3RD QUARTER 2006, VOLUME 8, NO. 3

IEEE COMMUNICATIONS SURVEYS
The Electronic Magazine of Original Peer-Reviewed Survey Articles

www.comsoc.org/pubs/surveys

A SURVEY OF ENVELOPE PROCESSES AND THEIR APPLICATIONS IN QUALITY OF SERVICE PROVISIONING

SHIWEN MAO, AUBURN UNIVERSITY
SHIVENDRA S. PANWAR, POLYTECHNIC UNIVERSITY

ABSTRACT

Provisioning of quality of service (QoS) guarantees has become an increasingly important and challenging topic in the design of the current and the next-generation Internet. The class of envelope processes (EPs) is one of the key elements for many QoS provisioning mechanisms. An arrival EP $\hat{A}(\tau)$ (or a service curve) bounds the cumulative traffic of a flow (or the cumulative service a flow receives) over any interval of length τ. Such bounds can be deterministic or probabilistic, and can be used for provisioning of deterministic or statistical service guarantees. In this article we provide a survey on arrival EPs and service curves. We provide an overview of various EPs proposed in the literature during the last 15 years and discuss their applications and performance in QoS provisioning. We aim to provide a big picture of the existing work. There is considerable research effort addressing QoS issues in resource-constrained access networks (such as wireless networks) and in the new multiprotocol label switching (MPLS) and peer-to-peer (P2P) networking paradigms. We aim to provide a comprehensive survey of existing work, which can yield useful insights, and help the development of new QoS metrics, mechanisms, and architectures for emerging network environments.

The Internet continues to provide a fast growing arena for new applications and services. Multimedia traffic is becoming an increasing portion of today's Internet traffic due to the proliferation of applications such as music/video streaming, video teleconferencing, IP telephony, and distance learning [1–3]. Such applications can have diverse quality of service (QoS) requirements, while the traffic generated is real-time and could be highly bursty. One major concern with regard to the design, implementation, operation, and management of the Internet is how to provide QoS guarantees for such applications, while achieving a high utilization of network resources.

QoS guarantees can be provisioned in the Internet using the Intserv architecture described in [4] and the Diffserv architecture described in [5]. However, due to advances in Dense Wavelength Division Multiplexing (DWDM) technology, overprovisioning in the network core has become popular among many service providers. Nevertheless, over-provisioning does not necessarily solve the QoS problem, for it may not be applicable to *all* segments of the network due to technical, regulatory, or capital investment limitations. In addition, over-provisioning in the core does not automatically provide service assurance due to the best-effort handling of some application traffic [6]. All these have made it difficult to guarantee application performance on the end-to-end basis. As a result, QoS mechanisms are still needed for the relatively resource constrained access networks (e.g., wireless access networks), even while applying over-provisioning in the core.

There has been tremendous work in the area of QoS research over the years, which is still an active area attracting considerable research efforts, as indicated by new conferences and journal special issues dedicated to this research [7]. Researchers have developed various QoS mechanisms, such as traffic shaping, admission control, signaling and resource reservation, scheduling, QoS routing, congestion control, and queue management (see [6] for a survey on these QoS "build-

ing blocks"). The class of envelope processes (EP), not only underpinned by rigorous theoretical analysis, but also widely implemented in practice, is one of the key elements for many of these QoS mechanisms.

EPs belong to the class of bounding traffic models. An arrival EP $\hat{A}(\tau)$ upper bounds the cumulative traffic $A(\tau)$ of a flow over any interval of length τ. Such traffic bounds could be deterministic (i.e., strict bounds) or probabilistic (i.e., violation is allowed, but with a small probability), and can be used for provisioning of deterministic or statistical service guarantees. Such a bounding traffic model is especially appealing since it is often not feasible to obtain an accurate statistical characterization of traffic sources, and an exact performance analysis with statistical traffic characterizations may be intractable. In addition to arrival EPs, the class-of-service envelope processes, termed *service curves* in the literature, provide deterministic or probabilistic bounds on the cumulative service a traffic flow receives at a network element. The use of these EPs can abstract not only traffic flow, but also network elements with complex scheduling disciplines. The resulting network calculus can greatly simplify performance analysis at a single network element, as well as for the entire end-to-end path.

