# PERFORMANCE BOUNDS IN COMMUNICATION NETWORKS WITH VARIABLE-RATE LINKS

*Kam Lee*

Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 U.S.A.

kaml@cs.cmu.edu

***Abstract***

*In most network models for quality of service support, the communication links interconnecting the switches and gateways are assumed to have fixed bandwidth and zero error rate. This assumption of steadiness, especially in a heterogeneous internetworking environment, might be invalid owing to subnetwork multiple-access mechanism, link-level flow/error control, and user mobility. Techniques are presented in this paper to characterize and analyze work-conserving communication nodes with varying output rate. In the deterministic approach, the notion of "fluctuation constraint," analogous to the "burstiness constraint" for traffic characterization, is introduced to characterize the node. In the statistical approach, the variable-rate output is modelled as an "exponentially bounded fluctuation" process in a way similar to the "exponentially bounded burstiness" method for traffic modelling. Based on these concepts, deterministic and statistical bounds on queue size and packet delay in isolated variable-rate communication server-nodes are derived, including cases of single-input and multiple-input under first-come-first-serve queueing. Queue size bounds are shown to be useful for buffer requirement and packet loss probability estimation at individual nodes. Our formulations also facilitate the computation of end-to-end performance bounds across a feedforward network of variable-rate server-nodes. Several numerical examples of interest are given in the discussion.*

## 1. INTRODUCTION

In the new generation of integrated service packet switching networks, there is the need to provide end-to-end Quality Of Service (QOS) support to individual virtual circuits. QOS provisioning is achieved via admission control, resource allocation, and packet scheduling at intermediate network nodes along the virtual circuit. Specifically, given the traffic descriptions (throughput, burstiness) and performance requirements (delay, packet loss probability) of a packet stream, the network has to determine how much link bandwidth and buffer space should be allocated to it. A connection request will be rejected if insufficient network resources are available to meet its demands. Performance analysis associated with call admission and resource allocation can, in principle, be done by using queueing theory [9, 10, 11]. But, a queueing formulation usually requires detailed knowledge of traffic and service characteristics, and, in most cases, are mathematically intractable or computationally intensive, thus not amenable to real-time processing. A possible alternative is to compute deterministic or statistical bounds on various performance parameters. Performance bound computations for communication networks are the subjects of [4, 5, 12, 15]. In these references, the communication links are assumed to have a *fixed* capacity. In reality, though, either the *actual* or *effective* output link rate at a communication node could be time-varying. For instance:

- Shared-Media Links — Two gateways could be interconnected via a shared-media subnet, access to which is governed by a contention-based MAC protocol. In this case, the transmission bandwidth available to the gateways would vary with the aggregate traffic load in the subnet.
- Flow/Error Controlled Links — Even if there is a dedicated fixed-speed point-to-point link connecting two nodes, the effective transmission rate could fluctuate as a result of some data-link-level flow control or error control mechanisms.
- Mobile Links — The capacity of a mobile channel over which a base station communicates with a mobile host might vary with time owing to multipath fading, environmental interference, and distance variations. In addition, the bandwidth allocated to a mobile connection could be subjected to changes during handoffs as the mobile host roams from cell to cell.

Variations in link capacity would affect the queue length and delay performance of network traffic. Therefore, it is useful to devise bounding schemes to characterize service rate fluctuations.

In this paper, techniques are developed to compute performance bounds for traffic flows across a connection-oriented packet switching network with variable-rate links. As shown in Figure 1, we model each node-link pair along a virtual circuit as a queueing unit consisting of a buffer and a server. The service rate of the server reflects the physical link characteristics as well as other protocol-dependent link control effects. Here, the term "server-node" is used to refer to a buffer-server pair, and a multiplexer is synonymous with a multiple-input server-node. Section 2 contains a brief review of related work. Deterministic and statistical characterization and analysis of an isolated work-conserving variable-rate server are presented and substantiated with numerical examples in Sections 3 and 4. Estimation of packet loss probability from queue size bounds is the focus of Section 5. Section 6 is concerned with the end-to-end performance bounds along a multiple-node virtual circuit. Following the discussions in Section 7, conclusions and possibilities for future work are given in Section 8. Appendix A and B contain formal proof of the theorems presented in this paper.

## 2. RELATED WORK

The concept of computing performance bounds for traffic flows in packet switching networks has drawn a lot of attention in recent years. Applying the "burstiness constraint" to source traffic, Cruz [4, 5] derived absolute delay bounds for packet streams traversing

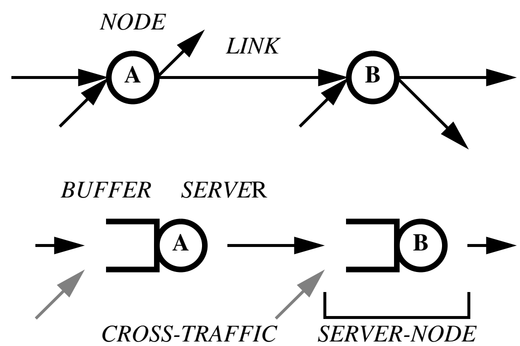


Figure 1: Physical network representation and logical virtual circuit model.

a network of work-conserving fixed-rate multiplexers. Parekh & Gallager [13, 14] proved that leaky-bucket-controlled sessions, which satisfy the "burstiness constraint," would experience bounded backlog and delay under Generalized Processor Sharing (GPS) scheduling. The problem of finding deterministic end-to-end delay and buffer occupancy bounds in ATM networks with rate-controlled non-work-conserving servers was tackled by Banerjea & Keshav [1]. In their analysis, a different parametric constraint was used for traffic characterization.

While Cruz tried to guarantee that "the delay of a session *i* packet in node *j* is less than *x*," Kurose [12] was interested in "the probability that the delay of a session *i* packet in node *j* being larger than *x* is less than *y*." He solved the problem by stochastically bounding the traffic stream with a series of random distributions. But this bounding technique still required the source traffic to have bounded burst size, a condition not met by most well-known arrival processes such as the Bernoulli and Poisson processes. To overcome this limitation, Yaron & Sidi [15] introduced the idea of characterizing traffic flows as "exponentially bounded burstiness" (EBB) processes, and were able to derive statistical bounds on queue size and delay in FCFS system. Moreover, they [16] applied their formulation to analyze the performance of a GPS server. Zhang [18] also adopted the EBB model to study the statistical behaviour of GPS scheduling. By introducing the notion of "feasible partitions," he obtained performance bounds which are generally tighter than those in [16]. Yates et al. [17] performed simulations to determine the per-session end-to-end delay distributions in a connection-oriented network using FCFS multiplexing under wide-area traffic conditions. Their results indicated that the worst case delay bounds predicted from Cruz's and Parekh's theories tend to be overly conservative, and as a result might lead to poor network utilization when used for call admission purposes.

In all the aforementioned work, the network nodes are assumed to be interconnected by perfectly reliable fixed-rate communications links. Chang [3] developed a theory of Minimum Envelope Process, which, incidentally, can be viewed as a generalization of the techniques proposed by Cruz, and Yaron & Sidi. While the focus of [3] was also on fixed-rate queueing node, Chang did mention the possibility of extending his formulation to cover the case of time-varying link capacity. Recently, Cruz [6, 8] has introduced the notion of "service burstiness" to deterministically model non-work-conserving service disciplines. In addition, he applied Chang's method to stochastically bound the service rate of a queueing node [7]. In Section 7, we will contrast our framework to these recent developments.

## 3. DETERMINISTIC SERVERS

Let $R(t)$ denote the instantaneous data rate of a traffic stream. In [4], the traffic stream is said to be "burstiness-constrained" (BC), $R \sim (\sigma, \lambda)$, if

$$\int_a^b R(t)\,dt \le \lambda(b-a) + \sigma \tag{3.1}$$

or, in discrete-time:

$$\sum_{n=a+1}^{b} R(n) \le \lambda(b-a) + \sigma \tag{3.2}$$

In other words, the amount of data generated by a BC input over any time interval $T$ is upper bounded by $\lambda T + \sigma$; $\lambda$ can be interpreted as the long term average data rate. Furthermore, it is proven that when all the inputs to a fixed-rate node are BC, the output delay is bounded, and the output stream is also BC. In this section, we will state, in the form of several theorems, that similar properties exist for a variable-rate node provided that it satisfies the "fluctuation constraint," introduced below. For simplicity of exposition, we assume that $L/v \ll 1$, where $L$ is the packet size, and $v$ is the average physical speed of the input or output link. Formal proof of the theorems is given in Appendix A.

