Coordinated Network Scheduling:
A Framework for End-to-End Services

Chengzhi Li   Edward W. Knightly
Department of Electrical and Computer Engineering
Rice University

Abstract

In multi-hop networks, packet schedulers at downstream nodes have an opportunity to make up for excessive latencies due to congestion at upstream nodes. Similarly, when packets incur *low* delays at upstream nodes, downstream nodes can *reduce* priority and schedule other packets first. The goal of this paper is to define a framework for design and analysis of *Coordinated Network Scheduling* (CNS) which exploit such inter-node coordination. We first provide a general CNS definition which enables us to classify a number of schedulers from the literature including, FIFO+, CEDF, and work-conserving CJVC as examples of CNS schedulers. We then develop a distributed theory of traffic envelopes which enables us to derive end-to-end statistical admission control conditions for CNS schedulers. We show that CNS schedulers are able to limit traffic distortion to within a narrow range resulting in improved end-to-end performance and more efficient resource utilization.

1 Introduction

During periods of congestion, a flow or class' end-to-end performance properties are strongly influenced by the choice of the packet scheduling algorithm employed at the network's routers. Consequently, recent advances in scheduler design can ensure properties such as fairness, performance differentiation, and performance isolation [3, 13, 15, 25]. Moreover, such performance properties are now achievable in high speed implementations [23, 28, 30] and scalable architectures in which core nodes do not maintain per-flow state [6, 24, 31].

Exploiting these scheduling mechanisms, admission control can limit congestion levels so that (for example) targeted latencies and throughputs are ensured, thereby providing services with predictable and controlled performance levels [21]. For example, statistical class-based admission control tests have been derived for Earliest Deadline First [4, 26, 29], Weighted Fair Queueing [12, 26, 36], Strict Priority [26], and Virtual Clock [19]. Moreover, techniques for providing multi-node or end-to-end statistical services have been developed for several classes of non-work-conserving schedulers [5, 27, 29, 35] and for Weighted Fair Queueing [36].

However, in both the data plane (scheduling) and control plane (admission control), none of the aforementioned techniques exploit a key property of multihop networks, namely, that a downstream node can compensate for excessive latency or unfairness incurred at an upstream node. Nor will downstream nodes *reduce* the priority of a packet which arrives ahead of schedule due to a lack of congestion upstream. In contrast, a number of service disciplines in the literature have been proposed which *do* exploit this property, which we refer to as *coordination*. Examples include the old customer first service discipline [8], global earliest deadline first (G-EDF) [8], modified first-in-first-out (FIFO+) [10], and coordinated earliest-deadline-first (CEDF) [1, 2].

The contributions of this paper are twofold. First, we devise a general framework for design and specification of a class of service disciplines which we refer to as Coordinated Network Schedulings (CNS). The key CNS property is that a packet's priority index at a downstream node is recursively expressed through the priority index of the same packet at the previous node, and therefore is a function of the packet's (perhaps virtual) entrance time into the network. We show that a broad class of schedulers from the literature, including CEDF, FIFO+, and others, can be characterized by this recursion and belong to the CNS class. We make several observations regarding coordinated network schedulers. (1) They can be core-stateless, in some cases quite trivially, and therefore can share the same scalability properties of architectures in which core nodes do not maintain per-flow state [31]. (2) The well known *traffic distortion* problem, in which provisioning of end-to-end services is hampered by complex traffic distortions due to multiplexing, e.g., [11, 22], can be avoided all together. (3) CNS inter-server cooperation can improve a flow's end-to-end performance, and consequently, improve the efficiency and utilization of the network at large.

Our second contribution is to devise a general theory for statistical analysis and admission control of coordinated servers. Our key technique is to devise a framework for end-to-end service provisioning that exploits the structural properties of coordinated network schedulers, thereby overcoming the traffic distortion problem and realizing the efficiency gains of coordination. To analyze CNS networks, we introduce the concept of *essential traffic*, which is the traffic that must be served before a time instant such that no local service violations will occur at that time. Using this concept and building on the interclass theory of [26], we derive expressions for the essential traffic and service *envelopes*, which provide a general statistical characterization of a CNS node's workload and service capacity. Within this framework, we establish an important property of the CNS discipline, namely, that traffic distortions in CNS networks are limited to within a narrow range. Therefore, the essential traffic and service envelopes at a CNS node can be evaluated as simple and minimally distorted functions of the flows' *original* (undistorted) traffic envelopes that characterized traffic before entrance into the network. We then derive CNS admission control conditions by transforming the problem of evaluating the service-violation probability into the problem of computing the essential traffic envelope and the essential service envelope.

Previous techniques for multi-node admission control include studies of non-work-conserving schedulers which shape and reshape traffic [5, 14, 16, 17, 27, 29, 34, 35]. While such schemes can have good performance properties, they require per-flow traffic processing in core nodes and do not exploit the coordination property. For work-conserving service disciplines, a key issue is traffic distortion. Previous approaches include bounding this distortion [7, 11, 22, 33] and exploiting isolation properties of GPS servers [18, 25, 36]. While such techniques are important for their generality, we will show that they can be conservative in practice. In contrast, our work develops a general framework for end-to-end services in CNS networks. Our solution applies to the broad class of (work-conserving) CNS servers, exploits the efficiency gains of coordination, and provides an end-to-end admission control algorithm that is quite general and achieves high utilization for multi-class multi-node services.

The remainder of this paper is organized as follows. In Section 2, we define the CNS discipline and show how scheduling algorithms from the literature can be classified within the CNS framework. Next, in Section 3 we develop a general theory for analysis and admission control for statistical end-to-end services. Finally, in Section 4, we provide numerical admission control results, and in Section 5 we conclude.

2 Framework for Coordinated Scheduling

In this section, we provide a formal definition of the CNS coordination property. We then use this definition to show how a number of schedulers from the literature possess this property so that our admission control tests derived in Section 3 apply to a broad class of schedulers.

2.1 CNS Definition

Definition 1 (Coordinated Network Scheduling) *Consider a multiplexer which services packets in increasing order of their priority indexes. A scheduler possesses the CNS property if*

$$d_{i,j}^k = \begin{cases} t_i^k + \delta_{i,1}^k, & j = 1 \\ d_{i,j-1}^k + \delta_{i,j}^k, & j > 1 \end{cases} \qquad (1)$$

where $d_{i,j}^k$ is the priority index assigned to the k^{th} packet of flow i at its j^{th} hop; t_i^k is the time when the k^{th} packet of flow i arrives at its first hop; and $\delta_{i,j}^k$ is the increment of the priority index of the k^{th} packet of flow i at its j^{th} hop.

Based on the selected method for incrementing the priority index, we sub-classify CNS service disciplines into delay- and rate-CNS. A service discipline belongs to the *delay-CNS* class if $\delta_{i,j}^k$ represents a delay parameter of the k^{th} packet of flow i at its j^{th} hop. For example, this delay parameter can be simply a local delay bound, or for other service disciplines (described below), can be a function of packet k's delay relative to the scheduler's mean delay.

