NREC_A$ and high $NREC_I$) with slightly larger values.

Another observation from our failure analysis is that $NREC_A$ and $NREC_I$ may differ significantly across topologies. For instance, for 66% utilization with a 0.1% g2g loss probability performance target, the maximum increase in $NREC_A$ due to a link failure is 7.1% and 3.6% for Abovenet and Tiscali respectively. The maximum increase in $NREC_I$ due to a link failure under the same scenario is 742.2% and 403.9% for Abovenet and Tiscali. This differing NREC behavior across ISP topologies is mainly driven by the

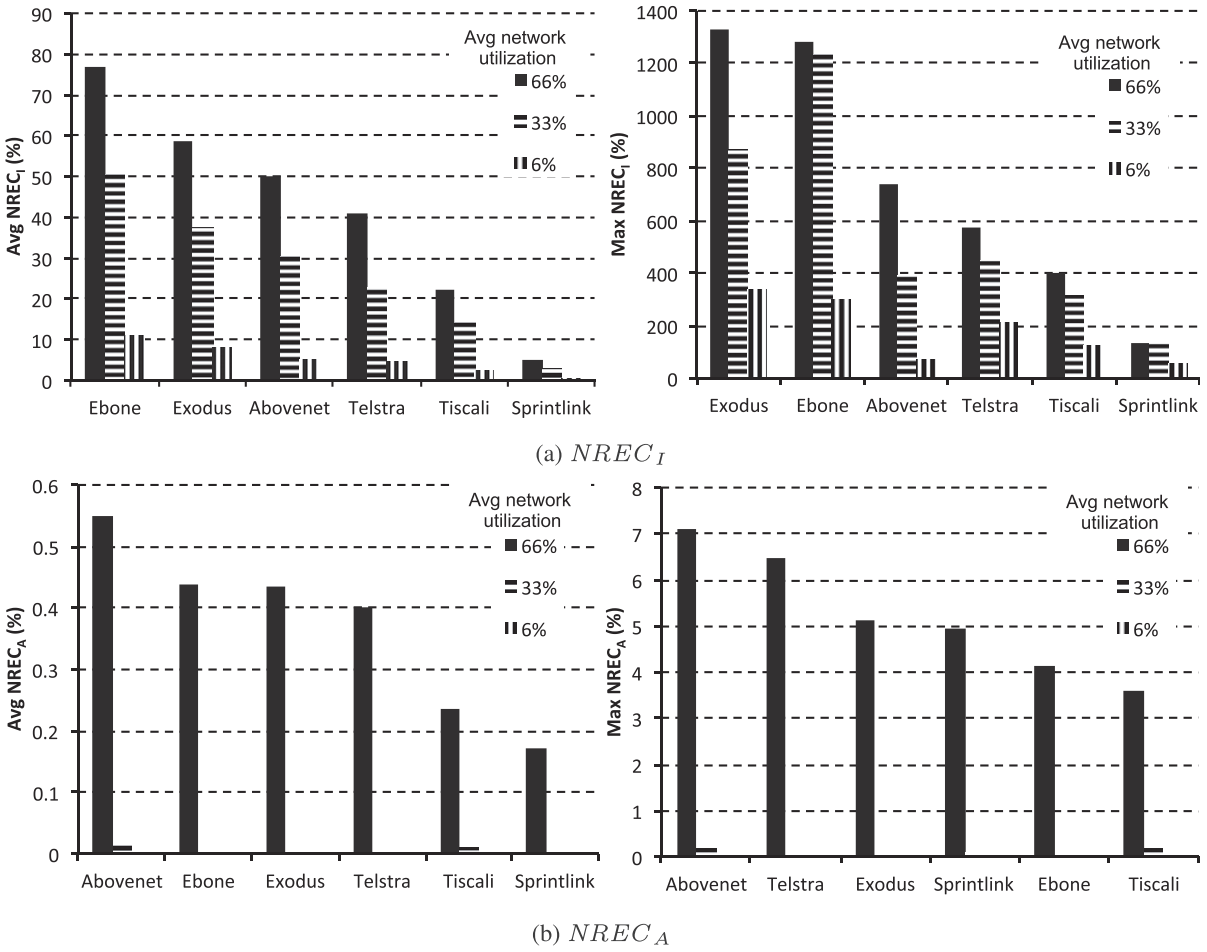


Fig. 15. NREC after a link failure: values are over-and-above the NREC for the topology without failure. The top and the bottom rows show the maximum and the average increase in NREC respectively. The traffic is MMPP with $a = 0.5$ and $r = 4$, and the buffer size at each link of the network is $K = 100$ ms. The solid and dashed lines show NRECs when the target average g2g loss probability is 0.1% and 10% respectively. Note that the dashed and the solid lines mostly overlap for $NREC_I$.

characteristics of the topologies, e.g., Exodus and Ebone are more hub-and-spoke in comparison to the others.

7. Summary and discussion

There has been considerable work and debate on the benefits of having a simple classless (i.e., best-effort) service. However, the quantification of the amount of extra capacity required of such a classless network to support traffic that requires delay and loss service-level assurances has not been explored in the past. In this paper, we have quantified the required extra capacity (REC) for a classless network to meet the same delay and loss assurances that would be provided by a relatively simple two-class network.

We first built an analytical framework to understand the nature of REC for a link. We demonstrated the nature of the ratio in simple analytic terms using an M/M/1 model. We then used a simulation-based calculation of link-level REC under a more bursty but *conservative* 2-state MMPP traffic arrival process, which exhibits a short-range dependent traffic pattern. We observed that REC grows with utilization, and is of particular concern when the proportion of premium class traffic requiring delay or loss assurances is small. To see the REC behavior under even more bursty traffic with long-range dependence (LRD), we used the same simulation-based REC calculation. We showed that the REC under LRD traffic is an order of magnitude higher than the REC estimates under the 2-state MMPP traffic. We also investigated REC behavior under closed-loop traffic composed of many TCP sources and showed that REC is even higher than what was observed under LRD traffic.