*Definition 3.1*: **Server Characterization** — Let $C(t)$ denote the instantaneous output transmission capacity of a variable-rate node. Then the node is said to be "fluctuation constrained" (FC), $C \sim (\delta, \mu)$, if[1]:

$$\int_a^b C(t)\,dt \ge [\mu(b-a) - \delta]^+ \tag{3.3}$$

or, in discrete-time:

$$\sum_{n=a+1}^{b} C(n) \ge [\mu(b-a) - \delta]^+ \tag{3.4}$$

❐

Remarks:

- If a constant bit stream is fed into a $(\delta, \mu)$-server at rate $\mu$, then the maximum backlog in the input buffer will never exceed $\delta$ bits; $\mu$ can be interpreted as the long term average service rate.
- The outgoing link of a FC node is said to be *deterministic* in the sense that the amount of data that the node can transmit over any time interval is lower bounded.
- As a special case, a node of fixed rate $\mu$ is FC: $C \sim (0, \mu)$.

*Theorem 3.1*: **Output Characteristics for Single-Input** — An input traffic stream $R \sim (\sigma, \lambda)$ enters a variable-rate node. If the maximum delay the traffic stream experiences in passing through this node is $D < \infty$, then the corresponding output stream will have a rate of $S \sim (\sigma + \lambda D, \lambda)$.

❐

Remarks:

- This input-output relationship is independent of the service-rate characteristics. In other words, the output stream is BC as long as the input itself is BC, and the maximum output delay is bounded.
- Later in this section, we will indicate that, under certain conditions, the delay of a BC input across a FC node is indeed bounded, and therefore, the resulting output stream is BC.
- This theorem was first derived by Cruz [4], and is included here for sake of completeness.

1. $[x]^+ = \max(0, x)$

*Theorem 3.2*: **Output Characteristics for Multiple-Input** — $N$ input traffic streams $R_i \sim (\sigma_i, \rho_i)$, $i = 1,...,N$, enter a variable-rate node. If the maximum delay stream $R_i$ experiences in passing through this node is $D_i < \infty$, then the corresponding output stream $S_i$ will have a rate of $(\sigma_i + \lambda_i D_i, \lambda_i)$. In addition, the aggregate output $S = \Sigma S_i$ also satisfies the burstiness constraint with rate $(\Sigma(\sigma_i + \lambda_i D_i), \ \Sigma\lambda_i)$

❐

Remarks:

- Again, this input-output relation is independent of the service-rate characteristics. The aggregate output as well as its individual components from a multiplexer is BC as long as the inputs are BC, and their delay is bounded.
- Later in this section, we will indicate that, under certain conditions, the delay of BC inputs across a FC node is indeed bounded, and therefore, the resulting individual output streams as well as the aggregate output are also BC.

*Theorem 3.3*: **Queue Length Bound for Single-Input Server** — Assume that a single input stream $R \sim (\sigma, \lambda)$ enters an infinite buffer work-conserving variable-rate server $C \sim (\delta, \mu)$. If $\lambda < \mu$, then the queue length (or backlog) in the buffer is bounded by:

$$Q \le \sigma + \frac{\lambda}{\mu}\delta \tag{3.5}$$

❐

Remarks:

- This theorem is independent of the queueing discipline exercised in the node.
- The stability criterion is $\lambda < \mu$, i.e the average input rate should be kept below the average service rate. Violation of this condition might lead to unbounded queue size.
- In the special case of $C \sim (0, \mu)$, (3.5) reduces to $Q \le \sigma$.
- Since $\lambda < \mu$, (3.5) implies that, in general, $Q \le \sigma + \delta$.

*Theorem 3.4*: **Queue Length Bound for Multiple-Input Server** — Assume that $N$ input streams, $R_i \sim (\sigma_i, \lambda_i)$, $i = 1, ..., N$, enter an infinite buffer work-conserving variable-rate server $C \sim (\delta, \mu)$. If $\Sigma\lambda_i < \mu$, then the queue length (or backlog) in the buffer is bounded by:

$$Q \le \Sigma\sigma_i + \frac{\Sigma\lambda_i}{\mu}\delta \tag{3.6}$$

❐

Remarks:

- This theorem, a generalization of Theorem 3.3, is applicable to any non-blocking (i.e. infinite-buffer) FC server-node, regardless of its queueing discipline.
- To maintain stability (i.e. bounded queue size), the aggregate average input rate must be less than the average link capacity.

*Theorem 3.5*: **Delay Bound for Single-Input General Service Server** — Assume that a single input stream $R \sim (\sigma, \lambda)$ enters an infinite buffer work-conserving variable-rate server $C \sim (\delta, \mu)$ with arbitrary queueing discipline (e.g. non-FCFS). If $\lambda < \mu$, then the delay of any data bit across this node is upper bounded by:

$$D \le \frac{\sigma + \delta}{\mu - \lambda} \tag{3.7}$$

❐

Remarks:

- It can easily be shown that this theorem also holds when there are more than one BC input streams. In that case, the delay bound can be computed from (3.7) by replacing $\sigma$ and $\lambda$ with $\Sigma\sigma_i$ and $\Sigma\lambda_i$, respectively.
- No particular queueing discipline is assumed in the derivation of (3.7). Otherwise, a tighter bound can possibly be obtained by taking into account of the queueing mechanism in the node. This is illustrated in the following theorem for the case of a FCFS variable-rate server-node.
- In the special case of $C \sim (0, \mu)$, (3.7) reduces to $D \le \sigma/(\mu - \lambda)$, in agreement with Equation 4.13 in [4].

*Theorem 3.6*: **Delay Bound for Single-Input FCFS Server** — Assume that an input stream $R \sim (\sigma, \lambda)$ enters an infinite buffer variable-rate server $C \sim (\delta, \mu)$ with FCFS queueing. If $\lambda < \mu$, then the delay of any data bit across this node is upper bounded by:

$$D = \frac{\sigma + \delta}{\mu} \tag{3.8}$$

❐

Remarks:

- Taking the queueing mechanism into consideration, this bound is tighter than the one specified by (3.7).
- In the special case of $C \sim (0, \mu)$, one gets $D \le \sigma/\mu$, in agreement with Equation 4.1 in [4].
- This theorem can be augmented to accommodate the case of multiple BC inputs by simply replacing $\sigma$ and $\lambda$ in (3.8) by $\Sigma\sigma_i$ and $\Sigma\lambda_i$, respectively.

*Example 3.1*: **Leaky-Bucket-Regulated Server** — A leaky-bucket type of arrangement may be used to ensure that the service characteristic of a server-node is FC. Token-bits enter the bucket at a constant rate $\mu$, and the bucket can hold a maximum of $\delta$ token-bits. Whenever the server *attempts* to retrieve a data bit from its input buffer, one token-bit, if present, will be drained from the bucket. If the service rate is controlled such that overflow never occurs in the bucket, then the server is FC ~ $(\delta, \mu)$.

*Example 3.2*: **Sinusoidal Link** — Suppose that the bit rate $C(t)$ of the output link of a communication node varies sinusoidally with time:

$$C(t) = \mu(1 + \sin\omega t) \tag{3.9}$$

Then,

$$\int_a^b C(t)\,dt \le \mu(b - a) - \frac{2\mu}{\omega} \tag{3.10}$$

So this variable-rate link is FC-compliant: $C \sim (2\mu/\omega, \mu)$. In fact, any server with periodically varying service rate can be characterized as a fluctuation-constraint process. Given next is an example of a communication node whose *effective* output capacity is periodic.