In contrast, a service discipline belongs to *rate-CNS* class if $\delta_{i,j}^k$ is a function of l_i^k and $r_{i,j}$, where l_i^k is the size of the k^{th} packet of flow i and $r_{i,j}$ is the reserved bandwidth for flow i at its j^{th} hop. The main characteristic of this class is that reserved bandwidths rather than delay bounds determine the service priority. Below we also describe examples of schedulers belonging to the rate-CNS class.

2.2 Discussion

The key property of the CNS discipline is that the priority index of each packet at a downstream server depends on its priority index at upstream servers, so that all servers in the network cooperate to provide the end-to-end service. For example, if a packet violates a local deadline at an upstream server, downstream nodes will increase the packet's priority thereby increasing the likelihood that the packet will meet its end-to-end delay bound. Similarly, if a packet arrives "early" due to a lack of congestion upstream, downstream nodes will reduce the priority of the packet.

As a simple example, consider a packet arriving to the network at $t = 0$ and traversing two nodes with $\delta_{i,1}^{k} = \delta_{i,2}^{k} = 10$ msec. At the second node, the packet's deadline (priority index) is 20 msec. If, for example, the packet misses its local delay bound at the first node and is queued for 15 msec, the second node prioritizes the packet with a deadline 5 msec after its arrival. This is contrast with EDF, in which each node schedules packets independently, so that the second node would prioritize the packet with a deadline *10 msec* after its arrival, making it far less likely that the packet would meet its 20 msec end-to-end requirement.

2.3 Example CNS Disciplines

The above definition of Coordinated Network Scheduling is quite general. Here, we show how several service disciplines from the literature, including G-EDF [8], FIFO+ [10], CJVC [32], and CEDF [1, 2] can be classified as instances of the CNS discipline.

Global EDF The Global Earliest Deadline First (G-EDF) service discipline was introduced in [8] to address the problem that reconstruction of continuous speech from voice packets is complicated by variable delays of packets due to multiplexing. In G-EDF, the priority index for a packet with age (time in network) α arriving at a server at time t is defined as $t + (D_{max} - \alpha - d)$, where D_{max} is the maximum allowable entry-to-exit delay and d is the estimated delay along the packet's remaining route in the network. If we rewrite the priority index assigned by G-EDF as:

$$d_{i,j}^{k} = \begin{cases} t_i^k + (D_i^{max} - \sum_{h=2}^{N_i} \delta_{i,h}), & j = 1 \\ d_{i,j-1}^{k} + \delta_{i,j}, & j > 1 \end{cases} \qquad (2)$$

where N_i is the path length of flow i, and $\delta_{i,j}$ is the expected delay suffered by the packets of flow i at its j^{th} hop, then it is clear that G-EDF is a delay-CNS discipline.

FIFO+ The modified first-in-first-out (FIFO+) service discipline [10] assigns a packet's priority index according to the difference between the average queueing delay seen by a packet and the particular queueing delay suffered by the packet at upstream servers. From the definition in [10], we can rewrite the recursive FIFO+ priority index as:

$$d_{i,j}^{k} = \begin{cases} t_i^k, & j = 1 \\ d_{i,j-1}^{k} + (\bar{d}_{i,j-1} - \hat{d}_{i,j-1}^{k}) & j > 1 \end{cases} \qquad (3)$$

so that $\delta_{i,j}^{k}$ is the difference between $\bar{d}_{i,j-1}$, the average queueing delay of flow i, and $\hat{d}_{i,j-1}^{k}$, the particular queueing delay at the $(j-1)^{th}$ hop. Comparing Equation (3) with Definition 1 shows that FIFO+ is also a delay-CNS discipline.

Work Conserving CJVC Core-Jitter Virtual Clock (CJVC) was proposed in [32] as a mechanism for achieving guaranteed service without per-flow state in the network core. CJVC uses "dynamic packet state" to store information in each packet header containing the eligible time of the packet at the ingress router and a slack variable that allows core routers to determine the local priority index of the packet. For a work-conserving variant of CJVC, the priority index of packet k of flow i at node j is given by:

$$d_{i,j}^{k} = \begin{cases} t_i^k + \frac{l_i^k}{r_i} + \eta_i^k, & j = 1 \\ d_{i,j-1}^{k} + \frac{l_i^k}{r_i} + \eta_i^k, & j > 1 \end{cases} \qquad (4)$$

where flow-i k^{th} packet size and reserved bandwidth are given by l_i^k and r_i respectively, and η_i^k is the slack variable assigned to the k^{th} packet of flow i before it enters the network. Thus, work-conserving CJVC is a rate-CNS service discipline.

Coordinated EDF In [1, 2], the Coordinated Earliest Deadline First (CEDF) service discipline is developed with the goal of minimizing end-to-end delays. The approach is to use EDF together with randomization of packet injection time and coordination of servers. There exist two ways to assign local deadline in CEDF service discipline.

In [2], the priority indexes are assigned as

$$d_{i,j}^{k} = \begin{cases} \tau_i^k + G_{i,1}, & j = 1 \\ d_{i,j-1}^{k} + G_{i,j}, & j > 1 \end{cases} \qquad (5)$$

where τ_i^k is the token arrival time chosen uniformly at random from interval $[(k-1)T_i, kT_i)$, $T_i = 2^{\lceil \frac{2L_i}{\epsilon \rho_i} \rceil}$, L_i is the maximum size of flow-i packets, ρ_i is the rate of flow i, ϵ is the utilization factor, and $G_{i,j}$ is a constant (expected local delay bound) determined for the j^{th} hop of flow i.

In [1], the priority indexes are assigned as

$$d_{i,j}^{k} = \begin{cases} \tau_i^k + \frac{(T_i - \tau_i^k + t_i^k) C_{i,1}}{\sum_{h=1}^{N_i} C_{i,h}}, & j = 1 \\ d_{i,j-1}^{k} + \frac{(T_i - \tau_i^k + t_i^k) C_{i,j}}{\sum_{h=1}^{N_i} C_{i,h}}, & j > 1, \end{cases} \qquad (6)$$

where T_i is the end-to-end delay bound for flow i, $\tau_i^k \in [t_i^k, t_i^k + T_i)$ is the arrival time of token for the k^{th} packet of flow i (similar to above), the C_{ij} is the capacity of the server in the j^{th} hop of flow i, and N_i is the path length of flow i. Thus, both variants of CEDF can be classified as delay-CNS disciplines in which the first priority index is randomized.

3 CNS Analysis and Admission Control

In this section, we develop a statistical multi-node analysis and admission control algorithm for CNS. We proceed in several steps. First, we introduce two key concepts needed for analysis: essential traffic and essential service. These concepts enable us to statistically bound the traffic that must be serviced in order to meet a flow's local quality-of-service constraints. We next show how the essential traffic at a core node can be computed based on a simple and minimally distorted transformation of the traffic at the *entrance* of the network. This result (Theorem 1) is a key to efficient end-to-end analysis. We then derive an expression for the statistical *service* envelope (Theorem 2): with this statistical description of service, we can characterize and control statistical sharing across traffic classes. Finally, we derive an end-to-end admission control test for coordinated schedulers (Theorem 3).