To show the behavior of REC network models for an IP backbone, we outlined a method of quantifying the network REC (NREC) from the network's link-level RECs and edge-to-edge performance targets. We observed that NREC increases with the average utilization of the network and as the relative proportion g of the premium traffic reduces. Moreover, NREC grows rapidly as the acceptable delay and

packet loss targets become tighter (smaller). So, for example, with conservative assumptions on the burstiness of the traffic (2-state MMPP parameters), NREC approaches 60% even at reasonable average link utilizations of 60%, for a relatively small proportion (e.g., 20%) of premium class traffic. NREC becomes even higher when we use more bursty traffic, like LRD. To understand the effect of failures on NREC, we recalculated the NREC of an ISP's topology after link failures. We found that, in order for the classless service to still satisfy the edge-to-edge performance targets after a link failure, the network's *aggregate* capacity (i.e., sum of all link capacities of the network) might have to be increased by an additional 7% in the worst case and 0.5% in the average case, *over-and-above* the NREC for the failure-free topology. We also found that each link's *individual* capacity would have to be increased by an additional 1,300% in the worst case and 70% in the average case.

In the future, as link speeds go up, the concern about queueing delay may be less important, because the dominant factor will be propagation delay. However, loss will still be a concern, especially when these links (in the core network) see a large aggregate number of flows, only some of which will be "adaptive", like TCP. One way to reduce loss is to increase buffering, but this increases delay [35]. Differentiating the delay-sensitive premium class traffic allows us to both control delay and loss for the premium traffic, without too much additional buffering. It also offers the flexibility for the provider to provision additional buffering for the BE traffic, to keep its loss probability within desired limits. On the other hand, in a classless network, the only option would be to increase the capacity enormously (as demonstrated by our results), in order to ensure that the loss is maintained at small levels while still not increasing queueing delay.

