Schedulability Criterion and Performance Analysis of Coordinated Schedulers

Chengzhi Li and Edward W. Knightly

Abstract—**Inter-server coordinated scheduling is a mechanism for downstream nodes to increase or decrease a packet's priority according to the congestion incurred at upstream nodes. In this paper, we derive an end-to-end schedulability condition for a broad class of coordinated schedulers that includes CJVC and CEDF. In contrast to previous approaches, our technique purposely allows flows to violate their local priority indexes while still providing an end-to-end delay bound. We show that under a simple priority assignment scheme, coordinated schedulers can outperform WFQ schedulers, while replacing per-flow scheduling operations with a simple coordination rule. Finally, we illustrate the performance advantages of coordination through numerical examples and simulation experiments.**

I. Introduction

In the past decade, there has been significant progress in the design of packet scheduling algorithms, including service disciplines which achieve performance isolation [20], [26], quality of service differentiation [10], [13], [19], and scalable core-stateless implementation [4], [23], [28].

Simultaneously, new theoretical tools have been devised to analyze the performance properties of such multi-class schedulers. For example, exact delay bounds for Earliest Deadline First (EDF) and Strict Priority (SP) schedulers are derived in [18]. Moreover, multi-node delay bounds have been developed for networks of elements characterized by service curves using "network calculus" [5], an approach which encompasses and generalizes previous results for networks of Weighted Fair Queueing (WFQ) servers [21] and rate-controlled servers [15], [26]. In general, such techniques provide *schedulability conditions*, i.e., constraints that, if satisfied, ensure that all packets of all flows will meet their respective delay bounds without violation or loss.

Recently, a class of schedulers that utilize *coordination* of priorities among nodes [2], [16], [29] has been studied. A scheduler that employs coordination can give a packet higher or lower priority at downstream nodes depending on whether the packet was serviced late or early at upstream nodes. This intuitively appealing concept has been applied in a number of service disciplines proposed in the literature including FIFO+ [7] and Global EDF [6]. Moreover, such schedulers have potential applications to multi-service networks since they can provide improved or guaranteed end-to-end performance using simple, work conserving, scheduling algorithms that do not require per-flow operations. Indeed, it was shown in [16] that core stateless service disciplines such as Core-stateless Jitter Virtual Clock (CJVC) [24] can also be expressed by a simple coordination mechanism.

The goal of this paper is to provide a schedulability condition and analytical framework for coordinated schedulers. Our approach represents a fundamental departure from previous techniques in two ways. First, our schedulability condition allows packets to *violate* local per-node constraints, while still ensuring delay bounds are satisfied end-to-end, i.e., by the final hop. Allowing such local violations is crucial to exploiting the key multi-node property of coordinated schedulers. Consequently, techniques that require all packets to satisfy their local constraints at each node to ensure end-to-end schedulability cannot be applied. Second, previous techniques rely on either per-flow traffic re-shaping [15], [26] or per-flow scheduling [3], [9], [21], [22] (such as in WFQ) to derive multi-node schedulability conditions. In contrast, we consider a scenario of work-conserving servers with no per-flow operations.

The contribution of this paper is as follows. First, we develop an end-to-end schedulability condition for a broad class of coordinated schedulers that includes Coordinated EDF (CEDF) and CJVC. Our key technique is to introduce a virtual partition of the traffic into *essential* and non-essential traffic, where only the former traffic can impede a packet in meeting its delay bound. With this concept, we derive a bound on the essential traffic at downstream nodes and show that distortion of the essential traffic is confined to within a narrow range. In other words, we show that *coordination* limits downstream distortion analogous to the way per-flow traffic reshaping eliminates distortion.

Second, we study the problem of assigning local priority indexes. We show that with a particular assignment scheme, coordinated schedulers can achieve not only the same end-to-end delay bound as WFQ, but also a tighter end-to-end delay bound than WFQ, yet without per-flow packet forwarding in the network core. In other words, we establish that any set of flows that can be scheduled in WFQ networks can also be scheduled in coordinated scheduling networks.

Finally, we illustrate and quantify the practical advantages of coordinated scheduling with a set of numerical examples and simulation experiments. We first devise a simple example with three flows to illustrate that coordinated schedulers can achieve a lower delay bound than WFQ schedulers. We then use simulations of exponential and Pareto on-off traffic flows and a 6-node network to illustrate statistical differences between coordinated scheduling, EDF, and WFQ.

The remainder of this paper is organized as follows. In Section 2, we provide background and a precise definition of inter-server coordination. In Section 3, we develop a key tool for multi-node analysis and show how to bound the essential traffic at downstream nodes. In Section 4, we use this traffic bound to provide a global schedulability criterion for networks using coordinated scheduling. Next, in Section 5, we study priority index assignment and its relationship to WFQ. Finally, in Section 6, we compare coordinated and non-coordinated service disciplines using numerical examples and simulations, and in Section

C. Li is with the University of Texas at Arlington, Arlington, TX 76010 USA. E. Knightly is with Rice University, Houston, TX 77005 USA. This research was supported by NSF Grant ANI-0085842, Intel Corporation, Hewlett Packard Laboratories, and by a Sloan Fellowship. A subset of this paper was published in the Proceedings of ITC-17 2001.

7 we conclude.

II. Background on Inter-Server Coordination

In this section, we provide a formal definition of coordination among servers. We then illustrate the generality of the definition by describing how service disciplines from the literature, namely CEDF and CJVC, can be characterized as examples of coordinated schedulers.

Term	Definition
M	number of servers
C_m	capacity of server m
$\mathcal{I}(m)$	set of all flows at server m
$\pi(i,j)$	j^{th} hop of flow i
i_m	the hop of flow i such that $\pi(i,i_m)=m$
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
$\delta_{i,j}$	mean value of $\delta_{i,j}^k$
$\eta_{i,j}$	variance of $\delta_{i,j}^k$
τ_i^k	increment of priority index of the k^{th} packet of flow i at its first hop
$d_{i,j}^k$	priority index of the k^{th} packet of flow i at its j^{th} hop
$F_{i,j}^k$	departure time of the k^{th} packet of flow i from its j^{th} hop
l_i^k	size of k^{th} flow-i packet
l_i^{max}	maximum flow-i packet size
$\tilde{l}_i^{max}$	maximum size of all packets except flow-i packets
$f_{i,j}(t)$	total flow-i traffic at its j^{th} hop during $[0,t)$
$b_i(I)$	flow i traffic envelope at its first hop
(β_i,ρ_i)	(burst, rate) leaky bucket parameters of $b_i(I)$
$f_{i,j}^*(t,s)$	flow-i traffic with priority index no larger 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
σ_i	burstiness of $\Gamma_{i,1}(I)$
γ_i	average rate of $\Gamma_{i,1}(I)$
$\tau_m(t,s)$	void time of server m before time t related to time s
$W_m^s(x)$	total traffic with priority index no larger than s queued at server m at time x
$D_{i,j}$	bound for priority index violation by flow i at server j
$\mathcal{D}_i$	upper bound on the end-to-end delay suffered by flow-i packets
T_{i,i_m}	$[\delta_{i,i_m}-D_{i,\pi(i,i_m-1)})]-2\sum_{h=2}^{i_m}\eta_{i,h}$
$\Theta_m(I)$	$\{i \mid T_{i,i_m} > I\}$
S_x	$\{i \mid T_{i,i_m} > T_{x,x_m}\}, \Omega_x=\{i \mid T_{i,i_m} < T_{x,x_m}\}$
Ω_x	$\{i \mid T_{i,i_m} < T_{x,x_m}\}$
Q_x	$\{i \mid T_{i,i_m} = T_{x,x_m}\}$

TABLE I
Notation

A. Definition and Properties

Definition 1 (Coordinated Multihop Scheduling) Consider a server which services packets in increasing order of their priority indexes. A scheduler possesses the CMS property if

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

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; τ_i^k and $\delta_{i,j}^k$ are the increment of the priority index of the k^{th} packet of flow i at the corresponding hops; $\delta_{i,j}^k$ $(j=2,3,\cdots,N_i)$ is determined when the k^{th} packet of the flow i arrives at its first hop and $\delta_{i,j}^k \in [\delta_{i,j}-\eta_{i,j}, \delta_{i,j}+\eta_{i,j}]$, $\forall k \geq 1$ where $\delta_{i,j}, \eta_{i,j} \geq 0$.[1]

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, which in turn depends on its entrance time into the network. Therefore, if a packet violates its priority index at an upstream server, downstream servers 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 servers will reduce the priority of the packet, enabling other packets to be serviced ahead of it. Thus, even though the distributed servers operate independently, the priority index of each packet is communicated downstream via insertion of a label into the packet header (e.g., as described in [23]) so that the servers (virtually) coordinate to provide an end-to-end service.