*Example 3.3*: **Go-Back-N Window-Based Flow-Controlled Link** — Suppose that gateway A is connected to gateway B via a perfectly reliable point-to-point transmission link which has a fixed raw speed of $s$ bits/msec. Packet flows from A to B are regulated by using the Go-Back-N sliding window protocol [2]. Assume that all packets are $p$ bits long, and let $w$ denote the window size, and $d$ the round trip delay between the gateways. Then, as a result of the flow control mechanism, the effective full service rate $C(t)$ of gateway A is periodic, alternating between busy and idle periods, with the average bit rate $\mu$ given by:

$$\mu = \frac{wp}{wp + ds}s \tag{3.11}$$

It follows that:

$$\int_{\Delta t} C(t)\,dt \ge \mu\Delta t - \mu d \tag{3.12}$$

So the gateway node is FC ~ $(\mu d, \mu)$. Simulations are performed to determine the delay distribution of one or more leaky-bucket filtered Poisson streams passing through an infinite buffer FCFS node whose output link is flow controlled. The flow control scheme is identical to the one just described, with $w = 80, d = 20p/s, p/s = 8$. Hence $C \sim (16$ packets, $0.1$ packets/msec). Each of the original Poisson sources has a mean rate of *0.01* packets/msec, and is fed into a leaky-bucket filter which has a bucket size of *10* and generates tokens at *0.012* tokens/msec. One token is consumed whenever a packet is admitted. The output of the filter is therefore BC-compliant: $R \sim (10$ packets, $0.010$ packets/msec$)$. In Table 1, the 99-percentiles of the delay distribution under various input load conditions are compared to the predicted deterministic delay bounds (i.e. Theorem 3.6).

| Load ($\Sigma\lambda/\mu$) | 99-percentile Delay (msec) | Deterministic Delay Bound (msec) |
|---|---|---|
| 0.1 | 96 | 260 |
| 0.3 | 150 | 460 |
| 0.5 | 163 | 660 |
| 0.7 | 175 | 860 |
| 0.9 | 276 | 1060 |

Table 1: Deterministic Delay Bounds.

## 4. STOCHASTIC SERVERS

The deterministic bounding scheme formulated in the previous section is applicable to a variable-rate node whose average effective output capacity over any time interval $T$ will never drop below $\mu$ by more than $\delta$. However, such absolute guarantee on service capacity is not always possible. For example, under some contention-based link access protocol such as CSMA, the channel access time is, in principle, unbounded. Similarly, there is no limit on the time taken to successfully deliver a packet across a noisy cellular link. So, a probabilistic bounding scheme is needed for these situations. Yaron & Sidi [15] introduced the idea of using decaying exponentials for statistical traffic characterization. First, a stochastic process $W(t)$ is "exponentially bounded" (EB), $W \sim (A, \alpha)$, if:

$$Pr\{W(t) \geq \sigma\} \leq Ae^{-\alpha\sigma} \tag{4.1}$$

Second, a traffic stream with data rate $R(t)$ is said to have "exponentially bounded burstiness" (EBB), $R \sim (\lambda, A, \alpha)$, if:

$$Pr\{\int_a^b R(t)\,dt \geq \lambda(b-a) + \sigma\} \leq Ae^{-\alpha\sigma} \tag{4.2}$$

or, in discrete-time:

$$Pr\left\{\sum_{n=a+1}^{b} R(n) \geq \lambda(b-a) + \sigma\right\} \leq Ae^{-\alpha\sigma} \tag{4.3}$$

We develop here the concept of an "exponentially bounded fluctuation" (EBF) process, which is the analog of the EBB model for statistical service characterization. In [15], it is shown that if all the input stream to an infinite buffer fixed-rate node are EBB, then the resulting output stream is EBB, and both the buffer occupancy and traffic delay are EB. Presented in this section are several theorems which claim that similar relationships hold when one or more EBB inputs enters a EBF-compliant variable-rate server-node. Detailed derivation of the theorems is given in Appendix B.

*Definition 4.1*: **Server Characterization** — A communication node with time varying output link capacity $C(t)$ is said to have "exponentially bounded fluctuation" (EBF), $C \sim (\mu, B, \beta)$, if:

$$Pr\{\int_a^b C(t)\,dt \leq \mu(b-a) - \delta\} \leq Be^{-\beta\sigma} \tag{4.4}$$

or, in discrete-time:

$$Pr\left\{\sum_{n=a+1}^{b} C(n) \leq \mu(b-a) - \delta\right\} \leq Be^{-\beta\sigma} \tag{4.5}$$

❐

Remarks:
- In the limit of $B \to 0$ and $\beta \to \infty$, the link capacity becomes deterministically bounded.
- A link with fixed-rate $\mu$ is EBF $\sim (\mu, 0, \infty)$.
- An EBF server is *stochastic* in the sense that it is only the service-rate distribution that is bounded.

*Theorem 4.1*: **Output Characteristics for Single-Input** — Suppose that an EBB traffic stream $R \sim (\lambda, A, \alpha)$ enters an infinite buffer EBF server $C \sim (\mu, B, \beta)$. If $\lambda < \mu$, then the output stream $S$ is EBB:

$$S \sim \left(\lambda, \frac{A+B}{1-e^{-\zeta(\mu-\lambda)}}, \zeta\right) \tag{4.6}$$

where $\zeta = (\alpha\beta)/(\alpha+\beta)$.

❐

Remarks:
- Comparing (4.6) to the result given by Proposition 5 in [15], an isolated variable-rate server is equivalent to the combination of a fixed-rate server $\mu$ and a *virtual* EBB input $\sim (0, B, \beta)$.

*Theorem 4.2*: **Output Characteristics for Multiple-Input** — Suppose that $N$ EBB traffic streams $R_i \sim (\lambda_i, A_i, \alpha_i)$, $i = 1, ..., N$, enter an infinite buffer EBF server $C \sim (\mu, B, \beta)$. If $\Sigma\lambda_i < \mu$, then each individual output stream $S_i$ is also EBB:

$$S_i \sim \left(\lambda_i, \frac{\Sigma A_i + B}{1-e^{-\zeta(\mu-\Sigma\rho_i)}}, \zeta\right) \tag{4.7}$$

where

$$\frac{1}{\zeta} = \sum_{i=1}^{N}\frac{1}{\alpha_i} + \frac{1}{\beta} \tag{4.8}$$

Furthermore, the aggregate output stream $S = \Sigma S_i$ is EBB as well:

$$S \sim \left(\Sigma\lambda_i, \frac{\Sigma A_i + B}{1-e^{-\zeta(\mu-\Sigma\rho_i)}}, \zeta\right) \tag{4.9}$$

where $\zeta$ is a function of the $\alpha_i$'s and $\beta$, as given by (4.8).

❐

Remarks:
- For N = 2, in the limit of $B \to 0$ and $\beta \to \infty$, (4.9) is identical to the results derived in Proposition 5 of [15] for a two-input fixed-rate multiplexer.

*Theorem 4.3*: **Queue Length in Single-Input Server** — Suppose an EBB input stream $R \sim (\lambda, A, \alpha)$ enters an infinite buffer EBF server $C \sim (\mu, B, \beta)$. If $\lambda < \mu$, then the queue length $Q(t)$ is EB:

$$Q \sim \left(\frac{A+B}{1-e^{-\zeta(\mu-\lambda)}}, \zeta\right) \tag{4.10}$$

where $\zeta = (\alpha\beta)/(\alpha+\beta)$.

❐

Remarks:
- In the limit of $B \to 0$ and $\beta \to \infty$, (4.10) becomes:

$$Q \sim \left( \frac{A}{1 - e^{-\alpha(\mu-\lambda)}}, \alpha \right) \tag{4.11}$$

which is consistent with Theorem 1 in [15].

- Imagine that an EBF variable-rate node $C \sim (\mu, B, \beta)$ has $N$ EBB input streams $R_i \sim (\lambda_i, A_i, \alpha_i)$, $i = 1,...,N$. If $\Sigma\lambda_i < \mu$, then it can readily be shown that the queue length $Q(t)$ is EB:

$$Q \sim \left( \frac{\Sigma A_i + B}{1 - e^{-\zeta(\mu - \Sigma\lambda_i)}}, \zeta \right) \tag{4.12}$$

where $\zeta$ is as given in (4.8).

- For N = 2, in the limit of $B \to 0$ and $\beta \to \infty$, (4.12) is equivalent to the result derived in Proposition 6 of [15] for a two-input fixed-rate multiplexer.
- Stability Criterion: As long as $\Sigma\lambda_i < \mu$, this queueing system is stochastically stable in the sense that the tail of the backlog distribution is bounded.

Theorem 4.4: **Delay Bound in Single-Input General Service Server** — Suppose that an EBB traffic stream $R \sim (\lambda, A, \alpha)$ enters an infinite buffer EBB server $C \sim (\mu, B, \beta)$. If $\lambda < \mu$, then the traffic delay $D(t)$ across this node is EB:

$$D \sim \left( \frac{A + B}{1 - e^{-\zeta(\mu-\lambda)}}, \zeta(\mu - \lambda) \right) \tag{4.13}$$

where $\zeta = (\alpha\beta)/(\alpha+\beta)$.

