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

Chengzhi Li and Edward W. Knightly

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 Multihop Scheduling* (CMS) which exploit such inter-node coordination. We first provide a general CMS definition which enables us to classify a number of schedulers from the literature including, G-EDF, FIFO+, CEDF, and work-conserving CJVC as examples of CMS schedulers. We then develop a distributed theory of traffic envelopes which enables us to derive end-to-end statistical admission control conditions for CMS schedulers. We show that CMS schedulers are able to limit traffic distortion to within a narrow range resulting in improved end-to-end performance and more efficient resource utilization. Consequently, our technique exploits statistical resource sharing among flows, classes, and nodes, and our results provide the first statistical multi-node multi-class admission control algorithm for networks of work conserving servers.**

I. 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], [26]. Moreover, such performance properties are now achievable in high speed implementations [24], [30], [32] and scalable architectures in which core nodes do not maintain per-flow state [6], [25], [33].

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 [22]. For example, statistical class-based admission control tests have been derived for Earliest Deadline First (EDF) [4], [27], [31], Weighted Fair Queueing (WFQ) [12], [27], [38], Strict Priority [27], and Virtual Clock [20]. Moreover, techniques for providing multi-node or end-to-end statistical services have been developed for several classes of non-work-conserving schedulers [5], [28], [31], [37] and for Weighted Fair Queueing networks with isolation among flows [38].[1]

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 oldest 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 Multihop Schedulers (CMS). The key CMS 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 CMS class. We make several observations regarding coordinated multihop schedulers. (1) 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], [23], can be addressed. (2) CMS inter-server cooperation can improve a flow's end-to-end performance, and consequently, improve the efficiency and utilization of the network at large. (3) 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 [33].

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 multihop schedulers, thereby overcoming the traffic distortion problem and realizing the efficiency gains of coordination. To analyze CMS 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 inter-class theory of [27], we derive expressions for the essential traffic and service *envelopes*, which provide a general statistical characterization of a CMS node's workload and service capacity. Within this framework, we establish an important property of the CMS discipline, namely, that traffic distortion in CMS networks is limited to within a narrow range. Therefore, the essential traffic and service envelopes at a CMS 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 CMS 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], [17], [18], [28], [31], [36], [37]. 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

C. Li is with the University of Virginia. E. Knightly is with Rice University. This research was supported by NSF Grants ANI-0085842 and ANI-9733610, by a gift from Texas Instruments, and by a Sloan Fellowship. A subset of this work appears in the Proceedings of ICNP 2000.

[1] That is, a statistical multiplexing among flows is not considered.

bounding this distortion [7], [11], [23], [35] and exploiting isolation properties of GPS servers [14], [19], [26], [38]. 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 CMS networks. Our solution applies to the broad class of (work-conserving) CMS 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 CMS discipline and show how scheduling algorithms from the literature can be classified within the CMS 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 admission control results obtained by simulations and numerical analysis, and in Section 5 we conclude.

II. Framework for Coordinated Scheduling

In this section, we provide a formal definition of the CMS 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 III apply to a broad class of schedulers.

A. CMS Definition

Denote $d_{i,j}^k$ as the priority index assigned to the k^{th} packet of flow-i with size l_i^k at its j^{th} hop. Moreover, let t_i^k denote the time when the k^{th} packet of flow i arrives at its *first* hop. Finally, let $\delta_{i,j}^k$ denote the increment of the priority index of the k^{th} packet of flow i at its j^{th} hop.[2]

Definition 1 (Coordinated Multihop Scheduling) Consider a multiplexer which services packets in increasing order of their priority indexes. A scheduler possesses the CMS property if the priority index of packet k of flow i at its j^{th} hop can be expressed as

$$d_{i,j}^k = \begin{cases} t_{i,1}^k + \delta_{i,1}^k, & j = 1 \\ d_{i,j-1}^k + \delta_{i,j}^k, & j > 1 \end{cases} \tag{1}$$

where $\delta_{i,j}^k$ is a non-negative function of $i, j, l_i^k, t_{i,1}^k$, and $d_{i,1}^{k-1}$, and for $j \geq 2$, the priority increments satisfy $\delta_{i,j}^k \in [\delta_{i,j} - \eta_{i,j}, \delta_{i,j} + \eta_{i,j}]$, $\forall k \geq 1$, for some constants $\delta_{i,j}$ and $\eta_{i,j}$.

In other words, at the *first* node, the priority index is added to the packet's arrival time, and the index may be a constant, or a function of the packet's arrival time, the packet size, the priority index of the flow's previous packet, or constants associated with the flow and/or node. In contrast, at *downstream* nodes, the priority index is computed recursively as a function of the upstream index rather than by using the local arrival time. Moreover, while this downstream index can also be dynamic, it must be bounded within a range such that $\delta_{i,j}^k \in [\delta_{i,j} - \eta_{i,j}, \delta_{i,j} + \eta_{i,j}]$.

Observe that the requirement that $\delta_{i,j}^k$ is a function of i, j, l_i^k, $t_{i,1}^k$, and $d_{i,1}^{k-1}$ can be interpreted as meaning that the priority index of flow-i's k^{th} packet at its j^{th} hop can be determined when the packet first enters the network. Consequently, FIFO, EDF, WFQ, and Virtual Clock are not coordinated schedulers. For example, the priority index assigned by FIFO can be written as $d_{i,j}^k = t_{i,j}^k$, where $t_{i,j}^k$ is the arrival time of the k^{th} packet of flow i at its j^{th} hop. For $j > 1$, $d_{i,j}^k$ cannot be determined when the packet arrives at its first hop. Similarly, for Virtual Clock, the priority index can be written as $d_{i,j}^k = \max\{t_{i,j}^k, d_{i,j}^{k-1}\} + \frac{l_i^k}{r_i}$. However, for $j > 1$, $d_{i,j}^k$ depends on the arrival time of the k^{th} packet of flow i at its j^{th} hop, and cannot be determined when the packet arrives at its first hop.

Based on the selected method for assigning the increments of the priority index, we sub-classify CMS service disciplines into delay- and rate-CMS. A service discipline belongs to the *delay-CMS* 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.

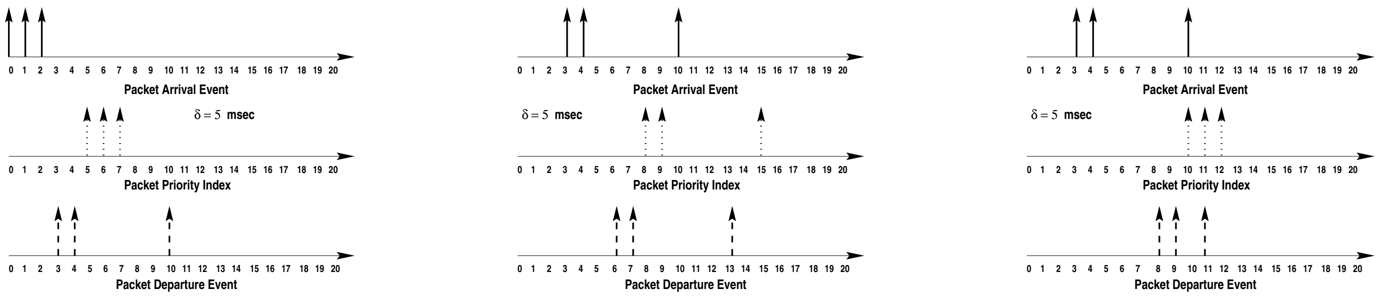
In contrast, a service discipline belongs to *rate-CMS* class if $\delta_{i,j}^k$ is a function of l_i^k and $r_{i,j}$, where l_i^k is the size of the k^{th} packet of flow i and $r_{i,j}$ is the reserved bandwidth for flow i at its j^{th} hop. The main characteristic of this class is that reserved bandwidths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

B. Discussion

The key property of the CMS 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.