Throughout, we denote $f_{i,j}(t)$ as the total traffic in $[0,t)$ arriving from traffic flow i at flow-i j^{th} hop, a node which is indexed by $\pi(i,j)$.[1] Without loss of generality, we ignore propagation delays so that the departure traffic of flow i from server $\pi(i,j)$ is the arrival traffic of flow i to server $\pi(i,j+1)$. As in [26], we call a sequence random of variables $\{B_i(I)\}_{I=0}^{\infty}$ a statistical traffic envelope of flow i if $\forall t, I > 0$[2]

$$f_{i,1}(t+I) - f_{i,1}(t) \leq B_i(I). \qquad (7)$$

Finally, We consider a discrete time model with both dropping (finite buffer) and non-dropping (infinite buffer) scheduling.

3.1 Essential Traffic

Here, we define *essential traffic* as a fundamental notion for analysis of coordinated schedulers that enables us to accurately evaluate a flow's delay-bound-violation probability. In particular, for a given local deadline s, all arriving traffic of server m arriving in $[0,t)$ can be virtually decomposed according to whether or not its local deadline is later than s. As only the portion of traffic with local deadline no later than time s affects the probability of violating the local deadline s, we refer to this traffic as essential traffic, which we formally define as follows.

Definition 2 (Essential Traffic) *The essential arrival traffic $f^*_{i,j}(t,s)$ of flow i at time t relative to time s at server $\pi(i,j)$ is defined as the total flow-i traffic with local deadline no later than time s arriving at server $\pi(i,j)$ in $[0,t)$, i.e.,*

$$f^*_{i,j}(t,s) = \min\{f_{i,j}(t), f_{i,1}(s - \sum_{k=1}^{j} \delta_{i,k})\}. \qquad (8)$$

Term	Definition
$\pi(i,j)$	j^{th} hop of flow i
N_i	path length of flow i
t_i^k	arrival time of the k^{th} packet of flow i at its first hop
$\delta_{i,j}^k$	increment of priority index of the k^{th} packet of flow i at its j^{th} hop
$f_{i,j}(t)$	total flow-i traffic at its j^{th} hop during $[0,t)$
$B_i(I)$	flow i statistical traffic envelope at its first hop
$f^*_{i,j}(t,s)$	flow-i traffic with local deadline no later than s arriving at server $\pi(i,j)$ during $[0,t)$
$\Gamma_{i,j}(I)$	flow i essential traffic envelope at its j^{th} hop
$S(I,\alpha)$	flow i essential service envelope at its j^{th} hop
$\tau_m(t,s)$	void time of server m before time t related to time s
$W_m^s(x)$	total traffic with local deadline no later than s queued at server m at time x
$\psi_{i,j}$	the difference between maximum and expected queueing delay for flow i traffic before arriving at its j^{th} hop
$\lambda_{i,j}$	probability of flow-i traffic missing its local deadline before arriving at its j^{th} hop
$d_{i,j}(t)$	(virtual) essential delay suffered by flow-i traffic at its j^{th} hop at time t
$D_{i,j}$	flow i delay bound at its j^{th} hop
$\epsilon_{i,j}$	upper bound on the probability that $d_{i,j}(t) > D_{i,j}$
$\tilde{\epsilon}_{i,j}$	upper bound on the probability of flow-i traffic missing its priority index at its j^{th} hop

Table 1: Notation

For a multiplexer (server) in a CNS network and given time t and local deadline s, an important instant previous to t is the time when the multiplexer does not service traffic with local deadlines later than s. As we see below, this is important for analysis because the traffic arriving at the multiplexer before this moment does not affect the probability of local deadline violation. We refer to this instant as the *void time* denoted as $\tau_m(t,s)$ and precisely define it as

$$\tau_m(t,s) = \max\{x \mid x \leq t \text{ and } W_m^s(x) = 0\}, \qquad (9)$$

where $W_m^s(x)$ is the total amount of traffic with local deadline no later than s backlogged at server m at time x.[3] Notice that server m is not necessarily idle at time $\tau_m(t,s)$

[1] Notation is summarized in Table 1.

[2] Throughout, $X \leq Y$ denotes almost sure inequality, $P[X \leq Y] = 1$.

[3] Without loss of generality, we assume that network is idle at time 0.

as it may be busy serving traffic with local deadlines later than s.

3.2 Essential Traffic Envelope

According to the definition of *void time*, we are not concerned with traffic arriving before $\tau_m(t,s)$ when computing the local delay-bound-violation probability. Thus, we define the essential traffic envelope as follows.

Definition 3 (Essential Traffic Envelope) *A sequence of random variables $\{\Gamma_{i,j}(I)\}_{I=0}^{\infty}$ is the essential traffic envelope of flow i at its j^{th} hop if $\forall t, s > 0$,*

$$f_{i,j}^*(t,s) - f_{i,j}^*(\tau_m(t,s),s) \le \Gamma_{i,j}(s - \tau_m(t,s)), \quad (10)$$

where $\pi(i,j) = m$ and $\tau_m(t,s)$ is defined in Equation (9).

A key challenge for provisioning multi-node services is evaluating the essential traffic envelope at core servers. The difficulty is due to the fact that a flow's traffic is unavoidably distorted after multiplexing with other flows so that the local deadline for a packet at a core server depends not only on its arrival time and local deadline increment, but also the queueing delay suffered by the packet at upstream servers.

The following lemma bounds the traffic missing its local deadline at upstream server for stable networks in which server m has a maximum queueing delay T_m.

Lemma 1 *For a given time s, the total amount of flow-i traffic that misses its local deadline (priority index) no later than s at the $(j-1)^{th}$ hop, and arrives at the j^{th} hop after s is bounded by $\lambda_{i,j}B_i(\psi_{i,j})$, where*

$$\psi_{i,j} = \sum_{h=1}^{j-1}(T_{\pi(i,h)} - \delta_{i,h})$$

and $\lambda_{i,j}$ is the probability of flow-i traffic missing its local deadlines before arriving its j^{th} hop, as computed in Corollary 2.