❒

Remarks:

- With proper substitutions into (4.13), a similar expression can be established for the case of multiple-input.
- Theorem 4.4 makes no assumption about the queueing or multiplexing mechanism within the node. Knowledge of the scheduling discipline might be exploited to obtain a tighter bound.

Theorem 4.5: **Delay Bound in Single-Input FCFS Server** — An EBB traffic stream $R \sim (\lambda, A, \alpha)$ is fed into an infinite buffer FCFS variable-rate server node $C \sim (\mu, B, \beta)$. If $\lambda < \mu$, then the traffic delay $D(t)$ across this node is EB:

$$D \sim \left( \frac{A + B}{1 - e^{-\zeta(\mu-\lambda)}}, \zeta\mu \right) \tag{4.14}$$

where $\zeta = (\alpha\beta)/(\alpha+\beta)$.

❒

Remarks:

- Since the decay factor in (4.14) is larger than the one in (4.13), this tail bound is tighter than the one given in the previous theorem.
- As before, this theorem can readily be extended to cover the case of multiple-input.
- In our analysis, independence between the traffic process and the server process is not assumed. If they are indeed independent, it is possible to derive a tighter set of queue size and delay bounds.

*Example 4.1*: **Exponential Server** — Suppose that the service time of fixed-size packets in a node is exponentially distributed with mean $\lambda$. Then the probability of $n$ packets departing from the node in a time interval $T$ is given by the Poisson formula:

$$Pr(n) = \frac{(\lambda T)^n}{n!} e^{-\lambda T} \tag{4.15}$$

Invoking Markov's Inequality [10], one gets:

$$\begin{aligned} Pr(n \le [(\lambda - \varepsilon)T - \sigma]) &= Pr\left( e^{-\theta n} \ge e^{-\theta[(\lambda-\varepsilon)T - \sigma]} \right) \\ &\le E\left[e^{-\theta n}\right] \cdot e^{\theta[(\lambda-\varepsilon)T-\sigma]} \\ &= e^{-\lambda T(1 - e^{-\theta})} \cdot e^{\theta[(\lambda-\varepsilon)T-\sigma]} \\ &= e^{-[\lambda(1-\theta-e^{-\theta}) + \varepsilon\theta]T} \cdot e^{-\theta\sigma} \end{aligned} \tag{4.16}$$

with $\varepsilon > 0$, and $\theta > 0$. So, the exponential server[2] is EBF $\sim (\lambda - \varepsilon, 1, \theta)$, where $\varepsilon$ and $\theta$ satisfies:

$$\lambda\left(1 - \theta - e^{-\theta}\right) + \varepsilon\theta = 0 \tag{4.17}$$

*Example 4.2*: **Round-Robin Polling Access** — A slotted communication channel is shared among N server-nodes. The nodes are polled in round-robin fashion on a slot-by-slot basis. If the packet queue of the selected node is not empty, it can transmit one packet in the time slot. Otherwise, the time-slot will be allocated to the next node in sequence that has a packet to send. This approach ensures that transmission will occur in any slots as long as outstanding packets are present in the system. In addition, a node is guaranteed to gain access to the channel within N slots' time. (This is essentially a centralized version of the well-known token-passing scheme [2].) Now suppose that the packet size is fixed, and that the polling latency is negligible compared to the slot-time. Then the minimum and maximum service time of a head-of-line packet in a node are 1 slot and N slots, respectively. Let $C(i)$ denote the effective service rate of a node in slot $i$, and further assume that the packet service time, $x$, is an i.i.d. random variable, *uniformly* distributed between 1 and N. Then, using the Markov Inequality, one obtains:

$$\begin{aligned} Pr\{\sum_{\Delta t} C(n) \le \mu\Delta t - \delta\} &\le Pr\left\{ \sum_{i=1}^{\mu\Delta t - \delta} x_i \ge \Delta t \right\} \\ &\le \left[E\left(e^{\theta x}\right)\right]^{\mu\Delta t - \delta} \cdot e^{-\theta\Delta t} \end{aligned} \tag{4.18}$$

with $\theta > 0$. Letting:

$$E\left(e^{\theta x}\right) = \frac{e^{\theta}\left[e^{\theta N} - 1\right]}{N\left[e^{\theta} - 1\right]} \le e^{\beta} \tag{4.19}$$

for some $\beta > 0$, (4.18) implies:

$$Pr\{\sum_{\Delta t} C(n) \le \mu\Delta t - \delta\} \le e^{(\beta\mu - \theta)\Delta t} e^{-\beta\delta} \tag{4.20}$$

Therefore, the service rate of a node in this round-robin polling system is EBF $\sim (\mu, 1, \beta)$ if there exist $\mu$ and $\beta$ such that:

$$\frac{e^{\beta\mu}\left[e^{\beta\mu N} - 1\right]}{N\left[e^{\beta\mu} - 1\right]} = e^{\beta} \tag{4.21}$$

For instance, with N = *19* and $\mu = 0.08$, $\beta$ is calculated to be *1.82*.

*Example 4.3*: **Stop & Wait Error Control** — $L$-bit long packets are transmitted from node A to node B across an unreliable slotted channel that has a bit error rate of $b$. Stop & Wait ARQ strategy [2] is adopted for error control purpose. Upon reception of a packet, node B will send node A a NACK if the packet is corrupted, otherwise an ACK will be sent instead. Any negatively acknowledged packet will be retransmitted by A in subsequent time slots until successful delivery. Assume that no more than one packet can be sent in each time slot, and that the acknowledgment will always be received before the end of the current time slot. Under these condi-

2. It can also be shown that a Poisson arrival process of rate $\lambda$ is EBB $\sim (\lambda + \varepsilon, 1, \theta)$, where $\lambda(1 + \theta - e^{\theta}) + \varepsilon\theta = 0$.

tions, the effective service rate $C(n)$ of node A can be modelled as a random Bernoulli process with parameter $p = (1 - b)^L$— the packet success rate. Then, for any $\varepsilon > 0$,

$$Pr\{\sum_{\Delta t} C(n) \le (p - \varepsilon)\Delta t - \delta\} \tag{4.22}$$
$$= Pr\{\sum_{\Delta t} \overline{C}(t) \ge (q + \varepsilon)\Delta t + \delta\}$$

where $\overline{C}$, the complementary process of $C$, is itself Bernoulli with parameter $q = (1 - p)$. Then according to Proposition 3 in [15], $\overline{C}(n)$ is EBB ~ ($q + \varepsilon$, 1, $\beta$), and therefore $C(n)$ is EBF ~ ($p - \varepsilon$, 1, $\beta$), for some $\beta$. The actual delay characteristics of a Bernoulli process (with parameter *0.05*) across this error control link (geometric server) are determined via simulations for the case of $p = 0.5$. Shown in Figure 2 are the results, along with analytical delay bound curves corresponding to three different EBF characterizations of link characteristics. The difference between the bound obtained for general service and the one for FCFS service are depicted in Figure 3. In these two graphs, the Bernoulli process is characterized as EBB ~ (*0.15, 1, 2.16*).

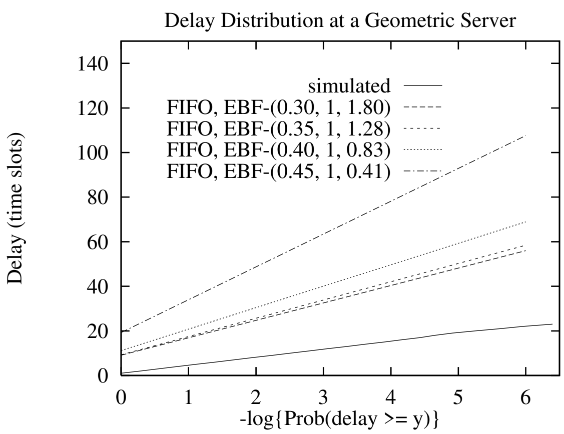


Figure 2: Simulated vs. analytical delay distribution of a Bernoulli input at a geometric server.