B. CJVC and CEDF

An example of a service discipline in the CMS class is Core-stateless Jitter Virtual Clock. CJVC was proposed in [24] 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:

$$d_{i,j}^k = \begin{cases} \max\{t_i^k, d_{i,1}^{k-1}\}+\frac{l_i^k}{r_i}, & j=1 \\ d_{i,j-1}^k+\frac{l_i^k}{r_i}+\xi_i^k, & j>1 \end{cases} \quad (2)$$

where flow-i k^{th} packet size and reserved bandwidth are given by l_i^k and r_i respectively, and ξ_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 $\frac{l_i^k}{r_i}+\xi_i^k \in [\delta_{i,j}-\eta_{i,j}, \delta_{i,j}+\eta_{i,j}]$, 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, work-conserving CJVC is a coordinated network service discipline in which the increment of the priority index is a function of the reserved bandwidth of the corresponding flow.

In [1], [2], [6], coordination within the context of EDF was studied. We refer to such schedulers as Coordinated Earliest Deadline First (CEDF) if the priority indexes are assigned as

$$d_{i,j}^k = t_i^k + G_{i,1} + G_{i,2} + \cdots + G_{i,j}, \quad (3)$$

clearly expressible in the form of Equation (1), where $G_{i,j}$ is a constant that refers to the per-node delay-bound increment for the j^{th} hop of flow i. Note that in this case, $\delta_{i,j}^k = \delta_{i,j} = G_{i,j}$ and $\eta_{i,j}=0$.

[1] Notation is summarized in Table I.

Our theoretical results address all schedulers satisfying the CMS definition, and throughout this paper, we use CEDF and CJVC as example service disciplines. Discussion of other schedulers can be found in [16], [28].

C. Example

Consider the 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 $\tau_i^k = \delta_{i,2}^k = 5$. 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 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 (excessively delayed) 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.

Thus, the example illustrates how distributed servers can (virtually) coordinate priority indexes to improve the likelihood of satisfying an end-to-end delay constraint.

III. Traffic Characterization in Downstream Nodes

In multi-node networks without traffic re-shaping, traffic characteristics are distorted at downstream nodes as compared to their properties at the network entrance. In this section, we derive a burstiness bound for arriving traffic at downstream nodes, which we use as a basis for deriving a global schedulability criterion in the next section.

A. Preliminaries

Let $f_{i,j}(t)$ denote the total arrival traffic of flow-i at its j^{th} hop (denoted as server $\pi(i,j)$) during time interval $[0,t)$. More precisely, we have

$$f_{i,j}(t) \quad = \quad \sum_{t_{i,j}^k < t} l_i^k, \tag{4}$$

where $t_{i,j}^k$ is the time when the k^{th} flow-i packet with size l_i^k arrives at server $\pi(i,j)$.[2] Ignoring propagation delay, the departure traffic of flow i from server $\pi(i,j)$ is the arrival traffic of flow i to server $\pi(i,j+1)$. To simplify notation, we use $f_{i,j+1}(t)$ to denote the departure traffic of flow i from server $\pi(i,j)$ as well as the arrival traffic of flow i to server $\pi(i,j+1)$. As in [8], we call a non-negative and non-decreasing function $b_i(I)$ the source traffic envelope of flow i if $\forall t, I > 0$,

$$f_{i,1}(t+I) - f_{i,1}(t) \leq b_i(I). \tag{5}$$

We also assume that the network is stable if for $m = 1, 2, \cdots, M$,

$$\overline{\lim}_{I \to \infty} \frac{\sum_{i \in \mathcal{I}(m)} b_i(I)}{I} < C_m, \tag{6}$$

where M is the number of servers in the network, $\mathcal{I}(m)$ is the set of all flows served by server m, and C_m is the capacity of server m. According to [21], [25], acyclic networks or cyclic networks with ring topology are stable if Inequality (6) is satisfied. Discussion of stability of networks with general topology can be found in [17].

B. Virtual Partition

Here, we define *essential traffic* as a fundamental notion for analysis of coordinated schedulers that enables us to accurately bound the queueing delay experienced by the traffic.[3] In particular, for a given priority index s, all arriving traffic of server m arriving in $[0,t)$ can be virtually decomposed according to whether or not its priority index is larger than s. As only the portion of traffic with priority index smaller than or equal to time s affects the time when traffic with priority index s is served, 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 s ($s \geq t$) at server $\pi(i,j)$ is defined as the total flow-i traffic with priority index no larger than s arriving at server $\pi(i,j)$ in $[0,t)$, i.e.,

$$f_{i,j}^*(t,s) = \sum_{t_{i,j}^k < t, d_{i,j}^k \leq s} l_i^k. \tag{7}$$

To illustrate, for the traffic with the arrival pattern described in Figure 1(a), example values of its essential traffic are given by: $f_{i,1}^*(3,s) = 0$ if $s \in [3,5)$; $f_{i,1}^*(3,s) = l_i^1$ if $s \in [5,6)$; $f_{i,1}^*(3,s) = l_i^1 + l_i^2$ if $s \in [6,7)$; $f_{i,1}^*(3,s) = l_i^1 + l_i^2 + l_i^3$ if $s \in [7,\infty)$.

In addition to essential traffic, an important characteristic of server m is the void time before a given time t, and relative to s ($s \geq t$), denoted by $\tau_m(t,s)$ and defined as

$$\tau_m(t,s) = \max\{x \mid x \leq t \text{ and } W_m^s(x) = 0\}, \tag{8}$$

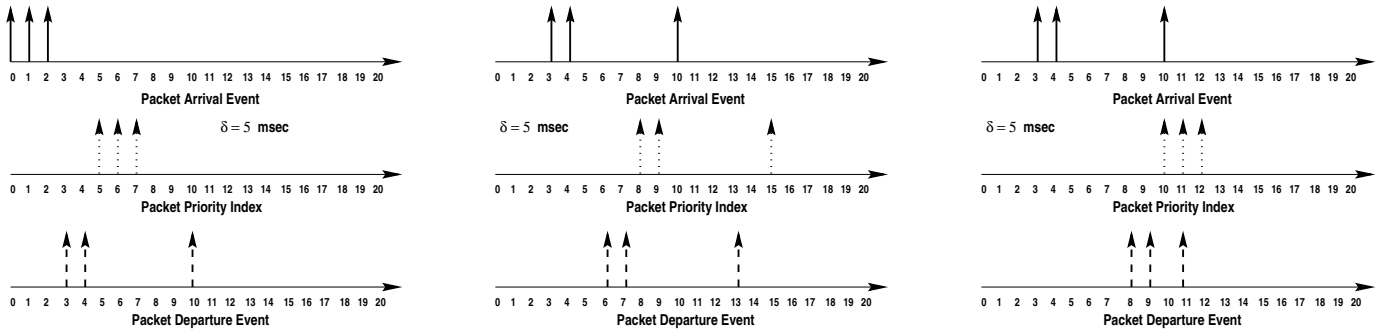
where $W_m^s(x)$ is the total amount of traffic, that has priority index smaller than or equal to s, queueing at server m at time x. In other words, void time refers to the largest time less than t such that there is no traffic backlogged with priority index smaller than or equal to s. Notice that for an initially idle network, the void time is guaranteed to exist.

C. Burstiness Bound at the Ingress Server

To compute the bounds of queueing delays suffered by the traffic with priority index s arriving at time t at server

[2] For convenience, we use $\sum_{t_{i,j}^k < t} l_i^k$ to denote $\sum_{k: t_{i,j}^k < t} l_i^k$. Furthermore, we consider the arrival and departure times of a packet to be the arrival and departure times of its last bit.

[3] A similar traffic function was used in the proof of theorem 8 in [14] albeit without a formal definition.



(a) CMS and EDF at the First Hop (b) EDF at the Second Hop (c) CMS at the Second Hop

Fig. 1. Illustration of Coordination

m, we only need to consider the essential traffic arriving in $[\tau_m(t,s),t)$. This is because $\tau_m(t,s)$ is the last time before t that there was no traffic with priority index smaller than or equal to s queued at server m. The envelope of the essential traffic of a flow in such an interval is defined as follows.

Definition 3 (Essential Traffic Envelope) A non-negative, non-decreasing function $\Gamma_{i,j}(I)$ is called the essential traffic envelope of flow-i traffic at its j^{th} hop if $\forall s \geq t > 0$,

$$f^*_{i,j}(t,s) - f^*_{i,j}(\tau_m(t,s),s) \leq \Gamma_{i,j}(s - \tau_m(t,s)), \quad (9)$$

where $\pi(i,j) = m$ and $\tau_m(t,s)$ is defined in Equation (8).[4]

Since the essential traffic at a downstream server depends on the corresponding essential traffic at the ingress server (i.e., the network entrance), we first provide an upper bound for the essential traffic envelope at the ingress server.