In our study, we have not collected distributional metrics (e.g., jitter or loss variation) during our simulation experiments and worked with averages while estimating REC. This is because we believe that average delay or loss provides a more conservative (i.e., smaller) estimate of REC in comparison to jitter or loss variation. We believe that REC estimates would be higher for distributional metrics particularly when traffic patterns with higher variability are considered. It would be useful to verify this assumption by extending our efforts to distributional metrics other than averages. Consideration of such distributional metrics along with fine-grained traffic modeling employing flow-level dynamics will shed more light into the trends for REC. Enriching the REC estimation effort with realistic models of flow arrivals and departures may give rise to a better understanding of REC when the impact on individual flows in a backbone are of concern. Though such flow-level QoS provisioning issues were heavily studied in the literature, efforts pertaining to REC estimations at that level are lacking.

Finally, *in terms of the capacity* needed to satisfy legacy and future network applications in the Internet, our results show that CoS in IP backbones is an order of magnitude better than the classless (i.e., over-provisioning) approach. However, further research is necessary to estimate the monetary costs of the two approaches, as scheduling and management costs need to be considered. Such economic considerations will also be necessary to fully quantify REC, or more generally "Required Extra Costs", for flows or a set of flows on an end-to-end (e2e) basis. Models involving inter-ISP economics and technological incompatibilities will be needed to understand monetary costs for attaining end-to-end performance measures. Though our work certainly relates to the e2e performance provisioning problem, we focus on what is at stake when only one part of the problem is considered, i.e., a single ISPs required investment into its network.

Acknowledgments

We sincerely thank Dr. Biplab Sikdar and the anonymous reviewers for their valuable feedback as they greatly helped us improve the paper.

This work is supported in part by the U.S. National Science Foundation awards 0721600 and 0721609.

Appendix A. Analytical derivation of REC under Poisson traffic

In this section, we derive the link-level REC in analytical terms for the two different performance targets: average delay and average loss probability.

A.1. Achieving a delay target: M/M/1 model

The first scenario we derive is the case when traffic is assumed to be Poisson and the performance target is queueing delay, i.e., t_{target}. Let μ_N be the required capacity for the classless link to be able to match the premium class performance with CoS. Since we are assuming that classless traffic is Poisson with a rate of λ_D, the delay achieved by the classless service for the aggregate traffic will be [47]:

$$t_{achieved} = \frac{1}{\mu_N - \lambda_D} \tag{A.1}$$

When the achieved delay $t_{achieved}$ is equal to the target delay (i.e., $t_{target} = t_{achieved}$), the classless link capacity μ_N equals the *minimum* required to satisfy the performance goal:

$$\mu_N = \frac{1}{t_{target}} + \lambda_D \tag{A.2}$$

This formulation for μ_N holds as long as the classless service performance is equal to or better than the premium class performance, i.e., $t_{achieved} \leqslant t_{target}$. For the purposes of our comparative model, we will assume $t_{achieved} = t_{target}$ holds.

Eq. (A.2) shows that the REC depends on the rigor of the performance goal $1/t_{target}$ and the aggregate traffic rate λ_D of the CoS link. However, not all values of t_{target} might be achievable for the premium class traffic at the CoS link. The average delay that premium class experiences at the CoS link is dependent on three factors: (i) the aggregate traffic rate λ_D, (ii) the fraction g of the premium class traffic in that aggregate, and (iii) the CoS link capacity μ_D.

By using non-preemptive priority queuing, we can formulate the delay for the premium class achieved by the CoS link as [48]:

$$t_{Prem} = \frac{1}{\mu_D}\frac{1+(1-g)\rho}{1-g\rho} \tag{A.3}$$

where $\rho = \lambda_D/\mu_D$ is the aggregate traffic load at the CoS link. By dividing (A.3) with the average packet service time (i.e., $1/\mu_D$) we get the average backlog in front of an arriving packet:

$$t_{Prem} = \frac{1+(1-g)\rho}{1-g\rho} \tag{A.4}$$

which expresses the delay in terms of "packets". This notion of delay is helpful especially for deriving conclusions on REC independent of the CoS link capacity μ_D and the average packet size. Thus, once we set values for g, and ρ, we also set the delay target, which is the delay achieved by the premium class. By setting $t_{target} = t_{Prem}$, we obtain the required classless capacity μ_N in terms of g,ρ, and μ_D:

$$\mu_N = \frac{\mu_D(1-g\rho)}{1+(1-g)\rho} + \lambda_D \tag{A.5}$$

from which REC in percentage can be written as:

$$REC_{delay} = 100\left[\frac{(1-g\rho)}{1+(1-g)\rho} + \rho - 1\right] \tag{A.6}$$