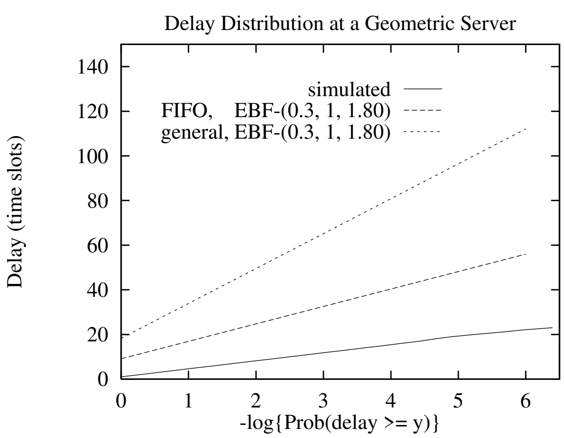


Figure 3: Delay bounds for general and FIFO queueing at a geometric server.

## 5. BOUNDS ON PACKET LOSS PROBABILITY

So far in our discussion, the emphasis has been on delay bound computation assuming the network node has an infinite amount of buffer space. In reality, though, a node has finite buffer capacity, and therefore packets may be dropped due to buffer overflow. Packet loss probability could be an important QOS measure, especially for applications that are loss-sensitive.

For deterministic server-nodes, packet drop is prevented if the amount of buffering available exceeds the queue size bound given by Theorems 3.3 or 3.4. So this allows one to determine the buffer allocation for traffic streams that requires zero loss.

For stochastic server-nodes, an estimated upper bound on packet loss probability may be obtained by applying Theorem 4.3. Specifically,

$$Pr\{loss | (buffer = B)\} \le Pr\{queue > B\} \le Ge^{-\eta B} \tag{5.1}$$

for some $G$ and $\eta$. Note that complete buffer sharing is assumed here when more than one input streams are present.

*Example 5.1*: **Finite Buffer Geometric Server** — To get an idea of how good the estimation is, the packet loss probability of a Bernoulli stream entering a geometric server is determined via simulations as a function of buffer size. The results are compared in Figure 4 to the performance estimates based on different EBF characterizations of the server. As before, the Bernoulli input has parameter *0.05* and is modelled as EBB ~ (*0.15, 1, 2.16*).

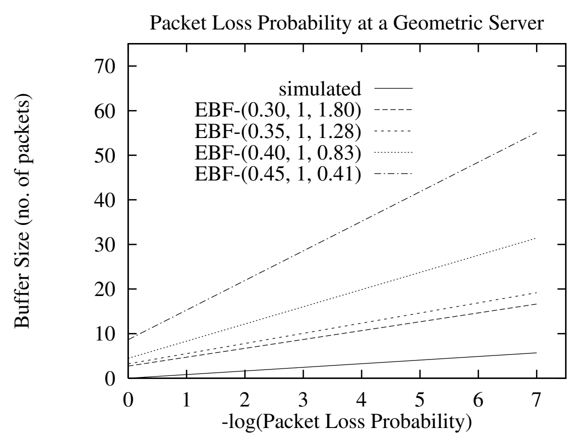


Figure 4: Estimation of packet loss probability using the EBB traffic model and EBF server model.

## 6. END-TO-END PERFORMANCE BOUNDS

Having developed techniques to compute deterministic and statistical performance bounds at a single variable-rate server-node, we are now concerned with the end-to-end performance of a traffic stream flowing across a feedforward network of variable-rate server-nodes. The strategy is to first identify any cross-traffic along its virtual circuit. Then, based on the appropriate characterizations of the source traffic, cross traffic, and communication servers, one can calculate the local performance bounds at each node using the techniques discussed earlier, proceeding in order from the first to the last node. In the deterministic model, the end-to-end delay bound ($D_{ee}$) can simply be calculated as the sum of all local delay bounds ($D_i$):

$$D_{ee} = \sum_{i=1}^{N} D_i \qquad (6.1)$$

In the stochastic model, to compute the end-to-end probabilistic delay bound, one can make use of the fact that the sum of multiple EB processes is again an EB process [15], regardless of their statistical dependencies. Specifically, if $S(t)$ is the sum of $N$ EB processes $R_i(t) \sim (A_i, \alpha_i)$, $i = 1, ..., N$, then $S(t)$ is also EB[3]:

$$S \sim \left( \sum A_i, \left( \sum \frac{1}{\alpha_i} \right)^{-1} \right) \qquad (6.2)$$

Finally, the overall packet loss probability ($P_{ee}$) of a connection can be estimated as:

$$P_{ee} \le 1 - \prod_{i=1}^{N} (1 - P_i) \qquad (6.3)$$

where $P_i$ is the packet loss probability at node $i$.

*Example 6.1*: **End-to-End Delay Across Two Stochastic Servers** — An EBB traffic stream $R \sim (\lambda, \sigma)$ flows across two packet switches in tandem. Each switch has infinite buffer and does FCFS queueing. Assume that all the communication links along its path have the same raw speed and bit error rate. Data link level error control is handled with the stop & wait protocol described in Example 4.3. For simplicity, further assume the absence of cross traffic in this circuit. Using the same numerical values as those in Example 4.3, simulations are done to obtain the end-to-end delay distribution. The results are graphed in Figure 5. Also shown are the delay bounds computed by recursively applying Theorems 4.1 and 4.5. The initial input is taken to be EBB ~ (*0.15, 1, 2.16*), and each switch is characterized as EBF ~ (*0.30, 1, 1.80*). The delay functions at switch 1 and switch 2 are calculated to be EB ~ (*14.6, 0.29*) and EB ~ (*172, 0.19*), respectively. Then, according to (6.2), the end-to-end delay distribution $D_{ee}(t)$ satisfies:

$$Pr\{D_{ee}(t) \ge \sigma\} \le 187 e^{-0.116\sigma} \qquad (6.4)$$

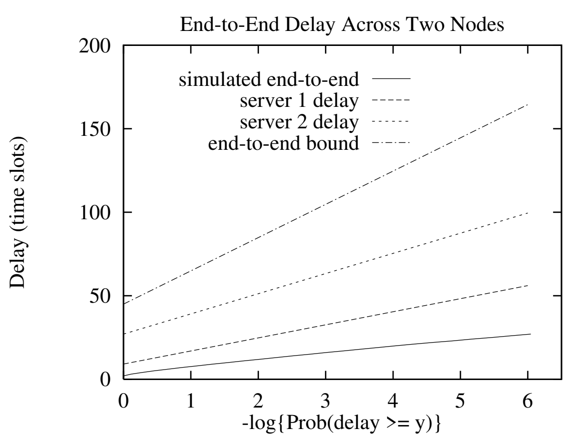


Figure 5: Delay distributions along a two-node virtual circuit.

3. If the added processes are independent, then a larger decay factor can be attained.

**7. DISCUSSIONS**

Equivalence and Duality

In our framework, as remarked in Section 3 and 4, a fixed-rate server of rate μ is merely an instance of a variable-rate server of FC ~ (0, μ) or EBF ~ (μ, 0, ∞). More interestingly, as far as queue/delay bounds are concerned, a variable-rate node of FC ~ (δ, μ) or EBF ~ (μ, B, β) is equivalent to a fixed-rate node of rate μ *coupled with* a zero-mean virtual input stream of BC ~ (δ, 0) or EBB ~ (0, B, β). Conversely, a fixed-rate node of rate μ subjected to input cross-traffic of BC ~ (σ, λ) or EBB ~ (λ, A, α) can be considered as a variable-rate node of FC ~ (σ, μ–λ) or EBF ~ (μ–λ, A, α) *without* input cross-traffic. This equivalence and duality relation is depicted in Figure 6.

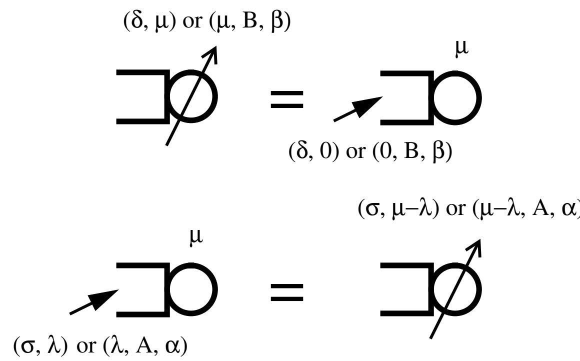


Figure 6: Transformation between variable-rate system and fixed-rate system.

Related Concepts

Recently, Cruz [6, 7, 8] has introduced the concept of "service burstiness" and "server availability process." The former enables the deterministic characterization of non-work-conserving packet scheduling disciplines that provides service guarantees to individual traffic streams. Service burstiness [6] is defined in terms of the output rate ($S$) of the stream. It basically imposes a lower bound on S over an interval of time. In contrast, the concept of "fluctuation constraint" devised in this paper characterizes the server in terms of its maximum output capacity ($C$). In general, $\int S \le \int C$; equality holds only if the server is occupied over the entire interval of observation. This approach decouples the server characteristics from input traffic characteristics, and can readily accommodate the case of multiple-input.