Lemma 1: An essential traffic envelope of flow i at its first hop is given by:

$$\Gamma_{i,1}(I) = \max_{t\geq 0} \sum_{t < d^k_{i,1} \leq t+I} l^k_i. \quad (10)$$

Proof: According to Definition 2, we have that $\forall s \geq t > 0$,

$$f^*_{i,1}(t,s) = \sum_{t^k_{i,1} < t, d^k_{i,1} \leq s} l^k_i \leq \sum_{d^k_{i,1} \leq s} l^k_i. \quad (11)$$

Moreover, since $\tau_m(t,s) \leq s$, we have

$$\begin{aligned} f^*_{i,1}(\tau_m(t,s),s) &= \sum_{t^k_{i,1} < \tau_m(t,s), d^k_{i,1} \leq s} l^k_i \\ &\geq \sum_{t^k_{i,1} < \tau_m(t,s), d^k_{i,1} \leq \tau_m(t,s)} l^k_i. \quad (12) \end{aligned}$$

Since $t^k_{i,1} < d^k_{i,1}$, we have

$$\sum_{t^k_{i,1} < \tau_m(t,s), d^k_{i,1} \leq \tau_m(t,s)} l^k_i = \sum_{d^k_{i,1} \leq \tau_m(t,s)} l^k_i. \quad (13)$$

That is $f^*_{i,1}(\tau_m(t,s),s) \geq \sum_{d^k_{i,1} \leq \tau_m(t,s)} l^k_i$. Therefore,

$$\begin{aligned} & f^*_{i,1}(t,s) - f^*_{i,1}(\tau_m(t,s),s) \\ \leq & \sum_{d^k_{i,1} \leq s} l^k_i - \sum_{d^k_{i,1} \leq \tau_m(t,s)} l^k_i \\ = & \sum_{\tau_m(t,s) < d^k_{i,1} \leq s} l^k_i \\ \leq & \max_{x\geq 0} \sum_{x < d^k_{i,1} \leq x + (s - \tau_m(t,s))} l^k_i \\ = & \Gamma_{i,1}(s - \tau_m(t,s)). \quad (14) \end{aligned}$$

Therefore, by Definition 3, $\Gamma_{i,1}(I)$ is an essential traffic envelope of flow i at its first hop. □

Based on Lemma 1, we have $\Gamma_{i,1}(0) = \lim_{I\to 0^+} \Gamma_{i,1}(I) \geq l^{max}_i$, which we use to derive the schedulability criterion in the next section.

D. Downstream Servers

At the output of a multiplexer, a traffic flow's characteristics (such as its traffic envelope) are distorted. Without additional mechanisms such as per-flow re-shaping, this distortion can become more severe at each downstream node. We now show that under coordinated network schedulers, the distortion of the essential traffic is limited due to coordination itself. That is, a flow's distortion is limited by downstream mechanisms to catch up late packets and delay early packets. Recall that we only consider stable networks, so that the queueing delay is bounded.

Lemma 2: If each flow-i packet has not missed its priority indexes at server $\pi(i,j-1)$ $(j \geq 2)$ by more than $D_{i,\pi(i,j-1)}$, then an essential traffic envelope $\Gamma_{i,j}(I)$ of flow i at its j^{th} hop (server m) is given by

$$\Gamma_{i,j}(I) = \Gamma_{i,1}(I - T_{i,j}), \quad (15)$$

where $T_{i,j} = [\delta_{i,j} - D_{i,\pi(i,j-1)}] - 2\sum_{h=2}^{j} \eta_{i,h}$.

Remark: The k^{th} packet of flow-i missing its priority index at server $\pi(i,j-1)$ by more than $D_{i,\pi(i,j-1)}$ indicates that $f^k_{i,j-1} - d^k_{i,j-1} \leq D_{i,\pi(i,j-1)}$, where $f^k_{i,j-1}$ is the departure time of the packet from server $\pi(i,j-1)$, and $d^k_{i,j-1}$ is the priority index of the packet at server $\pi(i,j-1)$.

Proof: In order to simplify notation, let $\pi(i,j) = m$, i.e., the j^{th} hop of flow i is server m. For all $s \geq t > 0$, consider the time interval $[\tau_m(t,s),t)$ and the quantity $f^*_{i,j}(t,s) - f^*_{i,j}(\tau_m(t,s),s)$. Since $\tau_m(t,s) \leq t$, $f^*_{i,j}(t,s) \geq f^*_{i,j}(\tau_m(t,s),s)$, there are two cases:

[4] To simplify notation, we assume that $\Gamma_{i,j}(I) = 0$, if $I < 0$; and $\Gamma_{i,j}(0) = \lim_{I\to 0^+} \Gamma_{i,j}(I)$.

Case 1: $f^*_{i,j}(t,s) > f^*_{i,j}(\tau_m(t,s),s)$
According to Definition 2, we have

$$f^*_{i,j}(t,s) = \sum_{t^k_{i,j}<t, d^k_{i,j}\le s} l^k_i \le \sum_{d^k_{i,j}\le s} l^k_i. \quad (16)$$

Since $d^k_{i,j} = d^k_{i,1} + \sum_{h=2}^{j}\delta^k_{i,h}$, we have

$$\begin{aligned} f^*_{i,j}(t,s) &\le \sum_{d^k_{i,1}+\sum_{h=2}^{j}\delta^k_{i,h}\le s} l^k_i \\ &= \sum_{d^k_{i,1}\le s-\sum_{h=2}^{j}\delta^k_{i,h}} l^k_i \\ &\le \sum_{d^k_{i,1}\le s-\sum_{h=2}^{j}[\delta_{i,h}-\eta_{i,h}]} l^k_i \quad (17)\end{aligned}$$

On the other hand, we have

$$f^*_{i,j}(\tau_m(t,s),s) \ge \sum_{d^k_{i,j-1}\le \tau_m(t,s)-D_{i,\pi(i,j-1)}} l^k_i$$

reasoned as follows.
1. Since $f^*_{i,j}(t,s) > f^*_{i,j}(\tau_m(t,s),s)$, at least one packet of flow i with priority index smaller than or equal to s arrives at server m in $[\tau_m(t,s),t)$. Thus, all flow-i packets arriving at server m before $\tau_m(t,s)$ have priority indexes less than s;
2. All flow-i packets with priority indexes (at server $\pi(i,j-1)$) smaller than or equal to $\tau_m(t,s) - D_{i,\pi(i,j-1)}$ have departed server $\pi(i,j-1)$ and arrived at server m in $[0,\tau_m(t,s)]$; otherwise the definition of $D_{i,\pi(i,j-1)}$ is violated. Thus, all flow-i packets with priority indexes (at server $\pi(i,j-1)$) smaller than or equal to $\tau_m(t,s)-D_{i,\pi(i,j-1)}$ have priority indexes (at server $\pi(i,j)$) smaller than or equal to s and have arrived at server $\pi(i,j)$ before time $\tau_m(t,s)$.

Similar to Equation (17), we have

$$\begin{aligned} &f^*_{i,j}(\tau_m(t,s),s) \\ &\quad\ge \sum_{d^k_{i,1}\le \tau_m(t,s)-\sum_{h=2}^{j-1}[\delta_{i,h}+\eta_{i,h}]-D_{i,\pi(i,j-1)}} l^k_i.\end{aligned}$$

Therefore,

$$f^*_{i,j}(t,s) - f^*_{i,j}(\tau_m(t,s),s) \le \sum_{d^k_{i,1}\in\Delta} l^k_i, \quad (18)$$

where $\Delta = \Big(\tau_m(t,s) - \sum_{h=2}^{j-1}[\delta_{i,h}+\eta_{i,h}] - D_{i,\pi(i,j-1)}, s - \sum_{h=2}^{j}[\delta_{i,h}-\eta_{i,h}]\Big]$. Moreover, according to Lemma 1,

$$\begin{aligned} &f^*_{i,j}(t,s) - f^*_{i,j}(\tau_m(t,s),s) \\ &\le \Gamma_{i,1}(s-\tau_m(t,s)-[\delta_{i,j}-D_{i,\pi(i,j-1)}] + 2\sum_{h=2}^{j}\eta_{i,h}) \\ &= \Gamma_{i,1}(s-\tau_m(t,s)-T_{i,j}). \quad (19)\end{aligned}$$

Case 2: $f^*_{i,j}(t,s) = f^*_{i,j}(\tau_m(t,s),s)$
In this case, Equation (19) still holds because of $\Gamma_{i,1}(I) \ge 0$.

Therefore

$$\Gamma_{i,j}(I) = \Gamma_{i,1}(I - T_{i,j}). \quad \square$$

This lemma characterizes a key property of coordinated schedulers, namely that a flow's traffic characteristics are minimally distorted at downstream servers. Specifically, if $T_{i,j}$ is a constant for $j = 2,3,\cdots,N_i$, we can use the same essential traffic envelope $\Gamma_{i,1}(I)$ to evaluate the local queueing delay suffered by flow i at each server along its path.