A.2. Achieving a loss target: M/M/1/K model

We now look at the case when the performance target is determined in terms of average packet loss probability, i.e., p_{target}. We establish our comparative model with one additional parameter, the buffer size K. We assume that CoS link provides an equal buffer of K packets to both traffic classes, and that the classless link uses both the buffers (i.e., total of $2K$ packets) for the aggregate traffic. The average loss probability achieved by the classless service for the aggregate traffic can be approximated by the tail probability of the queue (i.e., the probability that the queue occupancy would be larger than the buffer size of $2K$ packets) [47]:

$$p_{achieved} = \left(\frac{\lambda_D}{\mu_N}\right)^{2K} \tag{A.7}$$

This approximation has negligible deviation from the exact average loss probability formulation for M/M/1/2K when the K value is sufficiently large, e.g., $K > 10$. Also, the error in estimating the loss probability by means of this tail probability is conservative in terms of REC, as the tail probability is larger than the real loss probability for small K values.

When the achieved loss probability $p_{achieved}$ is equal to the target performance p_{target} (i.e., $p_{target} = p_{achieved}$), the classless link capacity is the *minimum* required to satisfy the performance goal:

$$\mu_N = \lambda_D \frac{1}{\sqrt[2K]{p_{target}}} \tag{A.8}$$

Similar to the previous case, Eq. (A.8) shows that the REC depends on the rigor of the performance goal $1/p_{target}$ and the aggregate traffic rate λ_D of the CoS link. It now depends on the available buffer size of the classless link, i.e., $2K$.

Not all p_{target} values might be achievable for the premium class traffic at the CoS link. The average loss probability that the premium class experiences at the CoS link is dependent on four factors: (i) the aggregate traffic rate λ_D, (ii) the fraction g of the premium class in the aggregate traffic, (iii) the CoS link capacity μ_D, and (iv) the available buffer size K. With non-preemptive priority queuing we assume for typical link speeds that the waiting time for an arriving higher priority packet for the completion of service of the current lower priority packet being transmitted will be negligible. Thus, we can safely approximate the loss probability for the premium class achieved by the CoS link by the M/M/1/K formula [47]:

$$p_{Prem} = \frac{1-g\rho}{1-(g\rho)^{K+1}}(g\rho)^K \tag{A.9}$$

where $\rho = \lambda_D/\mu_D$ is the aggregate traffic load at the CoS link. Once again, as with the delay case, the loss target is determined by the choice of g, ρ and K. By setting $p_{target} = p_{Prem}$, we obtain the required classless capacity μ_N in terms of g, ρ, μ_D, and K:

$$\mu_N = \frac{\lambda_D}{\sqrt{g\rho}}\sqrt[2K]{\frac{1-(g\rho)^{K+1}}{1-g\rho}} \tag{A.10}$$

from which REC in percentage can be written as:

$$REC_{loss} = 100\left[\frac{\rho}{\sqrt{g\rho}}\sqrt[2K]{\frac{1-(g\rho)^{K+1}}{1-g\rho}} - 1\right] \tag{A.11}$$

Appendix B. Edge router selection

To construct the set of edge routers in a given ISP topology, we first apply two criteria:

- *Criterion I:* We include any router in the topology with a degree less than *Max Degree* or BFS distance greater (see Section 5.3 and Table 1) than *Min BFS Distance*.
- *Criterion II:* For each PoP, include at least one node if Criterion I did not select one. Choose the node with minimum degree within the PoP.

The intuition behind Criterion I is that nodes with smaller degree or longer distance from the center of the topology are more likely to be edge routers. For each of the Rocketfuel topology, we identified *Max Degree* and *Min BFS Distance* values so that the number of edge routers corresponds to 75–80% of the nodes in the topology.