On the other hand, the concept of "server availability process" [7], like the concept of "exponentially bounded fluctuation," stochastically characterizes the available capacity ($C$) of a server. Using this concept, Cruz was able to determine performance bounds on the total backlog and end-to-end delay in a multi-hop path, assuming statistical independence among input and server processes. The formulation of "server availability process" is based on Chang's scheme [3], which is slightly different from the scheme we employ. Besides, our framework centers on single-node analysis without assumption of independence, and takes into account of cross-traffic.

Tightness of Bounds

In order to measure the "tightness" of the stochastic bound on parameter $W$, let's define the "*over-estimate factor*" simply as:

| Characterization | Deterministic | Statistical |
|---|---|---|
| Input (R) | $\int R \le \lambda\Delta t + \sigma$ | $Pr\,\{\int R \ge \lambda\Delta t + \sigma\} \le A\, e^{-\alpha\sigma}$ |
| Server (C) | $\int C \ge \mu\Delta t - \delta$ | $Pr\,\{\int C \le \mu\Delta t - \delta\} \le B\, e^{-\beta\delta}$ |
| Output (S) | $\int S \le \lambda\Delta t + f(\lambda, \sigma, \mu, \delta)$ | $Pr\,\{\int S \ge \lambda\Delta t + \eta\} \le g(\lambda, \alpha, \mu, \beta)e^{-h(\lambda, \alpha, \mu, \beta)\eta}$ |
| Stability Criterion | $\lambda < \mu$ | $\lambda < \mu$ |

Table 3: A Framework for Performance Bound Computation.

$$O(p) = \left\lceil \frac{predicted\ \ \sigma(p)}{actual\ \ \sigma(p)} \right\rceil \tag{6.5}$$

where

$$\sigma(p) = [\sigma \rightarrow Pr(W \ge \sigma) \le p] \tag{6.6}$$

Tabulated in Table 2 are the $O(10^{-6})$ values for the delay bounds in Example 4.3, corresponding to different EBF characterizations of the server process. In general, the tightness of the bound depends very much on the server characteristics. We also find that, in the multi-hop case of Example 6.1, the analytical end-to-end delay bound becomes looser as the number of hops increases.

| Server Characteristics | $O(10^{-6})$ [delay] |
|---|---|
| (0.30, 1, 1.80) | 2.55 |
| (0.35, 1, 1.28) | 2.66 |
| (0.40, 1, 0.83) | 3.13 |
| (0.45, 1, 0.41) | 4.89 |

Table 2: Over-Estimate Factors.

## 8. CONCLUSIONS

Motivated by the challenge of QOS support over time-varying communication channels, this paper is concerned with the characteristics and traffic effects of variable-rate communication servers. Foremost, we have developed server characterization schemes that are compatible with earlier work on traffic characterization (Table 3). In the deterministic case, the notion of "fluctuation constraint" (FC) is introduced, which imposes a lower bound on server rate variation. This, like the traffic "burstiness constraint" (BC) [4], is a form of envelope process [3]. In the stochastic case, the variable-rate server is characterized as an EBF process, consistent with the EBB method used in [15] to model statistical traffic. Under our framework, if all the input connections to a FC (EBF) work-conserving server-node are BC (EBB), and if the aggregate average input rate is below the average service rate (stability criterion), then:

1. all the corresponding output connections, as well as the aggregate output traffic, are also BC (EBB);
2. the queue size in the server-node is upper bounded (EB);
3. the packet delay across the server-node is also upper bounded (EB).

These results enable the computation of deterministic and statistical bounds on queue length and traffic delay in an isolated work-conserving variable-rate server-node. In addition, the probability of packet loss due to finite buffer size can be estimated from the analytical bound on queue length distribution. Finally, this paper shows how end-to-end performance bounds in a feedforward network of variable-rate server can be determined by applying the results derived for the single-node case. Several numerical examples with applications are given to illustrate the ideas presented.

The work presented in this paper can be extended in two directions. First, the theoretical framework proposed here may be expanded to cover the case of non-feedforward network, in ways similar to those discussed in [5, 12, 15] for fixed-rate servers. Second, it will be useful to investigate how efficient this kind of performance bound analysis can be used for call admission and resource allocation purposes. Yates et al. [17] did a simulation study to address this issue, but did not consider the effect of time-varying link capacity.

## ACKNOWLEDGMENT

The author would like to thank the anonymous reviewers, Hui Zhang, and Allan Fisher for their helpful suggestions; Richard Jordan for his mathematical insights; and, last but not the least, Rene Cruz for reading a draft of this paper and for providing preprints of his recent publications.

## REFERENCES

[1] A. Banerjea and S. Keshav, "Queueing Delays in Rated Controlled ATM networks," *Proc. INFOCOM '93*, pp. 547-556, Mar. 1993.

[2] D. Bertsekas and R. Gallager, *Data Networks*, Englewood Cliffs, NJ: Prentice Hall, 1991.

[3] C. S. Chang, "Stability, Queue Length and Delay, Part I: Deterministic Queueing Networks," *IEEE Trans. Auto. Control*, vol. 39, no. 5, pp. 913-931, May 1994.

[4] R. L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 114-131, Jan. 1991.

[5] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Trans. Inform. Theory*, vol. 37, no. 1, pp. 132-141, Jan. 1991.

[6] R. L. Cruz, "Service Burstiness and Dynamic Burstiness Measures: A Framework," *J. of High Speed Networks*, vol. 1, no. 2, pp. 105-1127, 1992.

[7] R. L. Cruz and H. N. Liu, "End-to-End Queueing Delay in ATM Networks," *J. of High Speed Networks*, vol. 3, no. 4,

pp. 413-428, 1994.
[8] R. L. Cruz, "Quality of Service Guarantees in Virtual Circuit Switched Networks," to appear in *IEEE JSAC*, 1995.
[9] F. P. Kelly, *Reversibility and Stochastic Networks*, New York: Wiley, 1979.
[10] L. Kleinrock, *Queueing Systems, Vol. 1: Theory*, New York: Wiley, 1975.
[11] L. Kleinrock, *Queueing Systems, Vol. 2: Computer Applications*, New York: Wiley, 1976.
[12] J. Kurose, "On Computing Per-Session Performance Bounds in High-Speed Multi-Hop Computer Networks," *Proc. ACM SIGMETRICS/PERFORMANCE '92*, pp. 128-139, 1992.
[13] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 344-357, June 1993.
[14] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case," *IEEE/ACM Trans. Networking*, vol. 2, no. 2, pp. 137-150, Apr. 1994.
[15] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks via Robust Exponential Bounds," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 372-385, June 1993.
[16] O. Yaron and M. Sidi, "Generalized Processor Sharing Networks with Exponentially Bounded Burstiness Arrivals," *Proc. IEEE INFORM '94*, pp. 5b.4.1-5b.4.7, June 1994.
[17] D. Yates, J. Kurose, D. Towsley, and M. G. Hluchyi, "On Per-Session End-to-End Delay Distributions and the Call Admission Problem for Real-Time Applications with QOS Requirements," *Proc. ACM SIGCOMM '93*, pp. 2-12, Sept. 1993.
[18] Z. L. Zhang, D. Towsley, and J. Kurose, "Statistical Analysis of Generalized Processor Sharing Scheduling Discipline," *Proc. ACM SIGCOMM '94*, pp. 68-77, 1994.