Proof: According to the definition of CNS, flow-i traffic arriving at server $\pi(i,j)$ after s and having local deadline before s at server $\pi(i,j-1)$, arrives at its first hop $\pi(i,1)$ during the interval

$$[s - \sum_{k=1}^{j-1}\delta_{i,k} - \psi_{i,j}, s - \sum_{k=1}^{j-1}\delta_{i,k}).$$

This is because if flow-i traffic arrives at its first hop *after* $s - \sum_{k=1}^{j-1}\delta_{i,k}$, the corresponding local deadline at server $\pi(i,j-1)$ is later than s; and if the traffic of flow i arrives at its first hop *before*

$$s - \sum_{k=1}^{j-1}\delta_{i,k} - \psi_{i,j} = s - \sum_{k=0}^{j-1}T_{\pi(i,k)},$$

then it should arrive at server $\pi(i,j)$ before time s, otherwise there is a contradiction with the definition of $T_{\pi(i,k)}, k = 1, \cdots, j-1$. Furthermore, the probability of flow-i traffic missing its local deadlines before arriving at server $\pi(i,j)$ is bounded by $\lambda_{i,j}$. Hence the total amount of flow-i traffic missing its local deadline no later than time s and arriving at server $\pi(i,j)$ after s is bounded by

$$\begin{aligned}\lambda_{i,j}[f_{i,1}(s - \sum_{k=1}^{j-1}\delta_{i,k}) - f_{i,1}(s - \sum_{k=1}^{j-1}\delta_{i,k} - \psi_{i,j})] \\ \le \quad \lambda_{i,j}B_i(\psi_{i,j}). \qquad \Box\end{aligned}$$

Based on this bound on the traffic missing its local deadline, we can evaluate the essential traffic envelope as follows.

Theorem 1 *The essential traffic envelope $\Gamma_{i,j}(I)$ of flow i is given by*

$$\Gamma_{i,j}(I) = B_i(I - \delta_{i,j}) + \lambda_{i,j}B_i(\psi_{i,j}), \quad (11)$$

Proof: Let $\pi(i,j) = m$. For all $t, s > 0$, consider the interval $[\tau_m(t,s), t)$. According to Equation (9), during this interval server m is busy serving traffic with local deadlines no later than s. According to Definition 2, we have

$$f_{i,j}^*(t,s) \le f_{i,1}(s - \sum_{k=1}^{j-1}\delta_{i,k} - \delta_{i,j}). \quad (12)$$

For $\tau_m(t,s) \le s - \delta_{i,j}$, we have

$$\begin{aligned}& f_{i,j}^*(\tau_m(t,s),s) \\ & \quad \ge \quad f_{i,1}(\tau_m(t,s) - \sum_{k=1}^{j-1}\delta_{i,k}) - \lambda_{i,j}B_i(\psi_{i,j}).\end{aligned}$$

This is because at time $\tau_m(t,s)$, the queue of server m does not contain packets with deadlines later than s. After $\tau_m(t,s)$, if all arriving packets have not missed their local deadline before arriving at server m, at least $f_{i,1}(\tau_m(t,s) - \sum_{k=1}^{j-1}\delta_{i,k})$ traffic of flow i with local deadline no later than s at server $\pi(i,j)$ has departed server $\pi(i,j-1)$ and arrived at server m before time $\tau_m(t,s)$. However, there may be packets missing their local deadline before arriving at server m. According to Lemma 1, we know that after time $\tau_m(t,s)$ and at server m, the arriving traffic of flow i with deadline less than $\tau_m(t,s)$ at server $\pi(i,j-1)$ is bounded by $\lambda_{i,j}B_i(\psi_{i,j})$. Therefore, we have

$$\begin{aligned}& f_{i,j}^*(t,s) - f_{i,j}^*(\tau_m(t,s),s) \\ & \quad \le \quad f_{i,j}(s - \sum_{k=1}^{j-1}\delta_{i,k} - \delta_{i,j}) - \\ & \qquad\quad f_{i,1}(\tau_m(t,s) - \sum_{k=1}^{j-1}\delta_{i,k}) + \lambda_{i,j}B_i(\psi_{i,j}) \\ & \quad \le \quad B_i(s - \tau_m(t,s) - \delta_{i,j}) + \lambda_{i,j}B_i(\psi_{i,j}).\end{aligned}$$

For $\tau_m(t,s) > s - \delta_{i,j}$, all flow-$i$ packets with local deadline no later than s arriving at server m during $[\tau_m(t,s), s)$ have missed their local deadline before arriving server m. Hence as above, we have

$$\begin{aligned} f^*_{i,j}(t,s) - f^*_{i,j}(\tau_m(t,s),s)) & \\ \leq \quad & \lambda_{i,j} B_i(\psi_{i,j}) \\ \leq \quad & B_i(s - \tau_m(t,s) - \delta_{i,j}) + \lambda_{i,j} B_i(\psi_{i,j}). \end{aligned}$$

Therefore, we have

$$\Gamma_{i,j}(I) = B_i(I - \delta_{i,j}) + \lambda_{i,j} B_i(\psi_{i,j}). \quad \square$$

This theorem reveals the important property of the CNS discipline, namely, that distortion of traffic in core servers is limited to within a narrow range. To illustrate, consider the special case in which packets that miss their local deadlines are discarded. In this case, $\psi_{i,j} = 0$ for any flow i and its j^{th} hop so that the essential traffic envelope at core servers, i.e., $\Gamma_{i,j}(I) = B_i(I - \delta_{i,j})$, is almost identical to that at the ingress server, $\Gamma_{i,1}(I) = B_i(I - \delta_{i,1})$.

3.3 Essential Service Envelope

The above result enables us to derive end-to-end admission control tests for CNS networks in the *single* class case. However, with multiple traffic classes with statistical sharing across classes, classes affect each others' performance. Consequently, characterizing the extent to which resources are shared across classes is the key to achieving high utilization in multi-class networks without worst-case allocation for each class [26]. Thus, we use statistical service envelopes as a tool for characterizing and controlling interclass resource sharing.

Definition 4 (Essential Service Envelope) *A sequence of random variables $\{S_{i,j}(I,\alpha)\}_{I=0}^{\infty}$ is called a (statistical) essential service envelope provided by server $\pi(i,j)$ to the traffic of flow i, if $\forall t, s > 0 \; \exists \; \theta$[4] such that*

$$f_{i,j+1}(t) \geq S_{i,j}(t - \theta, t - s) + f^*_{i,j}(\theta, s). \qquad (13)$$

Roughly, $S_{i,j}(I,\alpha)$ is a measure of the service provided in any busy interval with length I to flow-i traffic with local deadline α seconds before the end of the interval. Notice that the minimum service obtained by flow i during a busy interval of server $\pi(i,j)$ will depend on the other flows' arriving traffic. Furthermore, the essential service envelope provided by server $\pi(i,j)$ to flow i depends on the other flows' essential traffic. Using this definition, we can now derive an expression for the essential service envelope.

Theorem 2 *For all $I > 0$,*

$$S_{i,j}(I,\alpha) \quad = \quad (C_m I - \sum_{k \in \mathcal{I}(m), k \neq i} \Gamma_{k,j_k}(I - \alpha))^+,$$

where j_k is defined by $\pi(k, j_k) = \pi(i,j) = m$, $\mathcal{I}(m)$ is the set of flows served by server m, and C_m is the capacity of server m.