We would like to note that even after Criterion II there still remains a small portion of links empty in the ISP topology, as this was also observed in [49]. The reason behind this is that the link weights measured by Rocketfuel are just a snapshot of the real link weights which are changing over time due to dynamism of routing. Another reason could be the fact that ISPs are deploying such extra links for backup purposes to increase tolerance to link failures. After the edge selection criteria above, some of the possible paths are eliminated and only g2g paths are left for traffic generation. Table 1 shows the specific *Min BFS Distance* and *Max Degree* values we used for each topology, and the number of edge routers selected by these thresholds.

Appendix C. Feasible edge-to-edge traffic rates

Given that set of edge routers in an ISP topology is identified as in Appendix B, the next step in constituting a network traffic model is to compose a *feasible* traffic matrix that imposes a traffic flow on every g2g path. To do so, we first construct an initial traffic matrix based on the gravity model using populations of the cities, and then adjust the link capacities so that traffic load on individual links are feasible.

C.1. Gravity model

The essence of the gravity model is that the traffic between two routers should be proportional to the multiplication of the populations of the two cities where the routers are located. This is inspired from the proportionality of the attraction force to the masses of two objects. Briefly, we used city populations to calculate the "mass" of each edge router and then calculated "mass-product" for each g2g path. We also use BFS-based link capacity model (see Section 5.3) to guide assigning a traffic rate in Mb/s to each g2g flow. The following steps detail our method:

- *Step 1: Calculate city populations.* We used CIESIN [50] dataset to calculate the city populations. CIESIN provides global population data in terms of a geographic grid (i.e. longitudes and latitudes) with 2.5' resolution. In addition to the population of grid cell, CIESIN also provides the land area within each grid cell. To calculate the population of a city we started with the central location of the city and spun on squares until a total land area of $2500km^2$ is covered.
- *Step 2: Calculate the "mass" of all edge routers.* As there may be multiple edge routers in the same city, we equally divided the population of the city to each edge router residing in that city. Then, we normalized the population pertaining to each edge router with respect to the edge router with the minimum population. This normalized populations are the masses for edge routers.
- *Step 3: Calculate the "mass product" for each g2g flow.* Let the mass for edge router i be M_i. Given the masses M_i and M_j for the edge routers i and j,$T_{i,j}$ must be proportional to $M_{i,j} = M_iM_j$, which we call as the *mass product*.
- *Step 4: Find the min-mass-product g2g flow.* We identify the g2g flow with minimum mass-product value. Let this min-mass-product g2g flow, $f_{u,v}$, be in between the edge routers u and v. So, we represent the min-mass-product with $M_{u,v}$.
- *Step 5: Find the max possible rate of $f_{u,v}$.* We use the BFS-based initial link capacities from Section 5.3 in Mb/s to calculate the traffic rate for the min-mass-product g2g flow $f_{u,v}$. In other words, the traffic matrix entry $T_{u,v}$ is dependent upon the bottleneck capacity on the path from u to v. To assure that there is a maximum limit on link utilizations (to avoid links with 100% utilizations), we impose a constant factor to the bottleneck capacity of the path u to v, i.e.,

$$T_{u,v} = \hat{c}_{u,v}\ \texttt{MAX\_LINK\_UTIL} \tag{C.1}$$

  where $\hat{c}_{u,v}$ is the bottleneck capacity of the path from u to v, and `MAX_LINK_UTIL` is the maximum possible link utilization of our network model, which we set to 95% in this paper. Notice that $T_{u,v}$ is the basic unit flow rate for the complete network model.
- *Step 6: Assign g2g flow rates in Mb/s.* We calculate the g2g flow rate in Mb/s from edge router i to j as:

$$T_{i,j} = \frac{M_{i,j}}{M_{u,v}} T_{u,v}. \tag{C.2}$$

This method of generating traffic matrices based on gravity models yields a power-law behavior in the flow rates as was studied earlier [49,44].