## APPENDIX A

Proof of Theorem 3.1: Given $R \sim (\sigma, \lambda)$, $b \geq a$, and $D < \infty$,

$$\begin{aligned}\int_a^b S(t)\,dt &\leq \int_{(a-D)}^b R(t)\,dt \\ &\leq \lambda(b-(a-D)) + \sigma \\ &= \lambda(b-a) + (\sigma + \lambda D) \qquad \text{(A.1)}\end{aligned}$$

**Q.E.D.**

Proof of Theorem 3.2: The fact that each individual output stream $S_i$ is BC $\sim (\sigma_i + \lambda_i D_i, \lambda_i)$ follows directly from Theorem 3.1. As for the aggregate output $S = \Sigma S_i$: given $R_i \sim (\sigma_i, \lambda_i)$, $i = 1, 2, \ldots, N$, $b \geq a$, and $D_i < \infty$,

$$\begin{aligned}\int_a^b S(t)\,dt &= \sum_{i=1}^{N} \int_a^b S_i(t)\,dt \\ &\leq \sum_{i=1}^{N} \int_{(a-D_i)}^b R_i(t)\,dt \\ &\leq \sum_{i=1}^{N} [\lambda_i(b-a) + (\sigma_i + \lambda_i D_i)] \\ &= (b-a)\sum_{i=1}^{N}\lambda_i + \sum_{i=1}^{N}(\sigma_i + \lambda_i D_i) \qquad \text{(A.2)}\end{aligned}$$

**Q.E.D.**

Proof of Theorem 3.3: Let $Q(t)$ denote the amount of data (queue length) in the server-node at time $t$. If $Q(t) > 0$ for any $t_1 \leq t < t_2$, and $Q(t_1^-) = Q(t_2^+) = 0$, then $B = [t_1, t_2)$ is a **busy period**. Given $R \sim (\sigma, \lambda)$ and $C \sim (\delta, \mu)$, one observes that in $B$:

$$\int_{t_1}^{t} R(t)\,dt \leq \lambda(t - t_1) + \sigma \qquad \text{(A.3)}$$

$$\int_{t_1}^{t} S(t)\,dt = \int_{t_1}^{t} C(t)\,dt \geq [\mu(t-t_1) - \delta]^{+} \qquad \text{(A.4)}$$

By definition,

$$\begin{aligned}Q(t) &= \int_{t_1}^{t} R(t)\,dt - \int_{t_1}^{t} S(t)\,dt \\ &\leq [\lambda(t-t_1) + \sigma] - [\mu(t-t_1) - \delta]^{+} \\ &\leq \sigma + \frac{\lambda}{\mu}\delta \qquad \text{(A.5)}\end{aligned}$$

provided that $\lambda < \mu$.

**Q.E.D.**

Proof of Theorem 3.4: Given $R_i \sim (\sigma_i, \rho_i)$, $i = 1, 2, \ldots, N$, and $C \sim (\delta, \mu)$, the queue length $Q(t)$ within a busy period $B = [t_1, t_2)$ is:

$$\begin{aligned}Q(t) &= \int_{t_1}^{t} \sum_{i=1}^{N} R_i(t)\,dt - \int_{t_1}^{t} C(t)\,dt \\ &\leq \left[(t-t_1)\sum_{i=1}^{N}\lambda_i + \sum_{i=1}^{N}\sigma_i\right] - [\mu(t-t_1) - \delta]^{+} \\ &\leq \sum_{i=1}^{N}\sigma_i + \frac{\delta}{\mu}\sum_{i=1}^{N}\lambda_i \qquad \text{(A.6)}\end{aligned}$$

provided that $\Sigma\lambda_i < \mu$.

**Q.E.D.**

Proof of Theorem 3.5: The delay $D$ of an input stream $R \sim (\sigma, \lambda)$ across a general service server node $C \sim (\delta, \mu)$ is bounded by the duration $L = (t_2 - t_1)$ of a busy period $B = [t_1, t_2)$. Note that:

$$\int_{t_1}^{(t_1+L)} R(t)\,dt - \int_{t_1}^{(t_1+L)} C(t)\,dt = 0 \qquad \text{(A.7)}$$

So,

$$D \leq max\left\{L \rightarrow \int_{t_1}^{(t_1+L)} R(t)\,dt = \int_{t_1}^{(t_1+L)} C(t)\,dt\right\} \qquad \text{(A.8)}$$

Since both integrals in (A.8) are non-decreasing,

$$\begin{aligned}D &\leq max\{L \rightarrow [\lambda L + \sigma] = [\mu L - \delta]^{+}\} \\ &= \frac{\sigma + \delta}{\mu - \lambda} \qquad \text{(A.9)}\end{aligned}$$

provided that $\lambda < \mu$.

**Q.E.D.**

Proof of Theorem 3.6: Given $R \sim (\sigma, \lambda)$ and $C \sim (\delta, \mu)$, within a busy period $B = [t_1, t_2)$, the delay D across a FCFS server-node is upper-bounded by:

$$D \le max\{x \to \int_{t_1}^{t} R(t)\,dt = \int_{t_1}^{(t+x)} C(t)\,dt\} \quad \text{(A.10)}$$

for any $t_1 \le t < t_2$. Since both integrals are non-decreasing,

$$D_{max} \le max\{x \to [\lambda(t-t_1)+\sigma] = [\mu(t+x-t_1)-\delta]^+ \quad \}$$

$$= \frac{\sigma+\delta}{\mu} \quad \text{(A.11)}$$

**Q.E.D.**

## APPENDIX B

Theorems 4.1-4.6 are derived for the discrete-time case using similar techniques as in [15] based on busy-period analysis and union-bound approximation. Without loss of generality, the server queue is assumed to be empty at time = 0. In addition, *h(a)* is defined as the elapsed time since the queue was empty prior to *a*; so the current busy period begins at (*a* - *h(a)*) if the queue is not empty at *a*; and *h(a) = 0* if the queue is empty at *a*. For convenience, we let $k = b - a$, and use $F_{(a,b)}$ to denote $\sum_{i=a+1}^{b} F(i)$, for any function *F(i)*.

Proof of Theorem 4.1: Given $R \sim (\lambda, A, \alpha)$ and $C \sim (\mu, B, \beta)$,

$$Pr\{S_{(a,b)} \ge \lambda(b-a)+\sigma\}$$

$$= \sum_{i=0}^{a} Pr\{[S_{(a,b)} \ge \lambda k+\sigma] \cap [h(a)=i]\}$$

$$\le \sum_{i=0}^{a} Pr\{[R_{(a-i,b)} - C_{(a-i,a)}] \ge \lambda k+\sigma\} \quad \text{(B.1)}$$

Applying the union-bound approximation yields:

$$Pr\{[R_{(a-i,b)} - C_{(a-i,a)}] \ge \lambda k+\sigma\} \quad \text{(B.2)}$$

$$\le Pr\{R_{(a-i,b)} \ge \lambda(i+k) + p(\sigma+(\mu-\lambda)i)\}$$

$$+ Pr\{C_{(a-i,a)} \le \mu i - (1-p)(\sigma+(\mu-\lambda)i)\}$$

for some $0 \le p \le 1$. Since $R \sim (\lambda, A, \alpha)$ and $C \sim (\mu, B, \beta)$,

$$Pr\{[R_{(a-i,b)} - C_{(a-i,a)}] \ge \lambda k+\sigma\} \quad \text{(B.3)}$$

$$\le Ae^{-\alpha p(\sigma+(\mu-\lambda)i)} + Be^{-\beta(1-p)(\sigma+(\mu-\lambda)i)}$$

Equating $\alpha p = \beta(1-p) = \zeta$ gives:

$$\frac{1}{\zeta} = \frac{1}{\alpha} + \frac{1}{\beta} \quad \text{(B.4)}$$

Then, from (B.1), it follows that:

$$Pr\{S_{(a,b)} \ge \lambda(b-a)+\sigma\}$$

$$\le \sum_{i=0}^{a} \left[(A+B)e^{-\zeta(\sigma+(\mu-\lambda)i)}\right]$$

$$\le \sum_{i=0}^{\infty} \left[(A+B)e^{-\zeta(\sigma+(\mu-\lambda)i)}\right]$$

$$= \frac{A+B}{1-e^{-\zeta(\mu-\lambda)}} e^{-\zeta\sigma} \quad \text{(B.5)}$$

**Q.E.D**

Proof of Theorem 4.2: Given $R_i \sim (\lambda_i, A_i, \alpha_i)$, $i = 1, 2, ..., N$, and $C \sim (\mu, B, \beta)$,

$$Pr\{S_{1(a,b)} \ge \lambda_1(b-a)+\sigma\}$$

$$= \sum_{i=0}^{a} Pr\{[S_{1(a,b)} \ge \lambda_1 k+\sigma] \cap [h(a)=i]\}$$

$$\le \sum_{i=0}^{a} Pr\left\{R_{1(a-i,b)} + \sum_{j=2}^{N} R_{j(a-i,a)} - C_{(a-i,a)} \ge \lambda_1 k+\sigma\right\} \quad \text{(B.6)}$$