To illustrate this property, consider the simple example of Figure 1 in which three packets of flow i arrive to the network at $t = 0, 1, 2$ respectively, and traverse two hops with $\delta_{i,1}^k = \delta_{i,2}^k = \delta = 5$. In the example, all packets have identical size, the link speed is 1 packet per time unit, and cross traffic exists at both hops. At the first hop, these three packets are assigned priority indexes (deadlines) of 5, 6, and 7 respectively, by both CMS and EDF. Suppose further that these three packets depart from the first hop at times 3, 4, and 10 respectively, so that the the third packet misses its local deadline by 3 time units due to cross traffic with higher priority. According to the arrival times at the second hop, these three packets are assigned priority indexes of 8, 9, and 15 by EDF, whereas the indexes are 10, 11, and 12 for CMS. In the example, with further cross traffic at the second hop, the third packet has higher priority in the CMS network than the EDF network, and therefore is able to meet both its local delay bound and global delay bound. In contrast, in the EDF network, the third packet meets its local delay bound at the second hop, but is not able to "catch up", and meet its end-to-end delay bound.

[2] Notation is summarized in Table 1.



(a) CMS and EDF at First Hop

(b) EDF at Second Hop

(c) CMS at Second Hop

Fig. 1. Illustration of Coordination

C. Example CMS Disciplines

The above definition of Coordinated Multihop Scheduling is quite general. Here, we show how several service disciplines from the literature, including G-EDF [8], FIFO+ [10], CJVC [34], and CEDF [1], [2] can be classified as instances of the CMS discipline.

C.1 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:

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

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

C.2 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:

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

where $\bar{d}_{i,j-1}$ is the average queueing delay of flow i seen by the packet at the upstream server. Provided that $\bar{d}_{i,j}$,$j \geq 1$, is determined before the packet departs from its first hop and that the range of $\bar{d}$ is bounded, comparing Equation (3) with Definition 1 shows that FIFO+ is also a delay-CMS discipline.

C.3 Work Conserving CJVC

Core-Jitter Virtual Clock (CJVC) was proposed in [34] 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:

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

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. Furthermore, it can be verified that $\frac{l_i^k}{r_i} + \eta_i^k \in [\delta_{i,j} - \eta_{i,j}, \delta_{i,j} + \eta_{i,j}]$ for $j > 1$, where $\delta_{i,j} = \frac{l_i^{max} + l_i^{min}}{2r_i}$ and $\eta_{i,j} = \frac{l_i^{max} - l_i^{min}}{2r_i}$. Thus, considering $\delta_{i,1}^k = \frac{l_i^k}{r_i} + \max(0, d_{i,j}^{k-1} - t_i^k)$, it is clear that work-conserving CJVC is a rate-CMS service discipline.

C.4 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-CMS disciplines in which the first priority index is randomized.

III. CMS Analysis and Admission Control

In this section, we develop a statistical multi-node analysis and admission control algorithm for CMS. 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 downstream 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 $f_{i,j}(t) = \sum_{k:t^k_{i,j}\le t} l^k_i$ as the total traffic in $[0,t]$ arriving from traffic flow i at its j^{th} hop, a node which is indexed by $\pi(i,j)$. 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 $\pi(i,j+1)$. Similar to [27], we call a sequence of non negative random variables $\{B_i(I)\}_{I=0}^{\infty}$ a statistical traffic envelope of flow i if $\forall t, I > 0$ [3]

$$A_i[t,t+I] = \sum_{k:d^k_{i,1}\in[t,t+I]} l^k_i \le_{st} B_i(I), \qquad (7)$$

and assume that $A_i[\cdot,\cdot]$ and $A_j[\cdot,\cdot]$ are independent and $B_i(\cdot)$ and $B_j(\cdot)$ are independent if $i \neq j$. Furthermore, we consider a discrete time model with an infinite buffer in which traffic is treated as fluid. We next review several facts about stochastic ordering that is used later in this section.

Lemma 1: Let $X_1,\cdots,X_n$ be independent and $Y_1,\cdots,Y_n$ be independent. If $X_i{\le_{st}}Y_i$ for $i=1,\cdots,n$, then
1. $\sum_{i=1}^n X_i{\le_{st}}\sum_{i=1}^n Y_i$.
2. $c-\sum_{i=1}^n X_i{\ge_{st}}c-\sum_{i=1}^n Y_i$ for any real number c.
3. There exist independent random variables $\overline{Y_1},\cdots,\overline{Y_n}$ such that $\overline{Y_i}$ has the same distribution as Y_i and $X_i \le \overline{Y_i}$ for $i = 1,\cdots,n$.

Proof: See [29]. □

A. Essential Traffic

Here, we define *essential traffic* as a building block 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.

Term	Definition
$\pi(i,j)$	j^{th} hop of flow i
N_i	path length of flow i
$t^k_{i,j}$	arrival time of the k^{th} packet of flow i at its j^{th} hop
l^k_i	flow-i k^{th} packet size
$\delta^k_{i,j}$	increment of priority index of the k^{th} packet of flow i at its j^{th} hop
$\delta_{i,j}$	mean value of $\delta^k_{i,j}$
$\eta_{i,j}$	range of $\delta^k_{i,j}$ variance
$f_{i,j}(t)$	total flow-i traffic at its j^{th} hop during $[0,t)$
$B_i(I)$	flow i statistical traffic envelope at its first hop
$A_i[s,t]$	total amount of flow-i traffic with priority index in $[s,t]$, i.e.,$\sum_{k:d^k_{i,1}\in[s,t]} l^k_i$
$\overline{B_i(I)}$	a random variable with the same distribution as $B_i(I)$
$f^*_{i,j}(t,s)$	flow-i traffic with local deadline no later than s arriving at server $\pi(i,j)$ during $[0,t)$
$\mathcal{E}_{i,\pi(i,j)}$	maximum tolerable local deadline violation of flow i at server $\pi(i,j)$
$\Gamma_{i,j}(I)$	flow i essential traffic envelope at its j^{th} hop
$Q_{i,j}(t,s)$	the amount of flow-i traffic with local deadline at server $\pi(i,j)$ no later than s is discarded before arriving at server $\pi(i,j)$ during $[0,t]$
$\overline{\Gamma_{i,j}(I)}$	a random variable with the same distribution as $\Gamma_{i,j}(I)$
$S(I,\alpha)$	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
$\overline{S(I,\alpha)}$	a random variable with the same distribution as $S(I,\alpha)$
$W^s_m(x)$	total traffic with local deadline no later than s queued at server m at time x
$D_{i,j}$	flow i delay bound 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
$\epsilon_{i,j}$	upper bound on probability $P\{d_{i,j}(t) > D_{i,j}\}$