C.2. Handling infeasible links

After the initialization of the traffic matrix as outlined above, we still have to tackle the infeasible links as some links may have traffic loads larger than their estimated capacities. We increased the estimated capacity of the link so that the link capacity is just enough for the traffic load pertaining to it. We also assured that the highest link capacity is 40 Gb/s and there is always extra capacity so that the link utilization is never beyond `MAX_LINK_UTIL`.

References

[1] M. Yuksel, K.K. Ramakrishnan, S. Kalyanaraman, J.D. Houle, R. Sadhvani, Value of supporting class-of-service in IP backbones (short paper), in: Proc. of IEEE International Workshop on Quality of Service (IWQoS), Chicago, IL, June 2007.

[2] M. Yuksel, K.K. Ramakrishnan, S. Kalyanaraman, J.D. Houle, R. Sadhvani, Quantifying overprovisioning vs. class-of-service: Informing the net neutrality debate, in: Proceedings of IEEE International Conference on Computer Communication Networks (ICCCN), Zurich, Switzerland, August 2010.

[3] IPTV World Forum. <http://www.iptv-forum.com>.

[4] Global IP Traffic Forecast and Methodology, 2006–2011 (White Paper), Cisco Systems, 2007.

[5] Network Neutrality. <http://en.wikipedia.org/wiki/Network_neutrality>.

[6] J. Crowcroft, Net neutrality: the technical side of the debate: a white paper, ACM SIGCOMM Computer Communication Review 37 (1) (2007) 49–55.

[7] N. Spring, R. Mahajan, D. Wetherall, Measuring ISP topologies with Rocketfuel, in: Proceedings of Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM), 2002.