As a popular traffic model, EPs find wide applications in network operations and control, such as traffic specification, negotiation of a "traffic contract" between the user and service provider, admission control, and traffic policing, shaping, and pricing. In practice, deterministic EPs are implemented in most commercial routers [8, 9] and Linux operating systems [10, 11]. It is expected that both deterministic and probabilistic EPs will find their wider adoption in network operations as more multimedia applications and other mission-critical applications (e.g., distributed computing, e-commerce, and online stock exchanges) are supported.

In this article we provide a survey of arrival EPs and service curves. We overview various EPs proposed in the literature in the last 15 years and discuss their applications and performance in QoS provisioning. We believe such a survey is relevant and of importance due to the strong interest and considerable ongoing efforts in multimedia networking and distributed computing. Such a survey would be useful and timely for researchers and practitioners entering or working in this exciting area, to provide a big picture of the existing work and to facilitate their efforts along this line of research. We believe that significant future research is needed to address QoS issues in resource-constrained access networks, including 4G wireless networks, wireless mesh networks, mobile ad hoc networks, and wireless sensor networks, and in the new multiprotocol label switching (MPLS) and peer-to-peer (P2P) networking paradigms. Existing work surveyed in this article can provide useful insights and help the development of new QoS metrics, mechanisms and architectures for these new network environments.

There have been only a few surveys of QoS-related issues. In [6], Soldatos, Vayias, and Kormentzas provide a comprehensive survey of the "building blocks" for QoS provisioning, such as admission control and QoS routing, among others. In [12], Knightly and Shroff present a survey and comparison of representative admission control schemes. The EPs surveyed in the present article are actually key elements (or "building blocks") of the mechanisms surveyed in these related articles. In [13], Le Boudec and Thiran provides a comprehensive treatment of deterministic network calculus; however, the extensive work on probabilistic EPs is not included in this book. Finally, both deterministic guarantees and probabilistic guarantees are examined in Chang's textbook [14], but some of the latest (and important) advances in this area are not covered in this work, such as the class of EPs for self-similar traffic [15, 16] and statistical network calculus [17].

The remainder of this survey is organized as follows. We examine the class of deterministic EPs, which strictly bound the cumulative traffic of a source flow and can be used to provide deterministic services (such as a bounded delay). Starting with Cruz's $\{\sigma, \rho\}$ EP [18, 19], we first present a general definition for deterministic EPs and their key properties, and then introduce two classes of piecewise linear EPs, namely, the $\{\vec{\sigma}, \vec{\rho}\}$ EP and Deterministic Bounding Interval Dependent (D-BIND) [20]. Both these piecewise linear EPs are used to obtain a tighter bound for traffic sources exhibiting burstiness over multiple time-scales. Next, we examine the application of deterministic EPs in single-node performance analysis and admission control tests [13, 18, 21, 22] and also present a simulation study of their performance in bandwidth utilization [22]. Finally, we review the set of work on statistical multiplexing of deterministically regulated flows, and show that significant improvement in bandwidth utilization can be achieved by statistical multiplexing that exploits independence among the traffic flows [23–31].

The class of probabilistic EPs are presented. This class of EPs bounds the cumulative traffic of a source flow in a probabilistic manner: the source is allowed to exceed its probabilistic EP, but with a small probability. Such EPs are useful in providing probabilistic service assurances, e.g., a delay bound that is satisfied with a certain probability. By allowing a fraction of traffic to violate its QoS requirement, a user can easily trade-off the QoS received with the network resource required. More importantly, such a probabilistic approach can significantly improve network resource utilization. We review representative probabilistic EPs and their applications in probabilistic service assurance, including the class of bounded burstiness processes [32–34], Chang's log-moment generating function bound [14, 21], Kurose's bound using a family of random variables [35], H-BIND [20, 36], rate variance EPs [37, 36], and effective envelopes [23].

We present recent advances in EPs for self-similar traffic flows. It has been shown by many empirical studies that network data and video traffic are long-range dependent (LRD) or self-similar processes that exhibit high burstiness over multiple timescales [38–41]. The Weibull Bounded Burstiness EP [34, 42] and the fBm EP [15] are motivated by the fractal Brownian motion (fBm) traffic model for connectionless traffic [43]. We also introduce a self-similar leaky bucket for effective regulation of such traffic flows [15]. Finally, we introduce the multifractal Brownian motion (mBm) EP that provides a good probabilistic bound for multifractal traffic flows [16].