IV. End-to-End Schedulability Criterion

In this section, we derive a general end-to-end schedulability criterion for coordinated schedulers. In our approach, we allow packets to violate their local priority indexes and exploit the coordination property to obtain an end-to-end delay bound. Moreover, since priority indexes are not required to be equivalent to delay bounds, the approach provides flexibility in assignment of local priority indexes which we further exploit in the next section.

A. A Recursive Condition for Violating Packets

For an isolated EDF scheduler, the schedulability condition, which ensures that no packet violates its priority index, has been studied previously [12], [14], [18]. However, when the schedulability condition is not satisfied, it is important to bound the amount of time by which packets can miss their deadlines (priority indexes), especially for coordinated schedulers that allow packets to violate their local deadlines. Based on the property of coordinated schedulers exploited in Lemma 2, we provide a condition for bounding this time as follows.

Theorem 1: If $\forall i \in \mathcal{I}(m)$, each arriving packet of flow i at server m has not missed its priority indexes at the previous server by more than $D_{i,\pi(i,i_m-1)}$ and $d^k_{i,i_m-1} + D_{i,\pi(i,i_m-1)} \le d^k_{i,i_m}$ for $k \ge 1$,[5] then for a given flow $i^* \in \mathcal{I}(m)$, its packets will miss their priority index at server m by at most $D_{i^*,m}$ if $\forall I \ge T_{i^*,i^*_m}$,

$$\begin{aligned} &\sum_{i\in\mathcal{I}(m)} \Gamma_{i,1}(I - T_{i,i_m}) + \max_{i\in\Theta_m(I)} l^{max}_i \\ &\qquad\le C_m(I + D_{i^*,m}), \quad (20)\end{aligned}$$

where $l^{max}_i = \max_k l^k_i$, $T_{i,i_m} = [\delta_{i,i_m} - D_{i,\pi(i,i_m-1)}] - 2\sum_{h=2}^{i_m}\eta_{i,h}$, and $\Theta_m(I) = \{i \mid T_{i,i_m} > I\}$.
Remark: (1) The left hand side of Inequality (20) is an upper bound on the total amount of traffic needed to be served during any time interval with length $I + D_{i^*,m}$ if a flow-i packet does not violate its priority index by more than $D_{i^*,m}$ at the end of that time interval. The right hand side of Inequality (20) is the capacity of the server in that time interval. Inequality (20) with $\forall I \ge T_{i^*,i^*_m}$ implies that any flow-i packet does not violate its priority index by more than $D_{i^*,m}$. (2) While the schedulability criterion is presented within a theoretical framework as described in Theorem 1, the tests can indeed be implemented in computationally efficient ways. For end-to-end bandwidth or delay bound requirements, the simple priority assignment

[5] From now on,we use i_m to denote the i^{th}_m hop of flow i such that $\pi(i,i_m) = m$.

scheme provided in Section V will guarantee the required service for each flow provided that at each server, the total reserved bandwidth is not more than its capacity. Thus, after the priority index assignments are determined, admission control based on the scheme provided in Section V is quite simple: each server only needs to check that the total reserved bandwidth is not more than its capacity.

Proof: Without loss of generality, assume that the k^{th} packet of flow i^* with priority index $d^k_{i^*,i^*_m}$ arrives at server m at time t (i.e., $t = t^k_{i^*,i^*_m}$) and refer to this packet as the target packet. Since t is also the departure time of the k^{th} packet of flow i^* from server $\pi(i^*, i^*_m - 1)$ and $d^k_{i^*,i^*_m-1} + D_{i^*,\pi(i^*,i^*_m-1)} \leq d^k_{i^*,i^*_m}$, we have $t \leq d^k_{i^*,i^*_m-1} + D_{i^*,\pi(i^*,i^*_m-1)} \leq d^k_{i^*,i^*_m}$. Let $\tau = \tau_m(t, d^k_{i^*,i^*_m})$. According to Lemma 2, $\forall i \in \mathcal{I}(m)$, the total traffic coming from flow i with priority index smaller than or equal to $d^k_{i^*,i^*_m}$ during $[\tau, d^k_{i^*,i^*_m})$ is bounded by

$$\Gamma_{i,1}(d^k_{i^*,i^*_m} - \tau - T_{i,i_m}).$$

If at time τ server m is idle, then after time τ the total traffic that must be served before the departure of the target packet is bounded by $\sum_{i\in\mathcal{I}(m)} \Gamma_{i,1}(d^k_{i^*,i^*_m} - \tau - T_{i,i_m})$. Otherwise, at time τ, server m is serving a flow-j packet with priority index larger than $d^k_{i^*,i^*_m}$. Thus we can bound the total traffic that must be served after time τ and before the departure of the target packet by

$$\sum_{i\in\mathcal{I}(m)\wedge i\neq j} \Gamma_{i,1}(d^k_{i^*,i^*_m} - \tau - T_{i,i_m}) + l^{max}_j.$$

Furthermore, if $d^k_{i^*,i^*_m} - \tau - T_{j,j_m} \geq 0$, then by Definition 3 and Lemma 1, we have $\Gamma_{j,1}(d^k_{i^*,i^*_m} - \tau - T_{j,j_m}) \geq l^{max}_j$. On the other hand, if $d^k_{i^*,i^*_m} - \tau - T_{j,j_m} < 0$, we have $T_{j,j_m} > d^k_{i^*,i^*_m} - \tau$ and $j \in \{i \mid T_{i,i_m} > d^k_{i^*,i^*_m} - \tau\}$. Therefore,

$$\max_{i\,:\,T_{i,i_m} > d^k_{i^*,i^*_m} - \tau} l^{max}_i \geq l^{max}_j. \tag{21}$$

Thus, the total traffic that must be served after time τ and before the departure of the target packet is bounded by

$$\sum_{i\in\mathcal{I}(m)} \Gamma_{i,1}(d^k_{i^*,i^*_m} - \tau - T_{i,i_m}) + \max_{i\,:\,T_{i,i_m} > d^k_{i^*,i^*_m} - \tau} l^{max}_i. \tag{22}$$

Hence, the target packet misses its priority index at server m by at most $D_{i^*,m}$ if all such traffic expressed in Equation (22) can be serviced during time interval $[\tau, d^k_{i^*,i^*_m} + D_{i^*,m})$, i.e., if

$$\sum_{i\in\mathcal{I}(m)} \Gamma_{i,1}(d^k_{i^*,i^*_m} - \tau - T_{i,i_m}) + \max_{i\,:\,T_{i,i_m} > d^k_{i^*,i^*_m} - \tau} l^{max}_i \leq C_m(d^k_{i^*,i^*_m} - \tau + D_{i^*,m}). \tag{23}$$

Furthermore, if we replace $d^k_{i^*,i^*_m} - \tau$ by I and notice that $\Gamma_{i^*,1}(d^k_{i^*,i^*_m} - \tau - T_{i^*,i^*_m}) \geq l^k_{i^*}$ (the size of the target packet) indicating that $d^k_{i^*,i^*_m} - \tau \geq T_{i^*,i^*_m}$, then Equation (23) is satisfied if $\forall I \geq T_{i^*,i^*_m}$

$$\sum_{i\in\mathcal{I}(m)} \Gamma_{i,1}(I - T_{i,i_m}) + \max_{i\in\Theta(I)} l^{max}_i \leq C_m(I + D_{i^*,m}). \quad \square$$

Notice that if $\tau^k_i = \tau_i$, $\delta^k_{i,j} = \delta_{i,j}$, and $D_{i,\pi(i,j)} = 0$, Equation (20) is the schedulability condition provided in [18]. Thus, Theorem 1 is a generalization of the schedulability condition for an isolated EDF scheduler.

B. End-to-End Delay Bounds

Since the schedulability criterion given in Equation (20) decouples the priority index from the delay bound, the following corollary can be used to compute the end-to-end delay bound.