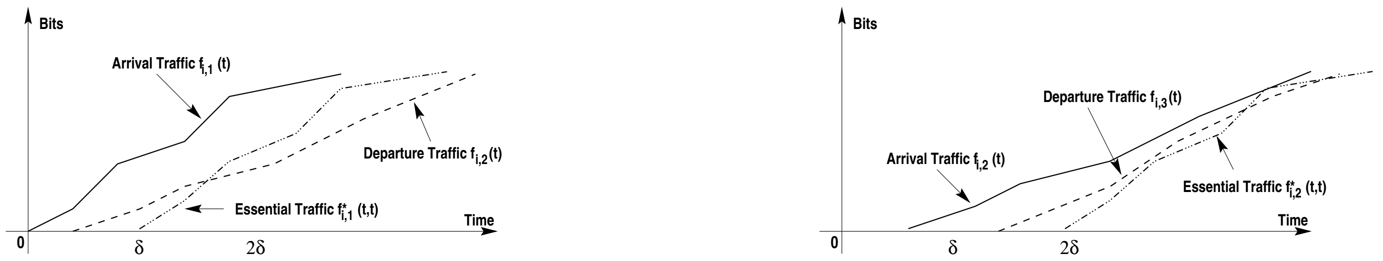
TABLE I

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

$$f^*_{i,j}(t,s) = \sum_{k:t^k_{i,j}\le t,\ d^k_{i,j}\le s} l^k_i. \qquad (8)$$

Figure 2 illustrates the relationship among arrival traffic, essential traffic, and departure traffic. If, for example, the priority index increment for flow i at each hop is δ, then flow-i's essential traffic $f^*_{i,1}(t,t)$ is simply $f_{i,1}(t-\delta)$ at node 1. Suppose further that some flow-i packets miss their local deadline at the first hop. Then, according to Definition 2, $f_{i,2}(t)$ will cross $f^*_{i,1}(t,t)$ as depicted in the figure. Next, ignoring propagation delay, the departure traffic of node 1 is the arrival traffic at node 2. Since

[3] $X{\le_{st}}Y$ (stochastic inequality) denotes $P[X>z]\le P[Y>z]$ for all z.



(a) First Hop

(b) Second Hop

Fig. 2. Example of Arrival, Essential, and Departure Traffic (priority index increment = δ)

at node 2, flow-i's arrival traffic has not missed its node-2 local deadline upon arriving, $f^*_{i,2}(t,t) = f_{i,1}(t-2\delta)$. The key point for Figure 2(b), is that the relationship between the arrival traffic $f_{i,2}(t)$ and the essential traffic $f^*_{i,2}(t,t)$ characterizes the advantage of CMS's coordination mechanism. In particular, the horizontal distance between $f_{i,2}(t)$ and $f^*_{i,2}(t,t)$ corresponds with the queueing delay incurred at node 1: the larger the queueing delay, the shorter the horizontal distance, and the higher the resulting priority index. Consequently, excessively delayed packets at upstream nodes can "catch up" at downstream nodes and still satisfy their end-to-end deadline requirement.

B. Essential Traffic Envelope

In each server of CMS networks, packets are served in increasing order of their priority indexes (local deadlines). For a given s, flow-i's essential traffic $f^*_{i,j}(t,s)$ must be characterized to compute the local deadline s violation probability. Furthermore, $f^*_{i,j}(s+\Delta,s) - f^*_{i,j}(t,s)$ affects the probability of violating the local deadline by no more than Δ for a packet with local deadline s arriving at server $\pi(i,j)$ during $[t, s+\Delta]$. Thus, we define the essential traffic envelope as follows.

Definition 3 (Essential Traffic Envelope) A sequence of non-negative random variables $\{\Gamma_{i,j}(I)\}^{\infty}_{I=-\infty}$ is an essential traffic envelope of flow i at its j^{th} hop if $\forall s, t > 0$ and $\forall \Delta$ such that $s + \triangle \geq t$,

$$f^*_{i,j}(s+\Delta,s) - f^*_{i,j}(t,s) \quad \leq_{st} \quad \Gamma_{i,j}(s-t). \tag{9}$$

A key challenge for provisioning multi-node services is characterizing the traffic at downstream servers. The difficulty is due to the fact that a flow's traffic is unavoidably distorted after multiplexing with other flows. Furthermore, the traffic distortion may be accumulated along the path of a flow in networks without coordinated scheduling disciplines. However, due to the coordination property, the accumulated distortion phenomenon of the *essential traffic* is mitigated. In networks with coordinated scheduling, the distortion of a flow's essential traffic after passing through a server depends only on the local deadline violation. If the flow's traffic does not violate its local deadlines, the distortion of the essential traffic is eliminated, even if the flow's traffic incurs a queueing delay. The following theorem precisely characterizes this advantage of the coordination property.

Let $\mathcal{E}_{i,\pi(i,j)}$ denote the maximum tolerable local deadline violation of flow-i traffic at server $\pi(i,j)$. That is, a flow-i packet will be discarded if it misses its local deadline at server $\pi(i,j)$ by more than $\mathcal{E}_{i,\pi(i,j)}$. Let $Q_{i,j}(t,s)$ denote the amount of flow-i traffic with local deadline at server $\pi(i,j)$ no later than s that is discarded during $[0,t]$ before arriving at server $\pi(i,j)$.

Theorem 1: An essential traffic envelope $\Gamma_{i,j}(I)$ of flow i at its j^{th} hop is given by

$$\Gamma_{i,j}(I) = B_i(I - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)}), \tag{10}$$

where $T_{i,j} = \eta_{i,j} + 2\sum_{h=2}^{j-1} \eta_{i,h}$.

Proof: To relate a flow's downstream essential traffic to its original arrival envelope, we statistically upper bound $f^*_{i,j}(s+\Delta,s)$ and lower bound $f^*_{i,j}(t,s)$.

Let $\pi(i,j) = m$. For all $t, s > 0$ and Δ such that $s+\Delta \geq t$, consider the interval $[t, s+\Delta]$. Without loss of generality, assume that there is at least one flow-i packet with local deadline no later than s arriving at server m during $[t, s+\Delta]$. Otherwise, $f^*_{i,j}(s+\Delta,s) - f^*_{i,j}(t,s) = 0$, a trivial case. Since for a given flow i, packets are always serviced in increasing order of their local deadlines, all flow-i packets arriving at server m before time t have local deadlines no later than s. According to Definition 2 and the definition of $Q_{i,j}(s+\Delta,s)$, we have

$$f^*_{i,j}(s+\Delta,s) + Q_{i,j}(s+\Delta,s) \leq \sum_{k:d^k_{i,j}\leq s} l^k_i. \tag{11}$$

Since $d^k_{i,j} = d^k_{i,1} + \sum_{h=2}^{j}\delta^k_{i,h}$ and $\delta^k_{i,h} \in [\delta_{i,h} - \eta_{i,j}, \delta_{i,h} + \eta_{i,j}]$, we have

$$\begin{aligned} & f^*_{i,j}(s+\Delta,s) + Q_{i,j}(s+\Delta,s) \\ \leq & \sum_{k:d^k_{i,1}+\sum_{h=2}^{j}\delta^k_{i,h}\leq s} l^k_i \\ = & \sum_{k:d^k_{i,1}\leq s-\sum_{h=2}^{j}\delta^k_{i,h}} l^k_i \\ = & A_i[0, s - \sum_{h=2}^{j}\delta^k_{i,h}] \\ \leq & A_i[0, s - \sum_{h=2}^{j}\delta_{i,h} - \eta_{i,h}]. \end{aligned} \tag{12}$$

Next, we lower bound $f^*_{i,j}(t,s)$ as follows:

$$f^*_{i,j}(t,s) + Q_{i,j}(t,s) \quad \geq \quad \sum_{k:d^k_{i,j-1}\leq t-\mathcal{E}_{i,\pi(i,j-1)}} l^k_i, \tag{13}$$

since at time t, all flow-i traffic with local deadline at server $\pi(i,j-1)$ no later than $t-\mathcal{E}_{i,\pi(i,j-1)}$ either has departed server $\pi(i,j-1)$ and arrived server $\pi(i,j)$ or has been discarded due to the maximum tolerable local deadline violation $\mathcal{E}_{i,\pi(i,j-1)}$ at server $\pi(i,j-1)$ for flow-i traffic. Thus, similar to Equation (12), we have

$$\begin{aligned} & f^*_{i,j}(t,s) + Q_{i,j}(t,s) \\ & \qquad \geq \quad A_i\big[0, t - \sum_{h=2}^{j-1} \delta^k_{i,h} - \mathcal{E}_{i,\pi(i,j-1)}\big] \\ & \qquad \geq \quad A_i\big[0, t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}\big]. \end{aligned} \tag{14}$$