Proof: According to Definition 4, we need to show that $\forall t, s > 0$, there exists a θ such that

$$f_{i,j+1}(t) \geq S_{i,j}(t - \theta, t - s) + f^*_{i,j}(\theta, s). \qquad (14)$$

If $f_{i,j+1}(t) \geq f^*_{i,j}(t,s)$, let $\theta = t$ and we have $f_{i,j+1}(t) \geq S_{i,j}(0, t - s) + f^*_{i,j}(t,s) = f^*_{i,j}(t,s)$. Otherwise $f_{i,j+1}(t) < f^*_{i,j}(t,s)$, and there exists some flow-i packets with local deadline less than s, queued at server m at time t. Since time $\tau_m(t,s)$ is the last time before time t when server m does *not* serve traffic with local deadline no later than s, there exists at most $\Gamma_{k,j_k}(s - \tau_m(t,s))$ traffic of flow k with local deadline no later than time s in the interval $[\tau_m(t,s), t)$. This is because $f^*_{k,j_k}(t,s) - f^*_{k,j_k}(\tau_m(t,s), s) \leq \Gamma_{k,j_k}(s - \tau_m(t,s))$.

Therefore, there at least exists $S_{i,j}(t - \tau_m(t,s), t - s) = (C_m(t - \tau_m(t,s)) - \sum_{k \neq i, k \in \mathcal{I}(m)} \Gamma_{k,j_k}(s - \tau_m(t,s)))^+$ flow-i traffic with local deadline no later than s served by server m during $[\tau_m(t,s), t)$. Furthermore, at time $\tau_m(t,s)$, the $f^*_{i,j}(\tau_m(t,s), s)$ of flow-i traffic has been served. Hence, let $\theta = \tau_m(t,s)$, and we have

$$\begin{aligned} f_{i,j+1}(t) \quad & \geq \quad S_{i,j}(t - \tau_m(t,s), t - s) + f^*_{i,j}(\tau_m(t,s), s) \\ & = \quad S_{i,j}(t - \theta, t - s) + f^*_{i,j}(\theta, s). \quad \square \end{aligned}$$

3.4 Admission Control

We now derive an end-to-end admission control condition for CNS networks. We use a concept of (virtual) delay due to the essential traffic at a particular node to derive the *local* delay-bound-violation probability as an intermediary step towards bounding the *end-to-end* probability. Thus, we define (virtual) essential delay $d_{i,j}(t,s)$ of flow-i traffic with local deadline no later than time s at server $\pi(i,j)$ at time t as

$$d_{i,j}(t,s) = \min\{\triangle : f^*_{i,j}(t,s) \leq f_{i,j+1}(s + \triangle)\}. \quad (15)$$

For a flow-i packet with local deadline s arriving at server $\pi(i,j)$ at time t, the event of this packet being served after time $s + D_{i,j}$ is contained in the event $\{d_{i,j}(t,s) > D_{i,j}\}$, and we henceforth consider this latter event. The following theorem shows how to evaluate this delay distribution.

[4] θ depends on t, s and $f_{i,j}(t)$.

Theorem 3 *The virtual delay distribution of flow i at its j^{th} hop is bounded by:*

$$\begin{aligned}&P[d_{i,j}(t,s) > D_{i,j}] \leq \\ &\quad P[\max_{I>0}\{\Gamma_{i,j}(I) - S_{i,j}(I + D_{i,j}, D_{i,j})\} > 0]. \quad (16)\end{aligned}$$

Proof: From Equation (15), we have

$$\{d_{i,j}(t,s) > D_{i,j}\} \equiv \{f^*_{i,j}(t,s) - f_{i,j+1}(s + D_{i,j}) > 0\}.$$

If $f^*_{i,j}(t,s) > f_{i,j+1}(s + D_{i,j})$ at time $s + D_{i,j}$ at server $\pi(i,j)$, there exists traffic with local deadline no later than s and arriving at t. Therefore $\tau_m(s + D_{i,j}, s) = \tau_m(t,s)$. Let $\theta = \tau_m(s + D_{i,j}, s)$, according to Theorem 2, we have

$$f_{i,j+1}(s + D_{i,j}) \geq S_{i,j}(s + D_{i,j} - \theta, D_{i,j}) + f^*_{i,j}(\theta, s).$$

Thus,

$$\begin{aligned}&\{f^*_{i,j}(t,s) - f_{i,j+1}(s + D_{i,j}) > 0\} \subseteq \{f^*_{i,j}(t,s) - \\ &\qquad [S_{i,j}(s + D_{i,j} - \theta, D_{i,j}) + f^*_{i,j}(\theta, s)] > 0\}.\end{aligned}$$

Furthermore, according to Theorem 1, we have

$$f^*_{i,j}(t,s) - f^*_{i,j}(\theta,s) \leq \Gamma_{i,j}(s - \theta).$$

Therefore,

$$\begin{aligned}&\{d_{i,j}(t,s) > D_{i,j}\} \\ &\quad \subseteq \quad \{\Gamma_{i,j}(s-\theta) - S_{i,j}(s + D_{i,j} - \theta, D_{i,j}) > 0\} \\ &\quad \subseteq \quad \{\max_{I>0}\{\Gamma_{i,j}(I) - S_{i,j}(I + D_{i,j}, D_{i,j})\} > 0\}. \quad \square\end{aligned}$$

Thus, according to Theorem 3, the problem of computing the flow-i delay distribution is transformed into the problem of finding flow-i essential traffic envelope and essential service envelope. Based on Theorem 1 and Theorem 2, we have the following results.

Corollary 1 *For all $t, s > 0$,*

$$P[d_{i,j}(t,s) > D_{i,j}] \leq \epsilon_{i,j}, \qquad (17)$$

where

$$\begin{aligned}\epsilon_{i,j} = P[\max_{I>0}\{&\sum_{k\in\mathcal{I}(m)} [B_k(I - D_{i,j} - \delta_{k,j_k}) + \\ &\lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_m I\} > 0], \qquad (18)\end{aligned}$$

and m and j_k are defined by $\pi(k, j_k) = \pi(i,j) = m$, and λ_{k,j_k} and ψ_{k,j_k} are defined in Lemma 1.