Corollary 1: Given the priority index increment assignments τ^k_i and $\delta^k_{i,j}$ ($j = 1, 2, \cdots, N_i$ and $k \geq 1$) for each flow i, if the conditions of Theorem 1 are satisfied for each flow at each server, then the end-to-end flow-i packet delay is bounded by

$$\max_{k\geq 1} \tau^k_i + \sum_{h=2}^{N_i} [\delta_{i,h} + \eta_{i,h}] + D_{i,\pi(i,N_i)}. \tag{24}$$

Proof: Let $F^k_{i,j}$ be the departure time of the k^{th} packet of flow i from its j^{th} hop. According to Theorem 1, we have

$$F^k_{i,N_i} \leq d^k_{i,N_i} + D_{i,\pi(i,N_i)}. \tag{25}$$

Furthermore, from Definition 1, we have

$$d^k_{i,N_i} = d^k_{i,1} + \sum_{h=2}^{N_i} \delta^k_{i,h}. \tag{26}$$

Therefore, the end-to-end delay of flow-i packets is bounded by

$$\begin{aligned}
&\max_{k\geq 1}\{F^k_{i,N_i} - t^k_{i,1}\} \\
&\leq \max_{k\geq 1}\{d^k_{i,1} - t^k_{i,1} + \sum_{h=2}^{N_i} \delta^k_{i,h} + D_{i,\pi(i,N_i)}\} \\
&\leq \max_{k\geq 1}\{d^k_{i,1} - t^k_{i,1}\} + \sum_{h=2}^{N_i} [\delta_{i,h} + \eta_{i,h}] + D_{i,\pi(i,N_i)} \\
&= \max_{k\geq 1} \tau^k_i + \sum_{h=2}^{N_i} [\delta_{i,h} + \eta_{i,h}] + D_{i,\pi(i,N_i)}. \quad \square
\end{aligned}$$

Observe that the maximum queueing delay of Equation (24) has three components. The first term has two interpretations which we illustrate by examples. If the network performs CEDF as in Equation (3), then τ^k_i is a constant and represents the local delay bound at the ingress node. Alternatively, if the network performs CJVC, then

$$\begin{aligned}
\max_k \tau^k_i = \max_k\{d^k_{i,1} - t^k_i\} &= \max_k\{\frac{l^k_i}{r_i} + [d^{k-1}_{i,1} - t^k_i]^+\} \\
&\leq \max_k \frac{l^k_i}{r_i} + \max_k [d^{k-1}_{i,1} - t^k_i]^+,
\end{aligned}$$

i.e., it is the maximum packet size divided by the guaranteed rate, plus the maximum amount of time a flow-i packet can arrive before its previous packet's priority index. The second term

is the sum of the upper bounds of the local priority indexes from the second to final hop. The third term represents the delay by which packets are allowed to violate the priority index at the *final* hop. As we will show in Section V, there is flexibility in how to assign all three of these components to obtain different end-to-end performance properties.

C. Leaky Bucket Flows

If the essential traffic envelopes at the ingress servers are bounded by affine functions, the schedulability criterion of Theorem 1 can be simplified. This scenario arises for both leaky bucket regulated traffic as well as virtual leaky-bucket smoothers as described in Section V-A.

Corollary 2: If each flow i has $\Gamma_{i,1}(I) = \sigma_i + \gamma_i I$ and $\sum_{i\in\mathcal{I}(m)} \gamma_i < C_m$ for $m = 1, 2, \cdots, M$, then Inequality (20) in Theorem 1 can be simplified as follows: if for any $x \in \mathcal{I}(m)$ with $T_{x,x_m} \geq T_{i^*,i^*_m}$,

$$\frac{\sum_{i\in Q_x}\sigma_i + \sum_{i\in\Omega_x}(\sigma_i - \gamma_i T_{i,i_m})}{C_m - \sum_{i\in\Omega_x}\gamma_i} + \frac{\max_{i\in S_x} l_i^{max} - C_m D_{i^*,m}}{C_m - \sum_{i\in\Omega_x}\gamma_i} \leq T_{x,x_m}, \quad (27)$$

where S_x[6] $= \{i \mid T_{i,i_m} > T_{x,x_m}\}$, $\Omega_x = \{i \mid T_{i,i_m} < T_{x,x_m}\}$, and $Q_x = \{i \mid T_{i,i_m} = T_{x,x_m}\}$.

Proof: Substituting $\Gamma_{i,1}(I) = \sigma_i + \gamma_i I$ into Inequality (20), we have

$$\sum_{i\in\mathcal{I}(m)} [\sigma_i + \gamma_i(I - T_{i,i_m})]1(I - T_{i,i_m}) + \max_{i\in\Theta_m(I)} l_i^{max} \leq C_m(I + D_{i^*,m}), \quad (28)$$

where $1(t) = 0$, if $t < 0$, and $1(t) = 1$, if $t \geq 0$. Since the left hand side of Inequality (28) is a piecewise-linear increasing function of I with finite discontinuous points $\{T_{i,i_m} \mid i \in \mathcal{I}(m)\}$, to verify Inequality (28) for all $I \geq T_{i^*,i^*_m}$, we only need to verify Inequality (28) for these discontinuous points in $[T_{i^*,i^*_m}, \infty)$. That is, for $x \in \mathcal{I}(m)$ with $T_{x,x_m} \geq T_{i^*,i^*_m}$,

$$\sum_{i\in\mathcal{I}(m)} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})]1(T_{x,x_m} - T_{i,i_m}) + \max_{i\in\Theta_m(T_{x,x_m})} l_i^{max} \leq C_m(T_{x,x_m} + D_{i^*,m}). \quad (29)$$

Since $\mathcal{I}(m) = S_x \cup Q_x \cup \Omega_x$ and $\Theta_m(T_{x,x_m}) = S_x$, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{I}(m)} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})]1(T_{x,x_m} - T_{i,i_m}) \\
&+ \max_{i\in\Theta_m(T_{x,x_m})} l_i^{max} \\
&= (\sum_{i\in S_x} + \sum_{i\in Q_x} + \sum_{i\in\Omega_x})[\sigma_i \\
&+\gamma_i(T_{x,x_m} - T_{i,i_m})]1(T_{x,x_m} - T_{i,i_m}) + \max_{i\in S_x} l_i^{max} \\
&= (\sum_{i\in Q_x} + \sum_{i\in\Omega_x})[\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})] + \max_{i\in S_x} l_i^{max} \\
&= \sum_{i\in Q_x} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})] \\
&+ \sum_{i\in\Omega_x} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})] + \max_{i\in S_x} l_i^{max} \\
&= \sum_{i\in Q_x} \sigma_i + \sum_{i\in\Omega_x} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})] + \max_{i\in S_x} l_i^{max}.
\end{aligned}$$

Therefore, Inequality (29) can be simplified as:

$$\sum_{i\in Q_x} \sigma_i + \sum_{i\in\Omega_x} [\sigma_i + \gamma_i(T_{x,x_m} - T_{i,i_m})] + \max_{i\in S_x} l_i^{max} \leq C_m(T_{x,x_m} + D_{i^*,m}). \quad (30)$$

By simple algebraic manipulation and noticing $C_m - \sum_{i\in\Omega_x}\gamma_i > 0$, Equation (30) can be further simplified as:

$$\frac{\sum_{i\in Q_x}\sigma_i + \sum_{i\in\Omega_x}(\sigma_i - \gamma_i T_{i,i_m})}{C_m - \sum_{i\in\Omega_x}\gamma_i} + \frac{\max_{i\in S_x} l_i^{max} - C_m D_{i^*,m}}{C_m - \sum_{i\in\Omega_x}\gamma_i} \leq T_{x,x_m}. \quad \square$$

In the next section, we apply this simplified schedulability criterion to assign priority indexes at downstream servers.

V. Priority Index Assignment for End-to-End Service

In this section, we develop a priority index assignment scheme and show that under this scheme, coordinated schedulers can achieve the same end-to-end delay bounds as WFQ.

A. At Ingress Servers

Suppose the ingress node services packets according to the virtual clock service discipline [11], [27]. Then the priority index increments at the ingress server are

$$\tau_i^k = [d_{i,1}^{k-1} - t_{i,1}^k]^+ + \frac{l_i^k}{\gamma_i}, \quad (31)$$

where, $d_{i,1}^0 = 0$ and γ_i is the reserved rate of flow i.

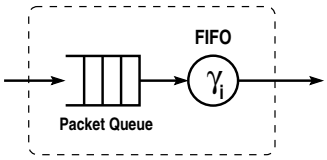


Fig. 2. Virtual Server

Conceptually, such virtual smoothing at the ingress node also spreads out the packets' priority indexes at downstream servers. Consequently, independent of the packet's service at upstream nodes, their priority indexes do not cluster at downstream nodes. Particularly, if $\delta_{i,j}^k = \delta_{i,j}$ and $\eta_{i,j} = 0$, from Equation (1), $d_{i,j}^k = d_{i,1}^k + \sum_{h=2}^j \delta_{i,h}$ and $d_{i,j}^{k_2} - d_{i,j}^{k_1} = d_{i,1}^{k_2} - d_{i,1}^{k_1}$.