Since $s+\Delta \geq t$, we have $Q_{i,j}(s+\Delta, s) \geq Q_{i,j}(t,s)$. Thus, according to Equations (12) and (14),

$$\begin{aligned} & f^*_{i,j}(s+\Delta,s) - f^*_{i,j}(t,s) \\ & \quad \leq \quad f^*_{i,j}(s+\Delta,s) + Q_{i,j}(s+\Delta,s) - f^*_{i,j}(t,s) \\ & \qquad\qquad\qquad - Q_{i,j}(t,s) \\ & \quad \leq \quad A_i\big[0, s - \sum_{h=2}^{j} (\delta_{i,h} - \eta_{i,h})\big] \\ & \qquad\qquad - A_i\big[0, t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}\big] \\ & \quad \leq \quad A_i\Big[t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}, \\ & \qquad\qquad\qquad s - \sum_{h=2}^{j} (\delta_{i,h} - \eta_{i,h})\Big]. \end{aligned}$$

Notice that the interval $\Big[t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}, s - \sum_{h=2}^{j} (\delta_{i,h} - \eta_{i,h})\Big]$ has duration

$$\begin{aligned} & \big[s - \sum_{h=2}^{j} (\delta_{i,h} - \eta_{i,h})\big] - \big[t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}\big] \\ & \quad = \quad s - t - \delta_{i,j} + \eta_{i,j} + 2\sum_{h=2}^{j-1} \eta_{i,h} + \mathcal{E}_{i,\pi(i,j-1)} \\ & \quad = \quad s - t - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)}. \end{aligned}$$

According to Equation (7),

$$\begin{aligned} & A_i\big[t - \sum_{h=2}^{j-1} (\delta_{i,h} + \eta_{i,h}) - \mathcal{E}_{i,\pi(i,j-1)}, s - \sum_{h=2}^{j} (\delta_{i,h} - \eta_{i,h})\big] \\ & \qquad \leq_{st} \quad B_i(s - t - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)}). \end{aligned}$$

Thus, we have

$$\begin{aligned} & f^*_{i,j}(s+\triangle,s) - f^*_{i,j}(t,s) \\ & \qquad\qquad \leq_{st} \quad B_i(s - t - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)}). \end{aligned}$$

That is, by Definition 3,

$$\Gamma_{i,j}(I) = B_i(I - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)}). \quad \square$$

This theorem describes an important properties of the CMS discipline, namely, that distortion of *essential* traffic downstream is limited to within a narrow range. To illustrate, consider the special case in which $\delta^k_{i,j} = \delta_i$ for all k so that $T_{i,j} = 0$ for $j = 1, \cdots, N_i$. In this case, for any flow i, its essential traffic envelope at downstream servers, namely, $\Gamma_{i,j}(I) = B_i(I - \delta_i + \mathcal{E}_{i,\pi(i,j-1)})$, is affected by the maximum tolerable local deadline violation $\mathcal{E}_{i,\pi(i,j-1)}$. Furthermore, if δ_i is chosen appropriately such that flow-i packets do not miss their local deadlines or flow-i packets that miss local deadlines are discarded, i.e., $\mathcal{E}_{i,\pi(i,j)} = 0$, flow-$i$'s essential traffic envelope at downstream servers is identical to that at the ingress server, namely, $\Gamma_{i,j}(I) = B_i(I - \delta_i)$. Finally, for two different flows i_1 and i_2, according to Equation (10), $\Gamma_{i_1,j_1}(\cdot)$ and $\Gamma_{i_2,j_2}(\cdot)$ are independent if $B_{i_1}(\cdot)$ and $B_{i_2}(\cdot)$ are independent. Thus, the property of independence between essential traffic envelopes at the network edge is preserved downstream.

C. Essential Service Envelope

The above result enables us to derive end-to-end admission control tests for CMS 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 [27]. Thus, we use statistical service envelopes as a tool for characterizing and controlling inter-class resource sharing.

Definition 4 (Essential Service Envelope) A sequence of nonnegative random variables $\{S_{i,j}(I,\Delta)\}_{I=0}^{\infty}$ is called a (statistical) essential service envelope provided by server $\pi(i,j)$ to the traffic of flow i, if $\forall t, s > 0$ and $\forall \Delta$ such that $s+\Delta \geq t$, the minimum available service, denoted as $Y_{i,j}[t, s+\Delta, s]$, from server $\pi(i,j)$ to flow-i traffic with local deadline at server $\pi(i,j)$ no later than s and arriving at server $\pi(i,j)$ during $[t, s+\Delta]$ is lower bound by

$$Y_{i,j}[t, s+\triangle, s] \quad \geq_{st} \quad S_{i,j}(s+\Delta-t, \Delta), \tag{15}$$

provided that during $[t, s+\Delta]$, server $\pi(i,j)$ only services the traffic arriving during $[t, s+\Delta]$.

Roughly, $S_{i,j}(I,\Delta)$ is a measure of the service provided in an interval with length I to flow-i traffic with local deadline Δ seconds before the end of the interval. Notice that the minimum available service for flow i from server $\pi(i,j)$ during an interval 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 a given $i^* \in \mathcal{I}(m)$, $\forall\Delta$ and $\forall I > 0$,

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

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

Proof: Since the amount of total traffic (except flow i^*) with local deadlines no later than s and arriving at server m during $[t, s+\Delta]$ is $\sum_{i\in\mathcal{I}(m),i\neq i^*}[f^*_{i,j_k}(s+\Delta,s) - f^*_{i,j_k}(t,s)]$ and the total service capacity of server m during the same time interval is $C_m(s+\Delta-t)$, we have that

$$Y_{i^*,j}[t, s+\Delta, s] \geq C_m(s+\Delta-t) - \sum_{i\in\mathcal{I}(m),i\neq i^*}\Big[f^*_{i,j_k}(s+\Delta,s) - f^*_{i,j_k}(t,s)\Big]. \quad (16)$$

$\forall i \in \mathcal{I}(m)$, according to Theorem 1, the total flow-i traffic with local deadlines no later than s arriving at server m during $[t, s+\triangle]$ is bound by

$$\begin{aligned} & f^*_{i,j_k}(s+\Delta,s) - f^*_{i,j_k}(t,s) \\ \leq \quad & A_i\Big[t - \sum_{h=2}^{j_k-1}(\delta_{i,h}+\eta_{i,h}) - \mathcal{E}_{i,\pi(i,j_k-1)}, \\ & \qquad s - \sum_{h=2}^{j_k}(\delta_{i,h}-\eta_{i,h})\Big] \\ \leq_{st} \quad & B_i(s - t - \delta_{i,j_k} + T_{i,j_k} + \mathcal{E}_{i,\pi(i,j_k-1)}) \\ = \quad & \Gamma_{i,j_k}(s-t). \end{aligned}$$