Proof: From Theorem 3, we have

$$\begin{aligned}&P[d_{i,j}(t,s) > D_{i,j}] \leq \\ &\qquad P[\max_{I>0}\{\Gamma_{i,j}(I) - S_{i,j}(I + D_{i,j}), D_{i,j})\} > 0].\end{aligned}$$

From Theorem 2, we have

$$\begin{aligned}&P[\max_{I>0}\{\Gamma_{i,j}(I) - S_{i,j}(I + D_{i,j}), D_{i,j})\} > 0] \leq \\ &\qquad P[\max_{I>0}\{\Gamma_{i,j}(I) - \\ &\qquad (C_m(I + D_{i,j}) - \sum_{k\in\mathcal{I}(m), k\neq i} \Gamma_{k,j_k}(I))^+\} > 0].\end{aligned}$$

From Theorem 1, we have

$$\begin{aligned}&P[\max_{I>0}\{\Gamma_{i,j}(I) - \\ &\qquad (C_m(I + D_{i,j}) - \sum_{k\in\mathcal{I}(m), k\neq i} \Gamma_{k,j_k}(I))^+\} > 0] \\ &\quad \leq \quad P[\max_{I>0}\{B_i(I - \delta_{i,j}) + \lambda_{i,j} B_i(\psi_{i,j}) \\ &\qquad -(C_m(I + D_{i,j}) - \sum_{k\in\mathcal{I}(m), k\neq i} [B_k(I - \delta_{k,j_k}) \\ &\qquad +\lambda_{k,j_k} B_k(\psi_{k,j_k})])^+\} > 0] \\ &\quad \leq \quad P[\max_{I>0}\{\sum_{k\in\mathcal{I}(m)} [B_k(I - D_{i,j} - \delta_{k,j_k}) + \\ &\qquad \lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_m I\} > 0]. \qquad (19)\end{aligned}$$

Therefore,

$$\begin{aligned}\epsilon_{i,j} \quad = \quad P[\max_{I>0}\{&\sum_{k\in\mathcal{I}(m)} [B_k(I - D_{i,j} - \delta_{k,j_k}) + \\ &\lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_m I\} > 0]. \quad \square\end{aligned}$$

Now, we can bound the probability of flow-i traffic missing its local deadline before arriving at server $\pi(i,j) = m$.

Corollary 2 *The probability of flow i missing a local deadline before arriving at server m is bounded by*

$$\lambda_{i,m} \leq \sum_{h=1}^{j-1} \tilde{\epsilon}_{i,h}, \qquad (20)$$

where j is determined by $\pi(i,j) = m$ and

$$\begin{aligned}\tilde{\epsilon}_{i,h} \quad = \quad P[\max_{s>0}\{&\sum_{k\in\mathcal{I}(\pi(i,h))} [B_k(s - \delta_{k,j_k}) \\ &+\lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_{\pi(i,h)} s\} > 0],\end{aligned}$$

and j_k is defined by $\pi(k, j_k) = \pi(i,h)$.

Proof: According to Equation (15), the probability of traffic of flow i missing its local deadline at server $\pi(i,j)$ at time t is bounded by $\tilde{\epsilon}_{i,j} = P[d_{i,j}(t,s) > 0]$.

According to Corollary 1,

$$\begin{aligned}\tilde{\epsilon}_{i,h} \quad = \quad P[\max_{s>0}\{&\sum_{k\in\mathcal{I}(\pi(i,h))} [B_k(s - \delta_{k,j_k}) \\ &+\lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_{\pi(i,h)} s\} > 0],\end{aligned}$$

Since $\lambda_{k,1} = 0$, we can compute λ_{k,j_k} in the order of j_k. Therefore, the upper bound for local deadline violation probability of flow-i traffic before arriving at server m is determined by

$$\lambda_{i,m} = \sum_{h=1}^{j-1} \tilde{\epsilon}_{i,h}. \quad \square$$

Thus, applying this result, each flow can be guaranteed an end-to-end delay bound along with its violation probability by using Corollaries 1 and 2 to compose per-node quality-of-service parameters into end-to-end ones.

4 Numerical Investigations

In this section, we evaluate the performance of the CNS discipline by performing a set of admission control experiments. We compute the admissible regions (set of admissible flows) under a set of scenarios for CNS networks, and as a baseline for comparison, we also study the admissible regions of GPS networks [25]. The goal of the section is not to provide a comprehensive performance analysis of CNS under a broad range of realistic scenarios, but rather to illustrate the performance implications of *coordination* in providing multi-node services.

4.1 Scenario

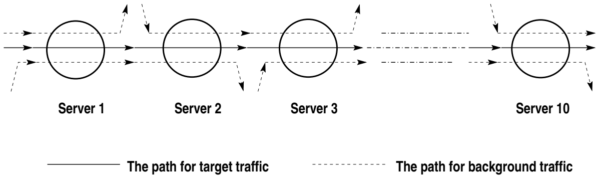


Figure 1: A Simple Tandem Network Topology.

We consider a simple tandem network as depicted in Figure 1. The network consists of 10 servers with the same capacity. There are M_0 flows entering the network from server 1 and exiting from server 10. These flows have the longest path and are chosen to be the target traffic for analysis. In addition, each multiplexer also serves two classes of cross traffic consisting of M_1 flows traversing one hop and M_2 flows traversing two hops as depicted. This cross traffic has the same characteristics as the target traffic (described below).

In order to simplify the evaluation, we assume that every flow has the same source traffic and is regulated by a dual leaky bucket before entering the network. If a source traffic of flow i is regulated by a dual leaky bucket with parameters (r_i, P_i, τ_i), then we can bound the statistical traffic envelope $B_i(I)$ of the source traffic as follows [20]: $E(B_i(I)) = r_i I$ and $Var(B_i(I)) \leq E(B_i(I)B_i(I)) - [E(B_i(I))]^2 \leq b_i(I) r_i I - r_i^2 I^2$, where

$$b_i(I) \quad = \quad \begin{cases} P_i I, & 0 \leq I \leq \frac{\tau_i}{P_i - r_i}, \\ \tau_i + r_i I, & \frac{\tau_i}{P_i - r_i} < I. \end{cases}$$

We study two classes of source traffic. The parameters for each class of source traffic are given in Table 2. Class 1's parameters represent on-off, periodic voice traffic whereas Class 2's parameters are similar to those of video traces.

	r (bps)	P (bps)	τ (b)
Class 1	32000	64000	23040
Class 2	150000	6000000	100000

Table 2: Parameters of Source Traffic.

4.2 GPS Admission Control

In a GPS network, the minimum bandwidth g_i^m guaranteed to flow i at server m is given by

$$g_i^m \quad = \quad \frac{r_i}{\sum_{k \in I(m)} r_k} C_m,$$

where r_i is the long term average rate of flow i, $I(m)$ is the set of flows that are served by server m. By simple extension of the results in [36], the probability of the end-to-end deadline D_i violation of the traffic of flow i can be bounded by

$$P[\max_{t > D_i} \{ \sum_{k \in \mathcal{S}} [B_k(t - D_i) - g_k t] \} > 0], \qquad (21)$$

where $g_k = \min_m \{g_k^m\}$ and $\mathcal{S}$ is the set of flows with the same source and destination as flow i.