We review the work on service curves, which provides deterministic or probabilistic bounds on the cumulative service a flow receives at a network element [13, 17, 44–56]. Such service curves are very useful in abstracting complex service disciplines and, when combined with arrival EPs, can greatly simplify the derivation of performance bounds at various network elements. More importantly, service curves are very useful in deriving end-to-end performance measures, where the entire path could be represented by a *network service curve*. We present the key deterministic network calculus results in this section [13, 18, 19, 21, 57], as well as an important probabilistic extension: statistical network calculus [17, 46], which can provide end-to-end statistical assurance using the min-plus algebra [58].

We conclude this article with a summary and qualitative comparison of the EPs surveyed in this article, and a discussion on future research directions.

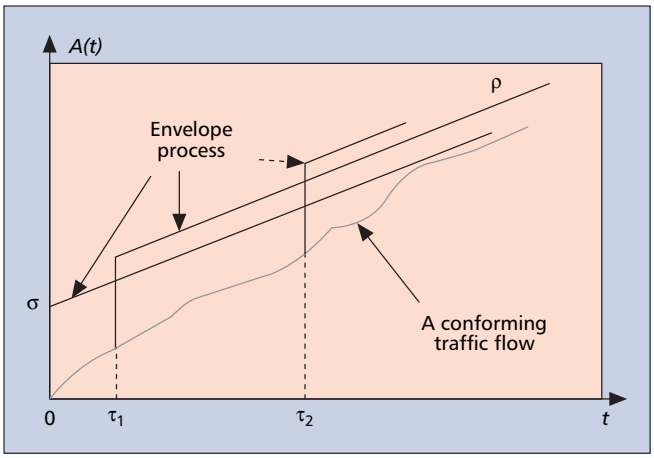

Figure 1. *Cruz's* {σ, ρ} *Envelope Process* $\hat{A}(t) = \sigma + \rho \cdot t$.

Deterministic Envelope Processes

In this section we review the class of deterministic EPs. We first present the definition and properties of such EPs, and then examine two representative deterministic EPs. Their applications in QoS provisioning is discussed and a performance study of such EPs is reviewed. Finally, we review the class of work on the statistical multiplexing of deterministically regulated flows.

Definition

As discussed, envelope processses (EPs) are popular traffic models used to bound the traffic generated by user sessions [18, 19]. Consider a source generating traffic at rate $a(t)$. The amount of traffic generated in the time interval $[t_1, t_2)$ is $A(t_1, t_2) = \int_{t_1}^{t_2} a(t)dt$. For discrete-time systems, $a(t)$ is the amount of arrival traffic in the t-th time slot, and the cumulative traffic during $[t_1, t_2)$ is $A(t_1, t_2) = \Sigma_{t_1}^{t_2-1} a(t)$. Throughout this article we consider *stationary* sources, that is, the statistical characteristics do not change over time.

For such traffic flows, Cruz's {σ, ρ} EP is defined as:

$$A(t_2–t_1) \le \rho \cdot (t_2 - t_1) + \sigma, \forall t_1 \le t_2, \quad (1)$$

where σ is the burstiness allowed, and ρ is an upper bound on the long term average rate of the traffic flow [18]. Although the cumulative traffic $A(t_1, t_2)$ could have various forms, it is upper bounded by the deterministic function $\hat{A}(t_2 - t_1) = \rho \cdot (t_2 - t_1) + s$ during $[t_1, t_2)$ for any $t_1 \le t_2$. Figure 1 illustrates such an EP and the corresponding traffic flow. The EP can be shifted along the time axis from $t = 0$ to $t = \tau_1$ (or to $t = \tau_2$), while the cumulative arrival should always be upper bounded by the shifted functions. That is, the EP is only a function of the time interval $\tau = t_2 - t_1$, regardless where the interval begins (i.e., t_1).