Since $A_i[t-\sum_{h=2}^{j_k-1}(\delta_{i,h}+\eta_{i,h})-\mathcal{E}_{i,\pi(i,j_k-1)}, s-\sum_{h=2}^{j_k}(\delta_{i,h}-\eta_{i,h})]$, $i\in\mathcal{I}(m)$, are independent and $B_i(s-t-\delta_{i,j_k}+T_{i,j_k}+\mathcal{E}_{i,\pi(i,j_k-1)})$,$i\in\mathcal{I}(m)$, are independent, according to Lemma 1, we have that

$$\begin{aligned} & \sum_{i\in\mathcal{I}(m),i\neq i^*}\Big[f^*_{i,j_k}(s+\Delta,s) - f^*_{i,j_k}(t,s)\Big] \\ \leq \quad & \sum_{i\in\mathcal{I}(m),i\neq i^*} A_i\Big[t - \sum_{h=2}^{j_k-1}(\delta_{i,h}+\eta_{i,h}) - \mathcal{E}_{i,\pi(i,j_k-1)}, \\ & \qquad s - \sum_{h=2}^{j_k}(\delta_{i,h}-\eta_{i,h})\Big] \\ \leq_{st} \quad & \sum_{i\in\mathcal{I}(m),i\neq i^*}\Gamma_{i,j_k}(s-t), \end{aligned}$$

and so

$$\begin{aligned} & Y_{i^*,j}[t, s+\Delta, s] \\ \geq \quad & C_m(s+\Delta-t) \\ & \quad - \sum_{i\in\mathcal{I}(m),i\neq i^*}\Big[f^*_{i,j_k}(s+\Delta,s) - f^*_{i,j_k}(t,s)\Big] \\ \geq \quad & C_m(s+\Delta-t) \\ & - \sum_{i\in\mathcal{I}(m),i\neq i^*} A_i\Big[t - \sum_{h=2}^{j_k-1}(\delta_{i,h}+\eta_{i,h}) \\ & \quad - \mathcal{E}_{i,\pi(i,j_k-1)}, s - \sum_{h=2}^{j_k}(\delta_{i,h}-\eta_{i,h})\Big] \\ \geq_{st} \quad & C_m(s+\Delta-t) - \sum_{i\in\mathcal{I}(m),i\neq i^*}\Gamma_{i,j_k}(s-t). \quad (17) \end{aligned}$$

That is, by Definition 4,

$$S_{i^*,j}(I,\Delta) \;=\; C_m I - \sum_{i\in\mathcal{I}(m),i\neq i^*}\Gamma_{i,j_k}(I-\Delta). \quad \square$$

According to this theorem and Lemma 1, for a given t, s, and Δ, there exists a random variable $\overline{S_{i,j}(s+\Delta-t,\Delta)}$ with the same distribution as $S_{i,j}(s+\Delta-t,\Delta)$ such that

$$Y_{i,j}[t, s+\Delta, s, s] \quad \geq \quad \overline{S_{i,j}(s+\Delta-t,\Delta)}. \quad (18)$$

Furthermore, according to Equation (17), (7), and (10), we have that

$$\begin{aligned} & Y_{i,j}[t, s+\Delta, s] \\ & \quad \geq \quad C_m(s+\Delta-t) - \sum_{k\in\mathcal{I}(m),k\neq i}\overline{\Gamma_{k,j_k}(s-t)}, \quad (19) \end{aligned}$$

where the random variable $\overline{\Gamma_{k,j_k}(s-t)}$ has the same distribution as $\Gamma_{k,j_k}(s-t)$ and $f^*_{k,j_k}(s+\Delta,s) - f^*_{k,j_k}(t,s) \leq \overline{\Gamma_{k,j_k}(s-t)}$. Finally, according to Equation (10) and $B_k(\cdot)$, $k\in\mathcal{I}(m)$, are independent, $\Gamma_{k,j_k}(s-t)$, $k\in\mathcal{I}(m)$, are independent. Furthermore, $A_k[\cdot,\cdot]$, $k\in\mathcal{I}(m)$, are independent, and from Lemma 1, we have that $\overline{\Gamma_{k,k_k}(s-t)}$, $k\in\mathcal{I}(m)$, are independent. Since $C_m(s+\Delta-t)-\sum_{k\in\mathcal{I}(m),k\neq i}\overline{\Gamma_{k,j_k}(s-t)}$ has the same distribution as $S_{i,j}(s+\Delta-t,\Delta)$,

$$\begin{aligned} & \overline{S_{i,j}(s+\Delta-t,\Delta)} \\ & \quad = \quad C_m(s+\Delta-t) - \sum_{k\in\mathcal{I}(m),k\neq i}\overline{\Gamma_{k,j_k}(s-t)}. \quad (20) \end{aligned}$$

D. Admission Control

We now derive an end-to-end admission control condition for CMS networks. The technique used to analyze the deadline-violation probability is to analyze per-server local delay-bound-violation probabilities and then compose them into end-to-end ones. When computing the local-deadline-violation probability at a given server, any arrival packet at the server is not considered as discarded even if it incurs a long queueing delay, i.e., packet discarding only occurs at the upstream servers. For a multiplexer (server) in a CMS 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 s violation. We refer to this instant as the *void time* denoted as $\tau_m(t,s)$ and precisely define it as

$$\tau_m(t,s) = \max\{x \mid x \leq t \text{ and } W^s_m(x) = 0\}, \quad (21)$$

where $W^s_m(x)$ is the total amount of traffic with local deadline no later than s backlogged at server m at time x.[4] Notice that server m is not necessarily idle at time $\tau_m(t,s)$ as it may be busy serving traffic with local deadlines later than s. We use a concept of (virtual) delay due to the essential traffic at a particular node to derive the *local* deadline-violation probability as an intermediary step towards bounding the *end-to-end* deadline violation 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

$$d_{i,j}(t,s) = \min\{\triangle : f^*_{i,j}(t,s) \leq f_{i,j+1}(s+\triangle)\}. \quad (22)$$

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

Observe that $f_{i,j+1}(s+\Delta)$ is the amount of flow-i traffic departing from server $\pi(i,j)$ during $[0, s+\Delta]$. If $f_{i,j+1}(s+\Delta) > f^*_{i,j}(t,s)$, all flow-$i$ traffic with local deadline no later than s at server m and arriving at server m before time t has departed server m before time $s + \triangle$. For one 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.

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

$$\begin{aligned} P\big[d_{i,j}(t,s) > D_{i,j}\big] \\ \leq \quad & P\Big[\max_{I>0}\{\overline{\Gamma_{i,j}(I)} - \overline{S_{i,j}(I+D_{i,j},D_{i,j})}\}> 0\Big], \quad (23)\end{aligned}$$

for $D_{i,j} \geq t - s$, where $\overline{\Gamma_{i,j}(I)}$ and $\overline{S_{i,j}(I+D_{i,j},D_{i,j})}$ are random variables with the same distribution as $\Gamma_{i,j}(I)$ and $S_{i,j}(I+D_{i,j},D_{i,j})$ respectively.

Proof: From Equation (22), we have

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

Since $D_{i,j} \geq t - s$, $s + D_{i,j} \geq t$. Thus, $f^*_{i,j}(s+D_{i,j},s) \geq f^*_{i,j}(t,s)$, so that [5]

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

If $d_{i,j}(t,s) > D_{i,j}$, there always exist packets with deadlines no later than s at server m during $[\tau_m(t,s), s+D_{i,j}]$. Since at $\tau_m(t,s)$ there is not traffic with deadline no later than s at server m, at least $f^*_{i,j}(\tau_m(t,s),s)$ amount of flow-i traffic has been served. Furthermore, during $[\tau_m(t,s), s+D_{i,j}]$, server m only serves the traffic with local deadline no later than s and arriving at server m after $\tau_m(t,s)$. Similar to Equation (16), we have

$$\begin{aligned} &Y_{i,j}[\tau_m(t,s), s+D_{i,j}, s] \\ &\quad \geq \quad C_m\big(s+D_{i,j} - \tau_m(t,s)\big) \\ &\qquad - \sum_{k\in\mathcal{I}(m),k\neq i} \big[f^*_{k,j_k}(s+D_{i,j},s) - f^*_{k,j_k}(\tau_m(t,s),s)\big].\end{aligned}$$