Using the union-bound approximation, one gets:

$$Pr\left\{\left[R_{1(a-i,b)} + \sum_{j=2}^{N} R_{j(a-i,a)} - C_{(a-i,a)}\right] \ge \lambda_1 k+\sigma\right\}$$

$$\le Pr\left\{R_{1(a-i,b)} \ge \lambda_1(i+k) + p_1\left(\sigma+\left(\mu-\sum_{m=1}^{N}\lambda_m\right)i\right)\right\}$$

$$+ \sum_{j=2}^{N} Pr\left\{R_{j(a-i,a)} \ge \lambda_j i + p_j\left(\sigma+\left(\mu-\sum_{m=1}^{N}\lambda_m\right)i\right)\right\}$$

$$+ Pr\left\{C_{(a-i,a)} \le \mu i - p_c\left(\sigma+\left(\mu-\sum_{m=1}^{N}\lambda_m\right)i\right)\right\}$$

$$\le \sum_{j=1}^{N} A_j e^{-\alpha_j p_j\left(\sigma+\left(\mu-\sum_{m=1}^{N}\lambda_m\right)i\right)} + Be^{-\beta p_c\left(\sigma+\left(\mu-\sum_{m=1}^{N}\lambda_m\right)i\right)} \quad \text{(B.7)}$$

where $\Sigma p_i + p_c = 1$. Equating $\alpha_1 p_1 = \alpha_2 p_2 = ... = \alpha_N p_N = \beta p_c = \zeta$ yields:

$$\frac{1}{\zeta} = \sum_{j=1}^{N} \frac{1}{\alpha_j} + \frac{1}{\beta} \quad \text{(B.8)}$$

Then, from (B.6), one gets:

$$Pr\{S_{1(a,b)} \ge \lambda_1(b-a)+\sigma\}$$

$$\le \sum_{i=0}^{a} \left[\left(\sum_{j=1}^{N} A+B\right) e^{-\zeta\left(\sigma+\left(\mu-\sum_{j=1}^{N}\lambda_j\right)i\right)}\right]$$

$$\le \frac{\sum_{j=1}^{N} A+B}{1-e^{-\zeta\left(\mu-\sum_{j=1}^{N}\lambda_j\right)}} e^{-\zeta\sigma} \quad \text{(B.9)}$$

thus indicating that each individual output stream $S_i$ is EBB. Next, let $R = \Sigma R_i$ denote the aggregate output of the server. Since the sum of multiple EBB processes is also EBB [15],

$$R \sim \left( \sum_{j=1}^{N} \lambda_j, \sum_{j=1}^{N} A_j, \left( \sum_{j=1}^{N} \frac{1}{\alpha_j} \right)^{-1} \right) \tag{B.10}$$

Then Theorem 4.1 asserts that the aggregate output $S = \Sigma S_i$ satisfies,

$$S \sim \left( \sum_{j=1}^{N} \lambda_j, \frac{\sum_{j=1}^{N} A_j + B}{1 - e^{-\zeta\left(\mu - \sum_{j=1}^{N} \lambda_j\right)}}, \zeta \right) \tag{B.11}$$

where $\zeta$ is as given in (B.8).

**Q.E.D**

*Proof of Theorem 4.3*: Given $R \sim (\lambda, A, \alpha)$ and $C \sim (\mu, B, \beta)$, and letting $Q(t)$ denote the queue length of this server-node,

$$Pr\{Q(t) \geq \sigma\} = \sum_{i=0}^{t} Pr\{[Q(t) \geq \sigma] \cap [h(t) = i]\}$$

$$= \sum_{i=0}^{t} Pr\{[R_{(t-i,t)} - C_{(t-i,t)}] \geq \sigma\}$$

$$\leq \sum_{i=0}^{t} Pr\{R_{(t-i,t)} \geq \lambda i + p(\sigma + (\mu - \lambda) i)\}$$

$$+ Pr\{C_{(t-i,t)} \leq \mu i - (1-p)(\sigma + (\mu - \lambda) i)\}$$

$$\leq \sum_{i=0}^{\infty} \left[A e^{-\alpha p(\sigma + (\mu - \lambda) i)} + B e^{-\beta(1-p)(\sigma + (\mu - \lambda) i)}\right] \tag{B.12}$$

for some $0 \leq p \leq 1$. Equating $\alpha p = \beta(1-p) = \zeta$ gives:

$$\frac{1}{\zeta} = \frac{1}{\alpha} + \frac{1}{\beta} \tag{B.13}$$

Thus,

$$Pr\{Q(t) \geq \sigma\} \leq \sum_{i=0}^{\infty} (A+B) e^{-\zeta(\sigma + (\mu - \lambda) i)}$$

$$= \frac{A+B}{1 - e^{-\zeta(\mu - \lambda)}} e^{-\zeta\sigma} \tag{B.14}$$

**Q.E.D.**

*Proof of Theorem 4.4*: Given $R \sim (\lambda, A, \alpha)$ and $C \sim (\mu, B, \beta)$, then the delay $D(t)$ in a general service server is bounded by the residual time of the busy period at time $t$. In particular,

$$Pr\{D(t) \geq \sigma\} = \sum_{i=0}^{t} Pr\{[D(t) \geq \sigma] \cap [h(t) = i]\}$$

$$= \sum_{i=0}^{t} Pr\{[R_{(t-i,t+\sigma)} - C_{(t-i,t+\sigma)}] \geq 0\}$$

$$\leq \sum_{i=0}^{t} Pr\{R_{(t-i,t)} \geq \lambda i + p(\mu - \lambda)(\sigma + i)\}$$

$$+ \sum_{i=0}^{t} Pr\{C_{(t-i,t)} \leq \mu i - (1-p)(\mu - \lambda)(\sigma + i)\}$$

$$\leq \sum_{i=0}^{\infty} \left[A e^{-\alpha p(\mu - \lambda)(\sigma + i)} + B e^{-\beta(1-p)(\mu - \lambda)(\sigma + i)}\right] \tag{B.15}$$

for some $0 \leq p \leq 1$. Equating $\alpha p = \beta(1-p) = \zeta$ gives:

$$\frac{1}{\zeta} = \frac{1}{\alpha} + \frac{1}{\beta} \tag{B.16}$$

So,

$$Pr\{D(t) \geq \sigma\} \leq \sum_{i=0}^{\infty} (A+B) e^{-\zeta(\mu - \lambda)(\sigma + i)}$$

$$= \frac{A+B}{1 - e^{-\zeta(\mu - \lambda)}} e^{-\zeta(\mu - \lambda)\sigma} \tag{B.17}$$

**Q.E.D.**

*Proof of Theorem 4.5*: Given $R \sim (\lambda, A, \alpha)$ and $C \sim (\mu, B, \beta)$, the delay $D(t)$ of a data bit entering the FCFS server at time $t$ is given by:

$$D(t) = \{d \rightarrow [R_{(t-i,t)} = C_{(t-i,t+d)}]\} \tag{B.18}$$

where $(t - i)$, $0 \leq i \leq t$, marks the beginning of the busy period. Therefore,

$$Pr\{D(t) \geq \sigma\}$$

$$= \sum_{i=0}^{t} Pr\{[D(t) \geq \sigma] \cap [h(t) = i]\}$$

$$= \sum_{i=0}^{t} Pr\{[R_{(t-i,t)} - C_{(t-i,t+\sigma)}] \geq 0\}$$

$$\leq \sum_{i=0}^{t} \left[Pr\{R_{(t-i,t)} \geq x\} + Pr\{C_{(t-i,t+\sigma)} \leq x\}\right]$$

$$\leq \sum_{i=0}^{\infty} \left[A e^{-\alpha(x - \lambda i)} + B e^{-\beta(\mu(i+\sigma) - x)}\right] \tag{B.19}$$

for some $x$. Setting $\alpha(x - \lambda i) = \beta(\mu(i + \sigma) - x)$ gives:

$$x = \frac{\beta\mu\sigma + (\alpha\lambda + \beta\mu) i}{\alpha + \beta} \tag{B.20}$$

It then follows that:

$$Pr\{D(t) \geq \sigma\} \leq \sum_{i=0}^{\infty} (A+B) e^{-\zeta((\mu - \lambda) i + \mu\sigma)}$$

$$= \frac{A+B}{1 - e^{-\zeta(\mu - \lambda)}} e^{-\zeta\mu\sigma} \tag{B.21}$$

where $\zeta = (\alpha\beta)/(\alpha+\beta)$.

**Q.E.D.**