Since $d_{i,1}^k = t_{i,1}^k + \tau_i^k = \max\{t_{i,1}^k, d_{i,1}^{k-1}\} + \frac{l_i^k}{\gamma_i}$ is the departure time of the k^{th} packet of flow i from the virtual server with

[6] $i^* \in S_x$ due to $T_{x,x_m} \geq T_{i^*,i^*_m}$.

capacity γ_i (see Figure 2), according to Definition 3 and Lemma 1,

$$\Gamma_{i,1}(I) \quad \leq \quad l_i^{max} + \gamma_i I. \tag{32}$$

If $\sum_{j\in\mathcal{I}(\pi(i,1))}\gamma_j \leq C_{\pi(i,1)}$, then according to Theorem 2 in [11],

$$F_{i,1}^k \leq d_{i,1}^k + \frac{l_i^{max}}{C_{\pi(i,1)}}, k = 1, 2, \cdots, \tag{33}$$

where $F_{i,1}^k$ is the departure time of the k^{th} packet of flow i from the ingress server $\pi(i,1)$, $l_x^{max} = \max_{k\geq 1} l_x^k$, $l_i^{max} = \max_{x\neq i} l_x^{max}$, and $C_{\pi(i,1)}$ is the capacity of server $\pi(i,1)$. Also, if $b_i(I) = \beta_i + \rho_i I$ and $\gamma_i \geq \rho_i$, using the results in [8], we have

$$d_{i,1}^k - t_{i,1}^k = \tau_i^k \leq \frac{\beta_i}{\gamma_i}. \tag{34}$$

Notice that in this case, $\frac{\beta_i}{\gamma_i}$ bounds the first term of the end-to-end delay bound of Corollary 1.

B. At Downstream Servers

At downstream servers and $j \geq 2$, we assign the priority index increments as:

$$\delta_{i,j}^k = \frac{l_i^{max}}{\gamma_i} + \frac{l_i^{max}}{C_{\pi(i,j-1)}} \tag{35}$$

where $C_{\pi(i,j-1)}$ is the capacity of server $\pi(i,j-1)$. It is easy to see that in this case, $\delta_{i,j}^k = \delta_{i,j}$,$\eta_{i,j} = 0$, and $T_{i,i_m} = \delta_{i,i_m} - D_{i,\pi(i,i_m-1)}$. This assignment is simpler than CJVC [24] and VTRS [28] because it does not require a slack variable or virtual time adjustment term for each packet. On the other hand, coordinated schedulers essentially treat all packets as having maximum size. The coordination property allows us to avoid this term and consequently to simplify the service discipline as well as obtain a tighter end-to-end delay bound.

We next show that with the above priority index assignment scheme, coordinated scheduling achieves the same end-to-end delay bound as WFQ.

Theorem 2: Consider the priority index increment assignment defined by Equations (31) and (35) and satisfying $\sum_{i\in\mathcal{I}(m)}\gamma_i < C_m$ for $m = 1, 2, \cdots, M$. If flow i^* satisfies $b_{i^*}(I) = \beta_{i^*} + \rho_{i^*} I$ and $\gamma_{i^*} \geq \rho_{i^*}$, then the end-to-end delay of flow-i^* packets is bounded by

$$\frac{\beta_{i^*}}{\gamma_{i^*}} + (N_{i^*} - 1)\frac{l_{i^*}^{max}}{\gamma_{i^*}} + \sum_{h=1}^{N_{i^*}} \frac{l_{i^*}^{max}}{C_{\pi(i^*,h)}}. \tag{36}$$

Proof: According to Corollary 2, to verify Inequality (20), we only need to verify Inequality (27). Since $\Gamma_{i,1}(I) = l_i^{max} + \gamma_i I$, substituting $\sigma_i = l_i^{max}$, $\delta_{i,i_m} = \frac{l_i^{max}}{\gamma_i} + \frac{l_i^{max}}{C_{\pi(i,i_m-1)}}$, $D_{i,\pi(i,i_m)} = \frac{l_i^{max}}{C_{\pi(i,i_m)}}$ (notice $d_{i,i_m-1}^k + D_{i,\pi(i,i_m-1)} = d_{i,i_m-1}^k + \frac{l_i^{max}}{C_{\pi(i,i_m-1)}} < d_{i,i_m-1}^k + \delta_{i,i_m} = d_{i,i_m}^k$ for all i, k, m), and $T_{i,i_m} = \delta_{i,i_m} - D_{i,\pi(i,i_m-1)} = \frac{l_i^{max}}{\gamma_i}$ into Inequality (27), we have

$$\frac{\sum_{i\in Q_x}\sigma_i + \sum_{i\in\Omega_x}(\sigma_i - \gamma_i T_{i,i_m})}{C_m - \sum_{i\in\Omega_x}\gamma_i}$$

$$+\frac{\max_{i\in S_x} l_i^{max} - C_m D_{i^*,m}}{C_m - \sum_{i\in\Omega_x}\gamma_i}$$

$$= \frac{\sum_{i\in Q_x} l_i^{max} + \sum_{i\in\Omega_x}(l_i^{max} - \gamma_i\frac{l_i^{max}}{\gamma_i})}{C_m - \sum_{i\in\Omega_x}\gamma_i}$$

$$+ \frac{\max_{i\in S_x} l_i^{max} - l_{i^*}^{max}}{C_m - \sum_{i\in\Omega_x}\gamma_i}$$

$$= \frac{\sum_{i\in Q_x} l_i^{max} + \max_{i\in S_x} l_i^{max} - \max_{i\neq i^*} l_i^{max}}{C_m - \sum_{i\in\Omega_x}\gamma_i}.$$

Since $i^* \bar{\in} S_x$, we have $S_x \subset \{i \mid i \neq i^*\}$ and $\max_{i\in S_x} l_i^{max} - \max_{i\neq i^*} l_i^{max} \leq 0$. Thus,

$$\frac{\sum_{i\in Q_x} l_i^{max} + \max_{i\in S_x} l_i^{max} - \max_{i\neq i^*} l_i^{max}}{C_m - \sum_{i\in\Omega_x}\gamma_i} \leq \frac{\sum_{i\in Q_x} l_i^{max}}{C_m - \sum_{i\in\Omega_x}\gamma_i}. \tag{37}$$

Since $\sum_{i\in\mathcal{I}(m)}\gamma_i < C_m$, we have $C_m - \sum_{i\in T_x}\gamma_i \geq \sum_{i\in Q_x}\gamma_i$. Thus,

$$\frac{\sum_{i\in Q_x} l_i^{max}}{C_m - \sum_{i\in\Omega_x}\gamma_i} \quad \leq \quad \frac{\sum_{i\in Q_x} l_i^{max}}{\sum_{i\in Q_x}\gamma_i}. \tag{38}$$

Notice that if $\frac{a_i}{b_i} = C, \forall i = 1, 2, \cdots, K$, then $\frac{\sum_{i=1}^K a_i}{\sum_{i=1}^K b_i} = C$; and $\forall i \in Q_x$, $\frac{l_i^{max}}{\gamma_i} = \frac{l_x^{max}}{\gamma_x}$, we have

$$\frac{\sum_{i\in Q_x} l_i^{max}}{\sum_{i\in Q_x}\gamma_i} = \frac{l_x^{max}}{\gamma_x} = \delta_{x,x_m} - D_{x,\pi(x,x_m-1)} = T_{x,x_m}.$$

Therefore,

$$\frac{\sum_{i\in Q_x}\sigma_i + \sum_{i\in\Omega_x}(\sigma_i - \gamma_i T_{i,i_m})}{C_m - \sum_{i\in\Omega_x}\gamma_i} + \frac{\max_{i\in S_x} l_i^{max} - C_m D_{i^*,m}}{C_m - \sum_{i\in\Omega_x}\gamma_i} \leq T_{x,x_m}$$

and Inequality (27) is verified. Furthermore, according to $d_{i,i_m-1}^k + D_{i,\pi(i,i_m-1)} = d_{i,i_m-1}^k + \frac{l_i^{max}}{C_{\pi(i,i_m-1)}} < d_{i,i_m-1}^k + \delta_{i,i_m} = d_{i,i_m}^k$ for all i, k, m and Corollary 1 and Corollary 2 and Equation (34), we know that the end-to-end delay bound for flow i^* is given by:

$$\frac{\beta_{i^*}}{\gamma_{i^*}} + (N_{i^*} - 1)\frac{l_{i^*}^{max}}{\gamma_{i^*}} + \sum_{h=1}^{N_{i^*}} \frac{l_{i^*}^{max}}{C_{\pi(i^*,h)}}. \quad \square$$

Notice that the end-to-end delay bound in Equation (36) is the same as that for WFQ [21] and VC [11].

Finally, we observe that coordinated schedulers can employ heterogeneously allocated per-node priority assignments in order to better utilize network resources. For example, flows could allocate a less stringent priority index to heavily loaded nodes. A general assignment scheme remains an important issue for future study in coordinated schedulers as well as other service disciplines.

flow 1	$\sigma_1 = L$	$l_1^{max} = L$	$\gamma_1 = \frac{C}{2}$	$\tau_1^k = 0$	$\delta_{1,2}^k = \delta_{1,2} = \frac{L}{C}, \eta_{1,2} = 0$	$\delta_{1,3}^k = \delta_{1,3} = \frac{2L}{C}, \eta_{1,3} = 0$
flow 2	$\sigma_2 = L$	$l_2^{max} = L$	$\gamma_2 = \frac{C}{2}$	$\tau_2^k = 0$	$\delta_{2,2}^k = \delta_{2,2} = \frac{3L}{C}, \eta_{2,2} = 0$	
flow 3	$\sigma_3 = L$	$l_3^{max} = L$	$\gamma_3 = \frac{C}{2}$	$\tau_3^k = 0$	$\delta_{3,2}^k = \delta_{3,2} = \frac{3L}{C}, \eta_{3,2} = 0$	

TABLE II
TRAFFIC PARAMETERS AND PRIORITY INDEX ASSIGNMENT.