Thus,

$$f_{i,j+1}(s+D_{i,j}) \geq Y_{i,j}[\tau_m(t,s), s+D_{i,j}, s] + f^*_{i,j}(\tau_m(t,s),s),$$

and so

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

Since the random variable $\tau_m(t,s) \in [0,t]$, we have

$$\begin{aligned} &\Big\{f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(\tau_m(t,s),s) \\ &\qquad - Y_{i,j}[\tau_m(t,s), s+D_{i,j}, s] > 0\Big\} \\ &\subseteq \Big\{\max_{x\in[0,t]}\{f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(x,s) \\ &\qquad - Y_{i,j}[x, s+D_{i,j}, s]\}> 0\Big\}.\end{aligned}$$

Notice that for a real number x, from Theorem 1 and Theorem 2, $[f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(x,s)] \leq_{st} \Gamma_{i,j}(s-x)$ and $Y_{i,j}[x, s+D_{i,j}, s] \geq_{st} S_{i,j}(s+D_{i,j}-x, D_{i,j})$. According to Lemma 1, we can find random variables $\overline{\Gamma_{i,j}(s-x)}$ and $\overline{S_{i,j}(s+D_{i,j}-x,D_{i,j})}$ with the same distribution as $\Gamma_{i,j}(s-x)$ and $S_{i,j}(s+D_{i,j}-x,D_{i,j})$ respectively such that $[f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(x,s)] \leq \overline{\Gamma_{i,j}(s-x)}$ and $Y_{i,j}[x, s+D_{i,j}, s] \geq \overline{S_{i,j}(s+D_{i,j}-x,D_{i,j})}$. Thus, we have

$$\begin{aligned} &\Big\{\max_{x\in[0,t]}\{f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(x,s) \\ &\qquad - Y_{i,j}[x, s+D_{i,j}, s]\}> 0\Big\} \\ &\subseteq \Big\{\max_{x\in[0,t]}\{\overline{\Gamma_{i,j}(s-x)} \\ &\qquad - \overline{S_{i,j}(s+D_{i,j}-x,D_{i,j})}\}> 0\Big\}.\end{aligned}$$

Therefore, we have

$$\begin{aligned} &P[d_{i,j}(t,s) > D_{i,j}] \\ &\quad = \quad P[f^*_{i,j}(t,s) - f_{i,j+1}(s+D_{i,j}) > 0] \\ &\quad \leq \quad P[f^*_{i,j}(s+D_{i,j},s) - f_{i,j+1}(s+D_{i,j}) > 0] \\ &\quad \leq \quad P\Big[f^*_{i,j}(s+D_{i,j},s) - f^*_{i,j}(\tau_m(t,s),s) \\ &\qquad\qquad - Y_{i,j}[\tau_m(t,s), s+D_{i,j}, s] > 0\Big] \\ &\quad \leq \quad P\Big[\max_{x\in[0,t]}\{f^*_{i,j}(t,s) - f^*_{i,j}(x,s) \\ &\qquad\qquad - Y_{i,j}[x, s+D_{i,j}, s]\}> 0\Big] \\ &\quad \leq \quad P\Big[\max_{x\in[0,t]}\{\overline{\Gamma_{i,j}(s-x)} \\ &\qquad\qquad - \overline{S_{i,j}(s+D_{i,j}-x,D_{i,j})}\}> 0\Big] \\ &\quad \leq \quad P\Big[\max_{I\geq 0}\{\overline{\Gamma_{i,j}(I)} - \overline{S_{i,j}(I+D_{i,j},D_{i,j})}\}> 0\Big]. \quad \square\end{aligned}$$

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

Corollary 1: For $t, s > 0$,

$$P[d_{i,j}(t,s) > D_{i,j}] \leq \epsilon_{i,j}, \qquad (24)$$

where

$$\begin{aligned} \epsilon_{i,j} = P\Big[\max_{I>0}\{ &\sum_{k\in\mathcal{I}(m)} \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})} \\ &\qquad - C_m(I+D_{i,j})\}> 0\Big],\end{aligned}$$

[5] When computing the probability of local deadline violation for a flow-i packet with local deadline s arriving at time t, $f^*_{i,j}(s+D_{i,j},s) = f^*_{i,j}(t,s)$, because the local deadline of any flow-i packet arriving after time t is larger than s, so that $\{f^*_{i,j}(t,s) - f_{i,j+1}(s+D_{i,j}) > 0\} \equiv \{f^*_{i,j}(s+D_{i,j},s) - f_{i,j+1}(s+D_{i,j}) > 0\}$.

and m and j_k are defined by $\pi(k, j_k) = \pi(i, j) = m$.
Proof: From Theorem 3, we have

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

According to Theorem 1,

$$\overline{\Gamma_{i,j}(I)} = \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})},$$

and according to Equation (20),

$$\begin{aligned} & \overline{S_{i,j}(I + D_{i,j}, D_{i,j})} = C_m(I + D_{i,j}) \\ & \quad - \sum_{k \in \mathcal{I}(m), k \neq i} \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})}. \quad (25) \end{aligned}$$

Therefore,

$$\begin{aligned} & P[d_{i,j}(t,s) > D_{i,j}] \\ & \quad \leq \quad P\Big[\max_{I>0}\{\overline{\Gamma_{i,j}(I)} - \overline{S_{i,j}(I + D_{i,j}, D_{i,j})}\} > 0\Big] \\ & \quad = \quad P\Big[\max_{I>0}\Big\{\overline{B_i(I - \delta_{i,j} + T_{i,j} + \mathcal{E}_{i,\pi(i,j-1)})} \\ & \qquad\qquad - \big[C_m(I + D_{i,j}) - \\ & \qquad \sum_{k \in \mathcal{I}(m), k \neq i} \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})}\big]\Big\} > 0\Big] \\ & \quad = \quad P\Big[\max_{I>0}\Big\{\sum_{k \in \mathcal{I}(m)} \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})} \\ & \qquad\qquad\qquad\qquad - C_m(I + D_{i,j})\Big\} > 0\Big]. \end{aligned}$$

That is,

$$\begin{aligned} \epsilon_{i,j} = P\Big[\max_{I>0}\Big\{ & \sum_{k \in \mathcal{I}(m)} \overline{B_k(I - \delta_{k,j_k} + T_{k,j_k} + \mathcal{E}_{k,\pi(k,j_k-1)})} \\ & \qquad - C_m(I + D_{i,j})\Big\} > 0\Big]. \quad \square \end{aligned}$$

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

IV. Simulation and Admission Control Experiments

In this section, we study the performance of the CMS discipline by performing a set of *ns-2* simulation and admission control experiments. Our goal is twofold. First, we establish the effectiveness of our admission control algorithm in properly controlling the number of admitted flows in CMS networks. To achieve this, we first perform a large set of simulation experiments, in which a fixed number of flows are established over various network paths (as described below). For each scenario, we perform numerous simulations and record average performance measures such as a end-to-end delay bound violation probability. We then run a set of *admission control* experiments in which we use an implementation of our admission control to determine the maximum number of admissible flows under a certain performance criteria. The results of these experiments yield experimental and predicted admissible regions, i.e., the true admissible region obtained by simulations and those obtained by our admission control algorithm.

Our second set of experiments explore the end-to-end performance of CMS networks as compared with non-CMS networks. In particular, we consider networks of FIFO, EDF, and WFQ schedulers and investigate the fraction of packets violating end-to-end delay targets under the different schedulers. These experiments illustrate the potential QoS improvements of coordinated network scheduling, that is, its ability to improve end-to-end performance under a particular network load, or conversely to improve the admissible region under a particular QoS requirement.