4.3 Computing Performance Bounds

To compute $P[\max_{t>0}\{\sum_{k \in \mathcal{I}(m)} [B_k(t - D_{i,j} - \delta_{k,j_k}) + \lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_m t\} > 0\}]$, we use the maximum variance approach developed in [9]. Let

$$\begin{aligned} \sigma_t^2 &= \text{var}\{ \sum_{k \in I(m)} [B_k(t - D_{i,j} - \delta_{k,j_k}) \\ &\quad + \lambda_{k,j_k} B_k(\psi_{k,j_k})] - C_m t\}, \\ m_t &= E\{C_m t - \sum_{k \in \mathcal{I}(m)} [B_k(t - D_{i,j} - \delta_{k,j_k}) \\ &\quad + \lambda_{k,j_k} B_k(\psi_{k,j_k})]\}, \\ \alpha &= \inf_t \frac{m_t}{\sigma_t}. \end{aligned}$$

Approximating $\sum_{k\in\mathcal{I}(m)}[B_k(s - D_{i,j} - \delta_{k,j_k}) + \lambda_{k,j_k}B_k(\psi_{k,j_k})] - C_m t$ as Gaussian, the following upper bound can been obtained.

$$P[\max_{t>0}\{\sum_{k\in\mathcal{I}(m)}[B_k(t - D_{i,j} - \delta_{k,j_k})+ \lambda_{k,j_k}B_k(\psi_{k,j_k})] - C_m t\} > 0\}] < e^{-\frac{\alpha^2}{2}}.$$

A detailed comparative performance study of this bound can be found [21].

Throughout the experiments, we report the mean utilization of the network as the mean rate of admitted flows divided by the link capacity, averaged over all nodes. We also report the two key quality of service measures: the end-to-end delay bound and its corresponding violation probability. In this way, we characterize the service disciplines' ability to simultaneously achieve quality-of-service objectives and efficiency objectives.

4.4 Numerical Results and Analysis

For the CNS discipline, we use two different methods to assign the increment of priority indexes. The first method is for G-EDF as described in [8] and in Section 2. The priority index increment for each packet at each hop is a constant C (2 msec for voice traffic and 6 msec for video traffic) except at the first hop, where the increment of is equal to the end-to-end delay bound minus (path length - 1)*C. The second method is a simplified version of CNS which we refer to as (S-CNS). Here, the priority index increment at each hop is simply a constant (2 msec for voice traffic and 6 msec for video traffic).

We consider two scenarios, voice traffic and video traffic and explore the three key performance parameters, utilization, delay bound, and violation probability. Figure 2(a) depicts utilization versus end-to-end queueing delay for voice traffic with the required violation probability less than 10^{-4}. Figure 2(b) depicts the end-to-end delay bound violation probability versus the end-to-end queueing delay bound for voice traffic when the network utilization is 89%.

We make two observations regarding the figure. First, both instances of CNS disciplines outperform GPS. For the same network topology and the same traffic pattern, the CNS network can support more flows than the GPS network while supporting the same QoS level. For example, with the end-to-end delay bound less than 100 msec, CNS can support at least 90% utilization as opposed to at most 82% for GPS. Thus, while GPS achieves local fairness of bandwidth sharing at each node, CNS uses the coordination property to minimize end-to-end delay and achieve *global* fairness.

Notice further that for a fixed network utilization (89%), i.e., a fixed number of flows in the network, the end-to-end delay bound violation probability for target traffic in the CNS network is always smaller than that in the GPS network. For example, with 100 msec end-to-end delay bound and 89% utilization, the delay bound violation probability for target traffic is larger than 0.009 for GPS whereas it is less than 10^{-5} for CNS. This is again a consequence of the coordination property.

Finally, notice that among CNS disciplines, G-EDF parameter allocation outperforms S-CNS. The reason for this is that G-EDF assigns an additional portion of the end-to-end delay budget to the first hop. As a consequence, the possibility of many packets having very small priority indexes at core servers at the same time is reduced. Thus, there is less traffic distortion at core servers and fewer packets miss their end-to-end delay bounds.

Figure 3 reports the results of an analogous set of experiments for the video traffic. While the conclusions to be drawn are largely the same as for voice, we do note that the advantages of CNS over GPS are even more pronounced. The reason for this is that the higher burstiness of this traffic (larger variance and longer burst lengths) places a heavier burden on the scheduler during periods of overload. CNS is better suited to meeting end-to-end delay objectives under such high congestion periods.

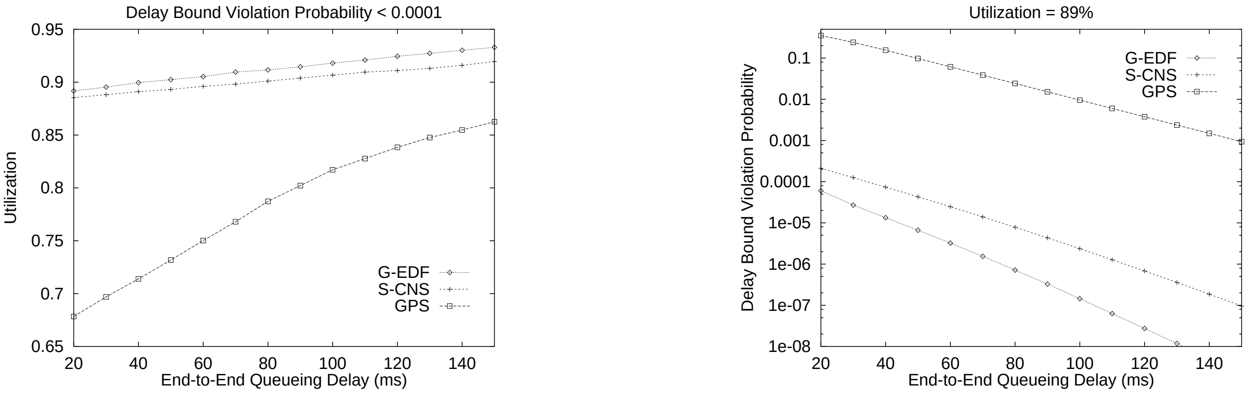
5 Conclusion

In this paper, we developed a framework for Coordinated Network Scheduling (CNS). With a definition of the fundamental coordination property, we showed how a number of schedulers from the literature can be characterized as CNS disciplines. We then developed a general theory based on traffic and service envelopes to analyze CNS networks and devised admission control tests for statistical end-to-end services. We showed that CNS disciplines limit traffic distortion to within a narrow range, thereby providing a foundation for efficient and scalable multi-node services. In admission control experiments with 10 hops, we found for example that utilization can be improved from 67% to 78% for highly bursty video flows requiring 150 msec end-to-end delay with violation probability 10^{-6}.