VI. COORDINATION VS. NON-COORDINATION: NUMERICAL EXAMPLES AND SIMULATIONS

In this section we illustrate the performance advantages of inter-server coordination by comparing the CMS service discipline with non-coordinated schedulers WFQ and EDF. For rate-guarantee oriented service disciplines such as WFQ and VC, we show via a numerical example that with appropriate selection of priority indexes, CMS can outperform WFQ and VC. It was previously established that EDF with traffic shaping can provide the same delay bbound as Weighted Fair Queueing [15]. However, per-flow traffic reshaping at each node may not be feasible if there are many traffic flows. In contrast, under CMS with a simple priority index assignments, Theorem 2 plus a simple example below provides the first demonstration that core-stateless schedulers can outperform WFQ schedulers.

Finally, we present a brief *ns-2* simulation study to illustrate performance differences in a scenario with six nodes and cross traffic.

A. Comparison of CMS and WFQ

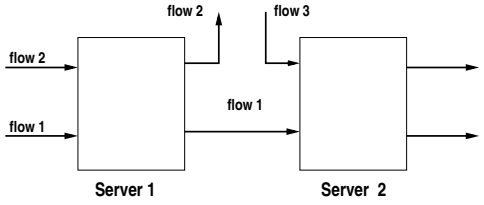


Fig. 3. Two Server System

Consider a simple system with 3 flows as described in Figure 3. Let server 1 and server 2 have capacity C and other servers have infinite capacity, each packet has size L and let $b_1(I) = b_2(I) = b_3(I) = L + \frac{C}{2}I$. The traffic parameters and the priority index assignments are summarized in Table II.

Since each packet does not suffer a queueing delay at its first hop, $D_{1,\pi(1,1)} = D_{2,\pi(2,1)} = D_{3,\pi(3,1)} = 0$ and the parameters that are needed when checking schedulability for flows at server 1 are given in Table III.

$T_{1,1_1} = T_{1,2} = \frac{L}{C}$	$Q_1 = \{1\}$	$S_1 = \{2\}$	$\Omega_1 = \emptyset$
$T_{2,2_1} = T_{2,2} = \frac{3L}{C}$	$Q_2 = \{2\}$	$S_2 = \emptyset$	$\Omega_2 = \{1\}$

TABLE III
PARAMETERS FOR SERVER 1.

According to Corollary 2, to verify flow-1 packets will miss their priority indexes at server 1 no more than $D_{1,\pi(1,2)} = \frac{L}{C}$, we need to verify that

$$\frac{\sum_{i\in Q_1}\sigma_i + \sum_{i\in\Omega_1}(\sigma_i - \gamma_i T_{i,i_1})}{C - \sum_{i\in\Omega_1}\gamma_i} + \frac{\max_{i\in S_1} l_i^{max} - CD_{1,\pi(1,2)}}{C - \sum_{i\in\Omega_1}\gamma_i} \le T_{1,2}, \quad (39)$$

and

$$\frac{\sum_{i\in Q_2}\sigma_i + \sum_{i\in\Omega_2}(\sigma_i - \gamma_i T_{i,i_1})}{C - \sum_{i\in\Omega_2}\gamma_i} + \frac{\max_{i\in S_2} l_i^{max} - CD_{1,\pi(1,2)}}{C - \sum_{i\in\Omega_2}\gamma_i} \le T_{2,2}. \quad (40)$$

To verify that flow-2 packets will miss their priority indexes at server 1 by no more than $D_{2,\pi(2,2)} = 0$, we need to ensure that

$$\frac{\sum_{i\in Q_2}\sigma_i + \sum_{i\in\Omega_2}(\sigma_i - \gamma_i T_{i,i_1})}{C - \sum_{i\in\Omega_2}\gamma_i} + \frac{\max_{i\in S_2} l_i^{max} - CD_{2,\pi(2,2)}}{C - \sum_{i\in\Omega_2}\gamma_i} \le T_{2,2}. \quad (41)$$

Using the parameters given in Table II and Table III, it is easy to verify Inequalities (39), (40), and (41). Similarly, using the parameters given in Table II and Table IV, it is easy to verify that flow-1 packets will miss their priority indexes at server 2 by no more than $D_{1,\pi(1,3)} = \frac{L}{C}$ and flow-3 packets will miss their priority indexes at server 2 by no more than $D_{3,\pi(3,2)} = 0$.

$T_{1,1_2} = T_{1,3} = \frac{L}{C}$	$Q_1 = \{1\}$	$S_1 = \{3\}$	$\Omega_1 = \emptyset$
$T_{3,3_2} = \frac{3L}{C}$	$Q_3 = \{3\}$	$S_3 = \emptyset$	$\Omega_3 = \{1\}$

TABLE IV
PARAMETERS FOR SERVER 2.

Then according to Corollary 1, we have the following end-to-end delay bounds that can be guaranteed by the CMS discipline.

$$\begin{aligned} \mathcal{D}_1 &= \max_{k\ge 1}\tau_1^k + \delta_{1,2} + \eta_{1,2} + \delta_{1,3} + \eta_{1,3} + D_{1,\pi(1,3)} \\ &= 4\frac{L}{C} \\ \mathcal{D}_2 &= 3\frac{L}{C}, \quad \mathcal{D}_3 = 3\frac{L}{C}. \end{aligned}$$

Alternatively, for WFQ, according to results provided in [21], [11], the end-to-end delay bound for flow 1 is given as:

$$\mathcal{D}_1 = \frac{L}{r_1} + \frac{L}{r_1} + \frac{L}{C} + \frac{L}{C},$$

where r_1 is the bandwidth (weight) reserved for flow 1 at servers 1,2. In order to guarantee $\mathcal{D}_1 = 4\frac{L}{C}$ for flow 1, r_1 must be C. Hence the bandwidths (weight) reserved for flow 2 at server 1 and flow 3 at server 2 must be zero. Therefore, in this case, WFQ degenerates to the strict priority service discipline (flow 1 has the highest priority). According to the result provided in [18], the minimum delay bounds guaranteed by the strict priority service discipline to flows 2 and 3 are $4\frac{L}{C}$.

B. Simulation Experiments

Throughout this paper, we have focused on schedulability conditions for coordinated schedulers. Here, we use *ns-2* simulations to illustrate potential performance improvements from coordination not only in the maximum end-to-end delay, but also in statistical delay properties.

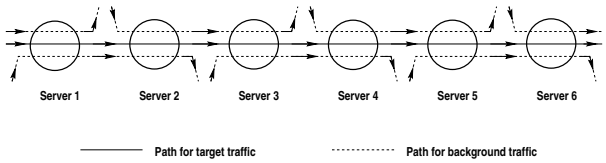


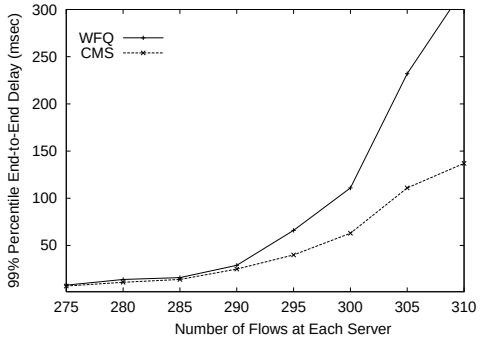
Fig. 4. Simulation Topology

We consider a simple tandem network topology as depicted in Figure 4. All link capacities are 10 Mb/sec, packet lengths are 100 bytes, and propagation delays are 0. There are several flows (varying from 25 to 60) 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 server also serves two classes of cross traffic consisting of 125 flows which traverse a single router and then exit the network, and 125 flows that traverse two routers and then exit. The cross traffic has the same characteristics as the target traffic.

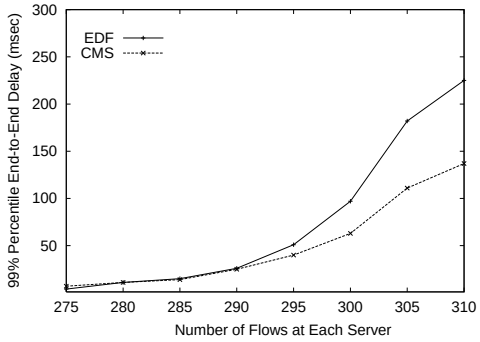
We simulate both exponential and Pareto on-off flows with on-rate 64 Kb/sec, mean on time 312 msec and mean off time 325 msec and Pareto shape parameter 1.9. The increment of the priority index at each server is 1 msec for the target traffic, 3 msec for the cross traffic with a 2-hop path, and 6 msec for cross traffic with a 1-hop path. We compare the 99-percentile end-to-end delay experienced by the target traffic for networks with CMS, EDF, and WFQ schedulers.