A. Scenario

We consider a simple tandem network topology as depicted in Figure 3. All link rates are 10 Mb/sec, packet lengths are 100 bytes, and propagation delays are 0 msec. There are M_0 flows entering the network from the first server and exiting from the last server. These flows have the longest path and are chosen to be the target class for analysis. In addition, each router also serves two classes of cross traffic consisting of M_1 flows which traverse a single router and then exit the network, and M_2 flows that traverse two routers and then exit the network. The cross traffic has the same characteristics as the target traffic (described below) and comprises approximately 80% of the total traffic.

We simulate exponential on-off flows with on-rate 64 kb/sec, mean on time 312 msec and mean off time 325 msec. For the CMS discipline, we choose

$$\eta_{i,j} = 0, \quad \delta_{i,j}^k = \delta_{i,j} = D,$$

for cross traffic flows with a 1 hop path;

$$\eta_{i,j} = 0, \quad \delta_{i,j}^k = \delta_{i,j} = \frac{D}{2},$$

for cross traffic flows with a 2 hop path;

$$\eta_{i,j} = 0, \quad \delta_{i,j}^k = \delta_{i,j} = \frac{D}{6},$$

for target traffic flows with a 6 hop path;

where D is the expected end-to-end queueing delay bound. In this case, $\sum_{k:d_{i,1}^k \in [t,t+I]} l_i^k \leq_{st} B_i(I), \forall t > 0$ is equivalent to $\sum_{k:t_{i,1}^k \in [t,t+I]} l_i^k \leq_{st} B_i(I), \forall t > 0$. Using [21], and the flows' mean rate, peak rate, and mean burst length given by the tuple (r_i, P_i, τ_i), we approximate the statistical traffic envelope $B_i(I)$ as $E(B_i(I)) = r_i I$ 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

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

Furthmore, when predicting the admissible region for CMS networks, we assume that a packet will be discarded if it suffers a queueing delay at a server more than 2 times its queueing delay budget at that server, i.e., $\mathcal{E}_{i,\pi(i,j)} = \delta_{i,j}$. Finally, for the EDF discipline, we choose the priority index for flow-i's k^{th} packet at its j^{th} hop as $t_{i,j}^k + \delta_{i,j}$, and for the WFQ discipline, we assign the same weight for each flow.

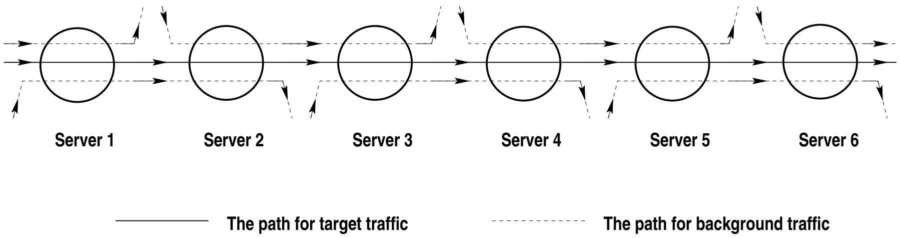


Fig. 3. Tandem Network Topology

B. WFQ Admission Control

To compare CMS with WFQ, we also implement WFQ admission control as follows. Consider flow i with guaranteed bandwidth g_i^m guaranteed at router m such that

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

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 [38], the probability of the end-to-end deadline D_i violation of the traffic of flow i can be bounded by

$$P\Big[\max_{t>0}\Big\{\sum_{k\in\mathcal{S}}\big[B_k(t)-g_k(t+D_i)\big]\Big\}>0\Big], \qquad (26)$$

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.

C. Computing Performance Bounds

To compute

$$P\Big[\max_{t>0}\Big\{\sum_{k\in\mathcal{I}(m)} \overline{B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})} - C_m(t+D_{i,j})\Big\}>0\Big],$$

we use the maximum variance approach developed in [9]. Let

$$\begin{aligned}
\sigma_t^2 \quad &= \quad \mathrm{var}\Big\{\sum_{k\in I(m)} \overline{B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})} \\
&\qquad\qquad - C_m(t+D_{i,j})\Big\}, \\
&= \quad \mathrm{var}\Big\{\sum_{k\in I(m)} B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)}) \\
&\qquad\qquad - C_m(t+D_{i,j})\Big\}, \\
m_t \quad &= \quad E\Big\{C_m(t+D_{i,j}) \\
&\qquad - \sum_{k\in\mathcal{I}(m)} \overline{B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})}\Big\}, \\
&= \quad E\Big\{C_m(t+D_{i,j}) \\
&\qquad - \sum_{k\in\mathcal{I}(m)} B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})\Big\}, \\
\alpha \quad &= \quad \inf_t \frac{m_t}{\sigma_t}.
\end{aligned}$$

Approximating $\sum_{k\in\mathcal{I}(m)} \overline{B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})}$ $-C_m(t+D_{i,j})$ as Gaussian, the following upper bound can been obtained.

$$P\Big[\max_{t>0}\Big\{\sum_{k\in\mathcal{I}(m)} \overline{B_k(t-\delta_{k,j_k}+T_{k,j_k}+\mathcal{E}_{k,\pi(k,j_k-1)})} - C_m(t+D_{i,j})\Big\}>0\Big] < e^{-\frac{\alpha^2}{2}}. \quad (27)$$

A proof of this bound can be found in [9] and a detailed comparative performance study in [22]. Roughly, the approach uses the dominant time scale, the value of t minimizing the expression above, to derive the exponential asymptotic upper bound of Equation (27). Our goal here is to use the technique as an efficient and accurate means to evaluate the admission control conditions. Regardless, refinements based on large deviations theory can also be applied within the context of statistical envelopes [4].

D. Admissible Regions

Here, we compare measured and predicted admissible regions for CMS and WFQ networks for the scenario described above, and also present measured admissible regions for EDF and FIFO networks. The results of the experiments are depicted in Figure 4. The figure shows network utilization vs. the end-to-end queueing delay of the target traffic, i.e., aggregate average traffic rate divided by link capacity vs. the delay target. The combinations of the target traffic and cross traffic and the corresponding expected end-to-end queueing delays are given in Table 2.

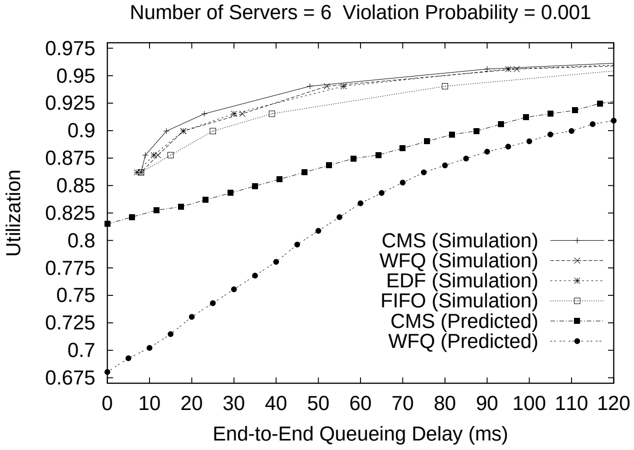


Fig. 4. Measured and Predicted Admissible Regions

The simulation curves depict the maximum number of flows

Utilization	Number of Target Flows	Number of Cross Flows	Expected Queueing Delay D
86.2%	55	220	3 msec
87.7%	56	224	6 msec
89.9%	57	230	12 msec
91.5%	58	234	20 msec
91.5%	59	236	30 msec
94.1%	60	240	45 msec
95.6%	61	244	60 msec

TABLE II

TRAFFIC COMBINATION AND EXPECTED END-TO-END QUEUEING DELAY

(scaled to utilization) that can be multiplexed such that the delay target (depicted on the horizontal axis) is satisfied with delay-bound-violation probability no more than 0.001. For these curves, 10 independent simulation runs of 400 simulated seconds are performed and mean results are reported. Similarly, the 'predicted' curve depicts the maximum number of flows admitted by our admission control algorithm, under the targeted delay depicted on the horizontal axis, and a delay-bound-violation probability of 0.001.