[8] W. Fischer, K. Meiher-Hellstern, The Markov-modulated poisson process (MMPP) cookbook, Performance Evaluation 18 (1992) 149–171.
[9] H. Jiang and C. Dovrolis, Why is the Internet traffic bursty in short time scales? in: Proc. of ACM SIGMETRICS, 2005.
[10] J. Beran, R. Sherman, M.S. Taqqu, W. Willinger, Long-range dependence in variable-bit-rate video traffic, IEEE Transactions on Communications 43 (1995) 1566–1579.
[11] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Transactions on Networking 2 (1) (1994) 1–15.
[12] M. Shreedhar, G. Varghese, Efficient fair queueing using deficit round robin, ACM SIGCOMM Computer Communication Review 25 (4) (1995) 231.
[13] R.J. Gibbens, S.K. Sargood, F.P. Kelly, H. Azmoodeh, R. Macfadyen, N. Macfadyen, An approach to service level agreements for IP networks with differentiated services, Philosophical Transactions of the Royal Society A358 (2000) 2165–2182.
[14] G. Cortese, R. Fiutem, P. Cremonese, S. D'Antonio, M. Esposito, S.P. Romano, A. Diaconescu, CADENUS: creation and deployment of end-user services in premium IP-networks, IEEE Communications Magazine 41 (1) (2003) 54–60.
[15] T. Engel, H. Granzer, B. Koch, M. Winter, P. Sampatakos, I. Venieris, H. Hussmann, F. Ricciato, S. Salsano, AQUILA: adaptive resource control for QoS using an IP-based layered architecture, IEEE Communications Magazine 41 (1) (2003) 46–53.
[16] F. Ciucu, A. Burchard, J. Liebeherr, A network service curve approach for the stochastic analysis of networks, in: Proc. of ACM SIGMETRICS, New York, NY, USA, 2005, pp. 279–290.
[17] J. Roberts, Internet traffic, QoS and pricing, in: Proc. of the IEEE, vol. 92, 2004, pp. 1389–1399.
[18] S. Oueslati, J. Roberts, A new direction for quality of service: flow aware networking, in: Proc. of NGI, 2005, pp. 1389–1399.
[19] S. Shenker, Fundamental design issues for the future Internet, IEEE Journal on Selected Areas of Communications 13 (1995) 1176–1188.
[20] F. Zhang, P.K. Verma, S. Cheng, Pricing, resource allocation and quality of service in multi-class networks with competitive market model, IET Communications 5 (1) (2011) 51–60.
[21] F. Zhang, P.K. Verma, A constant revenue model for packet switched network, in: Proceedings of IEEE Global Information Infrastructure Symposium, 2009.
[22] Y. Huang, R. Guerin, Does over-provisioning become more or less efficient as networks grow larger? in: Proc. of ICNP, 2005.
[23] S. Sahu, D. Towsley, J. Kurose, A quantitative study of differentiated services for the Internet, in: Proc. of IEEE GLOBECOM, 1999.
[24] F.P. Kelly, Models for a self-managed Internet, Philosophical Transactions of the Royal Society A358 (2000) 2335–2348.
[25] A.T. Andersen, B.F. Nielsen, A markovian approach for modeling packet traffic with long-range dependence, IEEE Journal on Selected Areas in Communications 16 (5) (1998) 719–732.
[26] A. Feldmann, W. Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, Performance Evaluation 31 (1998) 245–279.
[27] G.L. Choudhury, D.M. Lucantoni, W. Whitt, Squeezing the most out of ATM, IEEE Transactions on Communications 44 (2) (1996) 203–217.
[28] M.E. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes formulation and example, IEEE/ACM Transactions on Networking 5 (6) (1997) 835–846.
[29] V. Paxon, S. Floyd, Wide area traffic: the failure of Poisson modeling, IEEE/ACM Transactions on Networking 3 (3) (1995) 226–244.
[30] D. Burman, Insensitivity in queueing systems, Advances in Applied Probability 13 (1981).
[31] T. Bonald, A. Proutiere, Insensitivity in processor-sharing networks, Performance Evaluation 49 (2002) 193–209.
[32] Packet Length Distributions, CAIDA. <http://www.caida.org/analysis/AIX/plen_hist>.
[33] K.S. Trivedi, Probability and Statistics with Reliability, Queuing, and Computer Science Applications, Wiley Publishers, 2001.
[34] M. Yuksel, B. Sikdar, K.S. Vastola, B.K. Szymanski, Workload generation for ns simulations of wide area networks and the internet, in: Proc. of CNDS 2000, pp. 93–98.
[35] G. Appenzeller, I. Keslassy, N. McKeown, Sizing router buffers, in: Proc. of ACM SIGCOMM, 2004.
[36] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling tcp reno performance: a simple model and its empirical validation, IEEE/ACM Transactions on Networking (April) (2000).
[37] M. Mellia, I. Stoica, H. Zhang, TCP model for short lived flows, IEEE Communications Letters 6 (2) (2002) 85–87.
[38] E. Altman, T. Jimenez, Simulation analysis of RED with short lived TCP connections, Computer Networks 44 (5) (2004) 631–641.
[39] S. Floyd, V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Transactions on Networking 1 (4) (1993) 397–413.
[40] D. Ciullo, M. Mellia, M. Me, Two schemes to reduce latency in short lived TCP flows, IEEE Communications Letters 13 (10) (2009) 806–808.
[41] A. Kuzmanovic, The power of explicit congestion notification, in: Proceedings of ACM SIGCOMM, 2005, pp. 61–72.
[42] K.K. Ramakrishnan, S. Floyd, D. Black, The addition of explicit congestion notification (ECN) to IP, IETF Internet RFC 3168, September 2001.
[43] L. Li, D. Alderson, W. Willinger, J. Doyle, A first principles approach to understanding the Internet's router-level topology, in: Proc. of ACM SIGCOMM, 2004.
[44] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, C. Diot, Traffic matrix estimation: existing techniques and new directions, in: Proc. of ACM SIGCOMM, 2002.
[45] Y. Zhang, M. Roughan, C. Lund, D. Donoho, An information-theoretic approach to traffic matrix estimation, in: Proc. of ACM SIGCOMM, 2003.
[46] P. Goyal, A. Greenberg, C. Kalmanek, B. Marshall, P. Mishra, D. Nortz, K.K. Ramakrishnan, Integration of call signaling and resource management for IP telephony, IEEE Network 13 (3) (1999) 24–32.
[47] L. Kleinrock, Queueing Systems, Theory, vol. I, John Wiley and Sons, 1975.
[48] N.U. Prabhu, Foundations of Queueing Theory, Springer, 1997.
[49] R. Mahajan, D. Wetherall, T. Anderson, Negotiation-based routing between neighboring ISPs, in: Proc. of USENIX NSDI, 2005.
[50] The Center for International Earth Science Information Network (CIESIN). <http://www.ciesin.columbia.edu>.