The simulation results are depicted in Figures 5 and 6. Each point in the figure represents the result of a 200 second *ns-2* simulation run, with averages reported over multiple simulations. The figure shows the 99.9-percentile of the end-to-end delay distribution of the target traffic as a function of the number of flows passing each server.

We make three observations regarding the figures. First, coordination has reduced the 99-percentile end-to-end delay experienced by the target traffic: for example, in Figure 5, when each server supports 295 exponential on-off traffic flows (45 target flows and 250 cross traffic flows), the end-to-end delay experienced by the target traffic is 40 msec for CMS, 66 msec for WFQ, and 51 msec for EDF. 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



(a) Comparison of CMS and WFQ



(b) Comparison of CMS and EDF

Fig. 5. Exponential On-Off Traffic

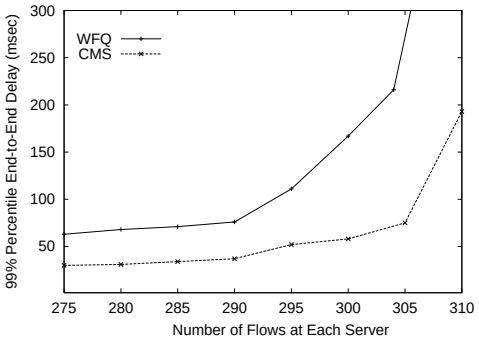
contrast, in EDF or WFQ networks, each router treats packets locally according to their arrival time, without regard to whether this arrival time is late or early.

Second, when traffic is more bursty, e.g., for Pareto on-off traffic, the advantage of CMS over WFQ or EDF is even more pronounced. For example, in Figure 6, when each server supports 295 Pareto on-off traffic flows (45 target flows and 250 cross traffic flows), the end-to-end delay experienced by the target traffic is 52 msec for CMS, 111 msec for WFQ, and 109 msec for EDF. The reason for this is that the heavy-tailed burst durations of this traffic place a heavier burden on the scheduler during periods of overload. Through inter-server coordination, CMS can better distribute this overload among network nodes and reduce a flow's end-to-end delay.

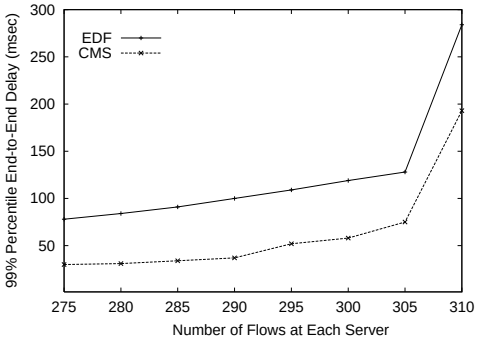
Third, the utilization at which queues build up and schedulers have an impact depends on the statistics of the arriving traffic. If the traffic is heavy-tailed e.g., Pareto ON-OFF, the coordinated scheduler significantly impacts performance for utilizations greater than 80%, i.e., more than 250 flows, whereas for Exponential On-Off traffic, it impacts at utilizations greater than 90%, i.e., more than 280 flows.

VII. Conclusion

In this paper, we derived an end-to-end schedulability criterion for a class of work conserving service disciplines termed coordinated schedulers. Exploiting the coordination property, we showed that the "essential traffic" for a flow incurs only minimal distortion at downstream nodes. Moreover, we showed that packets can be allowed to violate local priority indexes (such as local deadlines) and still satisfy an end-to-end requirement



(a) Comparison of CMS and WFQ



(b) Comparison of CMS and EDF

Fig. 6. Pareto On-Off Traffic

by "catching up" with higher priority downstream. We then devised a priority assignment scheme and showed that under the scheme, coordinated schedulers can outperform WFQ schedulers. Finally, we presented numerical and simulation results to quantify the performance gains of coordination.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees and the Associate Editor, Prof. Zhi-Li Zhang, for their constructive comments.

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. Le Boudec. Application of network calculus to guaranteed service networks. *IEEE Transactions on Information Theory*, 44(3):1087–96, May 1998.

[4] 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.

[5] C.-S Chang, R. Cruz, J.-Y Le Boudec, and P. Thiran. A min, + system theory for constrained traffic regulation and dynamic service guarantees. *IEEE/ACM Transactions on Networking*, 10(6):805–817, December 2002.

[6] 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.

[7] 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.

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

[9] R. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1048–1056, August 1995.

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

[11] N. Figueira and J. Pasquale. An upper bound on delay for the virtual clock service discipline. *IEEE/ACM Transactions on Networking*, 3(4):399–408, August 1995.

[12] V. Firoiu, J. Kurose, and D. Towsley. Efficient admission control for EDF schedulers. In *Proceedings of IEEE INFOCOM '97*, pages 310–317, Kobe, Japan, April 1997.

[13] 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.

[14] 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), September 1997.

[15] L. Georgiadis, R. Guérin, V. Peris, and K. Sivarajan. Efficient network QoS provisioning based on per node traffic shaping. *IEEE/ACM Transactions on Networking*, 4(4):482–501, August 1996.

[16] C. Li and E. Knightly. Coordinated multihop scheduling: A framework for end-to-end services. *IEEE/ACM Transactions on Networking*, 10(6):776–789, December 2002.

[17] C. Li, A. Rahal, and W. Zhao. Stability in ATM networks. In *Proceedings of IEEE INFOCOM 1997*, pages 160–167, Kobe, Japan, April 1997.

[18] J. Liebeherr, D. Wrege, and D. Ferrari. Exact admission control for networks with bounded delay services. *IEEE/ACM Transactions on Networking*, 4(6):885–901, December 1996.

[19] 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.

[20] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[21] 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.

[22] H. Sariowan, R. Cruz, and G. Polyzos. SCED: a generalized scheduling policy for guaranteeing quality-of-service. *IEEE/ACM Transactions on Networking*, 7(5):669–684, October 1999.

[23] 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.

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

[25] L. Tassiulas and L. Georgiadis. Any work-conserving policy stabilizes the ring with spatial re-use. *IEEE/ACM Transactions on Networking*, 4(2):205–208, April 1996.

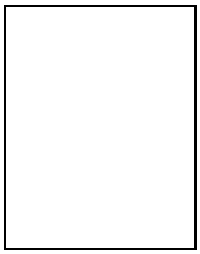
[26] H. Zhang and D. Ferrari. Rate-controlled service disciplines. *Journal of High Speed Networks*, 3(4):389–412, 1994.

[27] L. Zhang. Virtual clock: A new traffic control algorithm for packet switching networks. In *Proceedings of ACM SIGCOMM '90*, pages 19–29, Philadelphia, PA, September 1990.

[28] Z. Zhang, Z. Duan, and Y. Hou. Virtual time reference system: A unifying scheduling framework for scalable support of guaranteed services. *IEEE Journal on Selected Areas in Communications*, 18(12):2684–2695, December 2000.

[29] K. Zhu, Y. Zhuang, and Y. Viniotis. Achieving end-to-end delay bounds by EDF scheduling without traffic shaping. In *Proceedings of IEEE INFOCOM '01*, Anchorage, Alaska, April 2001.

Chengzhi Li (S'97-M'99) is currently a visiting assistant Professor at the University of Texas at Arlington. He received his B.S. degree in Applied Mathematics and M.S. degree in Operations Research from Fuzhou University and Xiamen University, China respectively. He received his Ph.D. in Computer Engineering from Texas A&M University in 1999. From 1999 to 2001, he was a postdoctoral fellow at Rice University. From 2001 to 2003, he was a research scientist at the University of Virginia. His research areas encompass wired and wireless networking.



Edward Knightly (S'91-M'96) is an associate professor in the ECE/CS Departments at Rice University. He received the B.S. degree from Auburn University in 1991 and the M.S. and Ph.D. degrees from the University of California at Berkeley in 1992 and 1996 respectively. He is an associate editor of the Computer Networks Journal and IEEE/ACM Transactions on Networking. He served as co-chair for IWQoS 1998 and as technical co-chair for IEEE INFOCOM 2005. He served on the program committee for numerous networking conferences including ICNP, INFOCOM, IWQoS, MOBICOM, and SIGMETRICS. He received the National Science Foundation CAREER Award in 1997 and the Sloan Fellowship in 2001. His research interests are in the areas of mobile and wireless networks, high-performance protocol design, quality of service, and performance evaluation.