We make the following observations regarding this figure. First, notice that the admissible region for the CMS network is larger than that of the WFQ, EDF, and FIFO networks. For example, for the same violation probability (≤ 0.001) and the same traffic load (60 target traffic flows and 240 cross traffic flows at each node, i.e., approximately 94.1% utilization), CMS can support an end-to-end delay of 48 msec whereas WFQ's delay is 52 msec; EDF's delay is 56 msec; ans FIFO's delay is 80 msec. Thus, while WFQ achieves local fairness of bandwidth sharing at each node [26] and EDF minimizes the queueing delay at a single server system [16], CMS uses the coordination property to minimize end-to-end delay and achieve *global* performance properties.

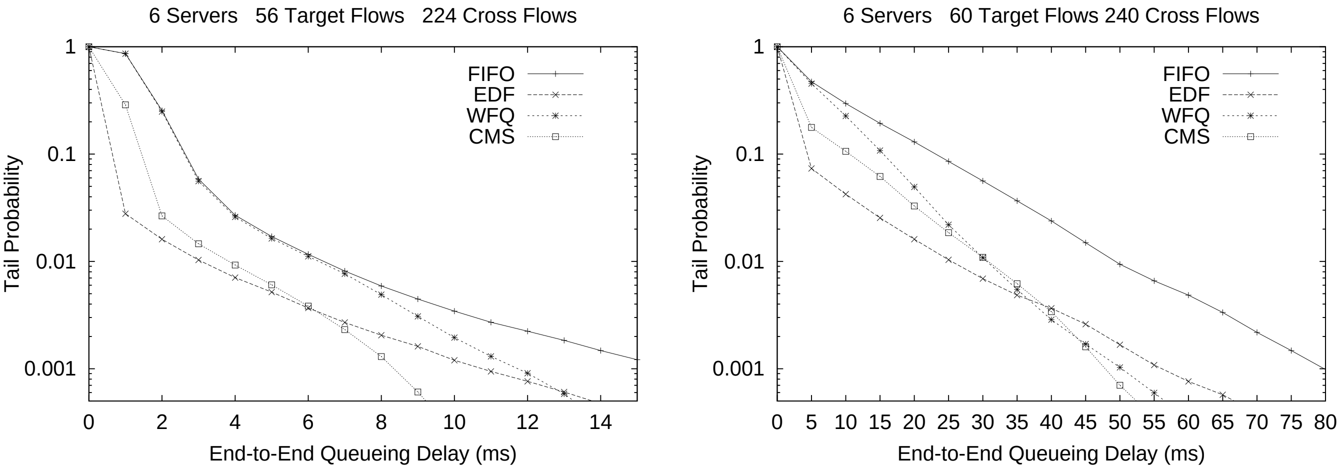
Second, we observe that our CMS admission control algorithm is able to exploit a large fraction of the available statistical multiplexing gain. For example, by an end-to-end delay bound of 60 msec, the CMS admission control algorithm admits a set of flows which contains 56 target traffic flows and 224 cross traffic flows at every server, within 6.1% of the actual utilization achievable in simulations (60 target traffic flows and 240 cross traffic flows). In contrast, for the WFQ network, 54 target traffic flows and 208 cross traffic flows are admitted at each router, which is more than 10% less than the utilization achievable in simulations (58 target traffic flows and 234 cross traffic flows). The reason for the more conservative nature of WFQ admission control is that traffic is treated as traversing the network on a guaranteed-rate "pipe" between an ingress and egress router, without taking into account inter-class resource sharing among pipes. Thus, with more traffic classes and more complex topologies, such an approach will suffer further utilization penalties.

E. End-to-End Delay Performance

Here, we compare the end-to-end queueing delays incurred by the target traffic for networks with CMS, WFQ, FIFO, and EDF schedulers. FIFO services provide baseline results for a scheduler with neither coordination nor QoS differentiation. In contrast, EDF provides differentiation and optimality at a single node, but does not employ coordination, so that gains of CMS vs. EDF are strictly due to coordination.

We measure end-to-end delay distributions for two scenarios: (a) 56 target traffic flows and 224 cross traffic flows at each server (the expected end-to-end queueing delay bound for each flow is 6 msec, and so the increments of the priority index at each server are 1 msec for target traffic, 3 msec for cross traffic with 2-hop path, and 6 msec for cross traffic with 1-hop path); (b) 60 target traffic flows and 240 cross traffic flows at each server (the expected end-to-end queueing delay bound for each flow is 45 msec, and so the increments of the priority index at each server are 7.5 msec for target traffic, 22.5 msec for cross traffic with 2-hop path, and 45 msec for cross traffic with 1-hop path). For this fixed number of flows and utilization, Figure 5 depicts the end-to-end queueing delay and its corresponding violation probability.

We make two observations regarding the figure. First, the QoS-violation probabilities for CMS, WFQ, and EDF are always smaller than those for FIFO. For example, in Figure 5 (b), the violation probability for a 50 msec end-to-end queueing delay bound is 0.0007 for CMS, 0.0016 for EDF, 0.001 for WFQ, and 0.0094 for FIFO. Second, notice that for smaller delays (less than 6 msec with 87.7% utilization or less than 35 msec with 94.1% utilization), the CMS violation probability larger than EDF's whereas for larger end-to-end queueing delays (greater than 6 msec with 87.7% utilization or greater than 35 msec with 94.1% utilization scenario) is smaller than EDF's. The reason for this is that in a CMS network, packets which suffer excessive queueing delays at upstream nodes have an opportunity to "catch up" at a downstream node, by having a higher (relative) priority index. In contrast, in EDF networks, each router treats packets locally according to their arrival time and local deadline, without regard to whether this arrival time is late or early. Thus, the experiments indicate that coordinated scheduling also has performance advantages, in addition to its other properties (e.g., tractability) established above.



(a) Utilization = 87.7% D = 6 msec.

(b) Utilization = 94.1% D = 45 msec.

Fig. 5. Comparison of FIFO, EDF, WFQ, and CMS Disciplines (Exponential On-Off Traffic)

V. Conclusion

In this paper, we developed a framework for Coordinated Multihop Scheduling (CMS). With a definition of the fundamental coordination property, we showed how a number of schedulers from the literature can be characterized as CMS disciplines. We then developed a general theory based on traffic and service envelopes to analyze CMS networks and devised admission control tests for statistical end-to-end services. We showed that CMS disciplines limit traffic distortion to within a narrow range, thereby providing a foundation for efficient and scalable multi-node services. We performed a set of simulation and admission control experiments to illustrate the accuracy of the approach to control multi-node admissions and demonstrated potential performance advantages of CMS as compared to WFQ, EDF, and FIFO. Our results present a new framework for understanding end-to-end statistical services in differentiating and work-conserving schedulers.

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.

[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, June 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 multiplex 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 A. Parekh. Optimal multiplexing on a single link: Delay and buffer requirements. *IEEE/ACM Transactions on Information Theory*, 43(5), 1997.

[17] 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.

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

[19] 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.

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

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] J. Qiu and E. Knightly. Inter-class resource sharing using statistical service

envelopes. In *Proceedings of IEEE INFOCOM '99*, New York, NY, March 1999.

[28] M. Reisslein, K. Ross, and S. Rajagopal. A framework for guaranteeing statistical qos. *IEEE/ACM Transactions on Networking*, 10(1):27–42, February 2002.

[29] Sheldon M. Ross. *Stochastic Processes*. Wiley, 1983.

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

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.