References

[1] M. Andrews. Probabilistic end-to-end delay bounds for earliest deadline first scheduling. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

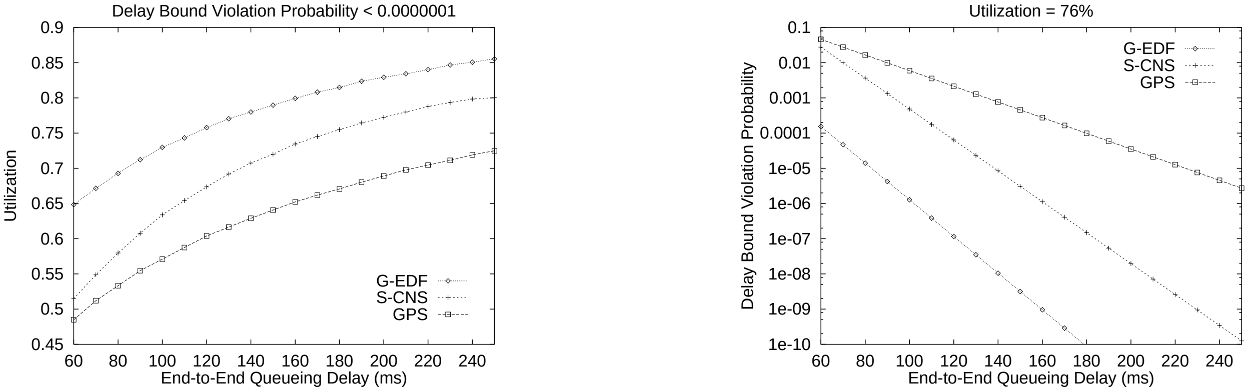
[2] M. Andrews and L. Zhang. Minimizing end-to-end delay in high-speed networks with a simple coordinated schedule. In *Proceedings of IEEE INFOCOM '99*, New York, NY, March 1999.



(a) Admissible region comparisons.

(b) End-to-end loss probability comparisons.

Figure 2: Periodic Voice Traffic (Server Capacity = 45 Mb/s).



(a) Admissible region comparisons.

(b) End-to-end loss probability comparisons.

Figure 3: Video Traffic (Server Capacity = 155 Mb/s).

[3] J. Bennett and H. Zhang. WF^{2}Q: Worst-case Fair Weighted Fair Queueing. In *Proceedings of IEEE INFOCOM '96*, San Francisco, CA, March 1996.

[4] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Effective envelopes: Statistical bounds on multiplexed traffic in packet networks. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[5] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Tradeoffs in networks with end-to-end statistical QoS guarantees. In *Proceedings of IWQoS 2000*, June 2000.

[6] Z. Cao, Z. Wang, and E. Zegura. Rainbow fair queueing: Fair bandwidth sharing without per-flow state. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[7] C. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[8] T. Chen, J. Walrand, and D. Messerschmitt. Dynamic priority protocols for packet voice. *IEEE Journal on Selected Areas in Communications*, 7(5):632–643, 1989.

[9] J. Choe and N. Shroff. A central limit theorem based approach to analyze queue behavior in ATM networks. *IEEE/ACM Transactions on Networking*, 6(5):659–671, October 1998.

[10] D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. In *Proceedings of ACM SIGCOMM '92*, pages 14–26, Baltimore, Maryland, August 1992.

[11] R. Cruz. A calculus for network delay, parts I and II. *IEEE Transactions on Information Theory*, 37(1):114–141, January 1991.

[12] G. de Veciana and G. Kesidis. Bandwidth allocation for multiple qualities of service using generalized processor sharing. *IEEE Transactions on Information Theory*, 42(1):268–272, January 1995.

[13] C. Dovrolis and P. Ramanathan. A case for relative differentiated services and the proportional differentiation model. *IEEE Network*, 13(5):26–35, September 1999.

[14] A. Elwalid, D. Mitra, and R. Wentworth. Design of generalized processor sharing schedulers which statistically mul-

tiplex heterogeneous QoS classes. In *Proceedings of IEEE INFOCOM '99*, March 1999.

[15] S. Floyd and V. Jacobson. Link-sharing and resource management models for packet network. *IEEE/ACM Transactions on Networking*, 3(4):365–386, August 1995.

[16] L. Georgiadis, R. Guérin, and V. Peris. The effect of traffic shaping in efficiently providing end-to-end performance guarantees. *IEEE/ACM Transactions on Networking*, 4(4), August 1996.

[17] S. Golestani. A stop-and-go queueing framework for congestion management. In *Proceedings of ACM SIGCOMM '90*, pages 8–18, Philadelphia, PA, September 1990.

[18] P. Goyal, S. Lam, and H. Vin. Determining end-to-end delay bounds for heterogeneous networks. In *Proceedings of IEEE Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'95)*, pages 287–298, Durham, NH, April 1995.

[19] P. Goyal and H. Vin. Statistical delay guarantee of virtual clock. In *Proceedings of IEEE Real-Time Systems Symposium*, December 1998.

[20] E. Knightly. Enforceable quality of service guarantees for bursty traffic streams. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[21] E. Knightly and N. Shroff. Admission control for statistical QoS: Theory and practice. *IEEE Network*, 13(2):20–29, March 1999.

[22] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proceedings of ACM SIGMETRICS '92*, pages 128–139, Newport, RI, June 1992.

[23] A. Mekkittikul and N. McKeown. A practical scheduling algorithm to achieve 100% throughput in input-queued switches. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[24] T. Nandagopal, N. Venkitaraman, R. Sivakumar, and V. Bharghavan. Relative delay differentation and delay class adaptation in core-stateless networks. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[25] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. *IEEE/ACM Transactions on Networking*, 2(2):137–150, April 1994.

[26] J. Qiu and E. Knightly. Inter-class resource sharing using statistical service envelopes. In *Proceedings of IEEE INFOCOM '99*, New York, NY, March 1999.

[27] M. Reisslein, K. Ross, and S. Rajagopa. Guaranteeing statistical QoS to regulated traffic: The multiple node case. In *Proceedings of IEEE Conference on Decision and Control*, pages 531–538, Tampa, FL, December 1998.

[28] M. Shreedhar and G. Varghese. Efficient fair queueing using deficit round-robin. *IEEE/ACM Transactions on Networking*, 4(3):375–385, June 1996.

[29] V. Sivaraman and F. Chiussi. Providing end-to-end statistical delay guarantees with earliest deadline first scheduling and per-hop traffic shaping. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[30] D. Stephens and H. Zhang. Implementing distributed packet fair queueing in a scalable switch architecture. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[31] I. Stoica, S. Shenker, and H. Zhang. Core-Stateless Fair Queueing: A scalable architecture to approximate fair bandwidth allocations in high speed networks. In *Proceedings of ACM SIGCOMM '98*, Vancouver, British Columbia, September 1998.

[32] I. Stoica and H. Zhang. Providing guaranteed services without per flow management. In *Proceedings of ACM SIGCOMM '99*, Cambridge, MA, August 1999.

[33] O. Yaron and M. Sidi. Performance and stability of communication networks via robust exponential bounds. *IEEE/ACM Transactions on Networking*, 1(3):372–385, June 1993.

[34] H. Zhang. Providing end-to-end performance guarantees using non-working-conserving disciplines. *Computer Communications: Special Issue on System Support for Multimedia Computing*, 18(10), October 1995.

[35] H. Zhang and E. Knightly. Providing end-to-end statistical performance guarantees with bounding interval dependent stochastic models. In *Proceedings of ACM SIGMETRICS '94*, pages 211–220, Nashville, TN, May 1994.

[36] Z. Zhang, D. Towsley, and J. Kurose. Statistical analysis of generalized processor sharing scheduling discipline. *IEEE Journal on Selected Areas in Communications*, 13(6):368–379, August 1995.