Murat Yuksel is an Associate Professor at the CSE Department of The University of Nevada – Reno (UNR), Reno, NV. He was with the ECSE Department of Rensselaer Polytechnic Institute (RPI), Troy, NY as a Postdoctoral Research Associate and a member of Adjunct Faculty until 2006. He received a B.S. degree from Computer Engineering Department of Ege University, Izmir, Turkey in 1996. He received M.S. and Ph.D. degrees from Computer Science Department of RPI in 1999 and 2002 respectively. His research interests are in the area of computer communication networks with a focus on protocol design, network economics, wireless routing, free-space-optical mobile ad hoc networks (FSO-MANETs), and peer-to-peer. He is a senior member of IEEE, life member of ACM, and a member of Sigma Xi and ASEE.

Kadangode K. Ramakrishnan is a Distinguished Member of Technical Staff at AT&T Labs – Research in Florham Park, New Jersey. He joined AT&T Bell Labs in 1994 and has been with AT&T Labs Research since its inception in 1996. Prior to 1994, he was a Consulting Engineer and Technical Director in Networking at Digital Equipment Corporation. Between 2000 and 2002, he was at TeraOptic Networks, Inc. as Founder and Vice President. His current research interests are in networking and communications, including congestion control, multimedia distribution, content dissemination and problems associated with large scale distributed systems. He has published over 100 papers and has over 100 patents issued. He is an IEEE Fellow and an AT&T Fellow, recognized for his work on congestion control and VPN services. His contributions on congestion control, channel access protocols (for Ethernet and Cable), network interfaces, operating system support for network I/O, VPN services and IP Telephony have been adopted and implemented in the industry. K.K. has an MS degree from the Indian Institute of Science (1978), an MS (1981) and Ph.D. (1983) in Computer Science from University of Maryland, College Park. K.K. has been on the editorial board of the IEEE/ACM Transactions on Networking

and IEEE Network Magazine and has been a member of the National Research Council Panel on Information Technology for NIST. He has participated in numerous standards bodies working on communication networks.

Shivkumar Kalyanaraman is a Senior Manager at IBM Research-India. He was a Professor at the Department of Electrical, Computer and Systems Engineering at Rensselaer Polytechnic Institute in Troy, NY. He received a B.Tech. degree from the Indian Institute of Technology, Madras, India in July 1993, followed by M.S. and Ph.D. degrees in Computer and Information Sciences at the Ohio State University in 1994 and 1997 respectively. His research interests are in network traffic management topics such as congestion control architectures, quality of service (QoS), high-speed wireless, free-space optical networking, network management, multicast, pricing, multimedia networking, and performance analysis. His special interest lies in developing the inter-disciplinary areas between traffic management, wireless communication, optoelectronics, control theory, economics, scalable simulation technologies, and video compression. He is a member of the ACM and IEEE.

Joseph D. Houle has been with AT&T for 25 years. He has extensive experience in Data Communication with a background in equipment design, service definition and network implementation. Joe is most recently working on the network capabilities to enable cloud computing. Previously Joe had focused on Content delivery technologies including studies on the economics of Net Neutrality. Joe also contributed to early IPv6 service provider industry evolution plans and some preliminary Wi-Fi offers. Joe has an M.S. in Computer Science from Johns Hopkins University and a B.S. in IE/OR from Rutgers University.

Rita Sadhvani is an Associate Director at Verizon Wireless, Basking Ridge, NJ. Prior to this, she was with AT&T Consumer Services until 2008. She received a B.S. degree in Electrical Engineering from Indian Institute of Technology, Kanpur, India in 1987. She received a master's degree in Economics at Rutgers University, New Brunswick, NJ in 1994. Her research interests include network economics and global band diversity.