Cruz's EP is a simple linear function defined by two parameters σ and ρ.[1] In fact, a deterministic EP does not have to follow such a specific form, and could be any nondecreasing, nonnegative function of time $\hat{A}(\tau)$, as long as the cumulative traffic is bounded as follows [21]:

$$A(t_1, t_2) \le \hat{A}(t_2 - t_1), \forall t_1 \le t_2. \quad (2)$$

The choice for a specific $\hat{A}(\tau)$ depends on how easy it is to enforc and analyze, and how tight it is in bounding the traffic flow (and thus how efficient it is in resource utilization). We will discuss examples of generalized deterministic EPs in the following sections.

It is worth noting that for a given traffic flow, its EP is not unique. For example, if we have $A(t) \le \hat{A}(\tau)$, then we have $A(t) \le k \cdot \hat{A}(\tau)$ for any $k > 1$. For QoS provisioning, it is therefore important to examine the tightest one among all the bounding functions. It is shown in [21] that if $\hat{A}(\tau)$ is increasing and subadditive,[2] its long-term average rate $\hat{\rho}$ exists:

$$\hat{\rho} \overset{def}{=\!=} \lim_{t\to\infty} \frac{\hat{A}(t)}{t} = \inf_{t\ge 1} \frac{\hat{A}(t)}{t} \quad (3)$$

The limit $\hat{\rho}$ is referred to as the *envelope rate* of $\hat{A}(\tau)$. The minimum EP (i.e., the tightest bounding function) is called the minimum envelope process (MEP) of $A(t)$ and is found to be

$$A^*(t) = \sup_{s\ge 0} A(s, s+t). \quad (4)$$

That is, $A^*(t)$ is the maximum amount of traffic that could possibly be generated by the given source in a time interval of length t. It is shown in [21] that $A^*(t)$ is increasing and subadditive. The tightest bound on the average rate, ρ*, called the minimum envelope rate (MER), is the envelope rate of $A^*(t)$ [see Eq. 3].

The $\{\vec{\sigma}, \vec{\rho}\}$ Envelope Process

The {σ, ρ} EP has the advantages of being simple and easy to enforce with a token bucket. However, it is relatively loose because it enforces a burstiness constraint σ over all the timescales, while typically a source tends to exhibits smaller burstiness over larger time-scales. For example, consider a variable bit rate (VBR) video source. The burstiness is largest at the timescale of a frame time. Its average bit rate over an interval gradually decreases as the interval gets larger, and is equal to the long-term average rate at the timescale of the entire video trace length. For a given traffic process, a tighter EP indicates a more accurate estimate of the actual traffic load and implies that less network resources are required to accommodate it. To achieve a high utilization, it is thus desirable to use an EP that bounds the traffic over different timescales using different burstiness factors.

The discussion above suggests that a tighter traffic bound can be obtained by deploying a concave, piecewise-linear envelope. In [22, 31], the {σ, ρ} EP was extended to the $\{\vec{\sigma}, \vec{\rho}\}$ EP, which maintains n {σ, ρ} pairs. The amount of traffic in a time interval t is restricted by $\hat{A}(t) = \min_{1\le i\le n}\{\sigma_i + \rho_i t\}$. Since each term $\sigma_i + \rho_i t$ is an *affine* function and the minimum of n affine functions is concave, $\hat{A}(\tau)$ is concave and subadditive.

Figure 2 illustrates such a $\{\vec{\sigma}, \vec{\rho}\}$ EP, which consists of four linear segments and tightly bounds the arrival process $A(t)$. The $\{\vec{\sigma}, \vec{\rho}\}$ EPs can be enforced using a number of cascaded leaky buckets, while a conforming flow keeps all the leaky buckets from overflowing. The T-SPEC in Intserv [5] adopts a dual leaky bucket model for traffic specification. The corresponding EP is $\hat{A}(\tau) = \min\{M + Pt, B + rt\}$, where M, P, B, and r are parameters with constant values. This is a special case of the $\{\vec{\sigma}, \vec{\rho}\}$ EP with $n = 2$.

D-BIND

Another piece-wise linear EP, D-BIND was introduced by Knightly *et al.* to address this multiple time-scale burstiness problem in traffic flows [20]. To further illustrate this prob-

[1] *An arbitrary traffic flow can be policed to be conforming to the* {σ, ρ} *EP by using a* leaky bucket *with a token rate* ρ *and a token bucket size* σ.

[2] *A process* A(t) *is subadditive if* $A(t_1 + t_2) \le A(t_1) + A(t_2)$.

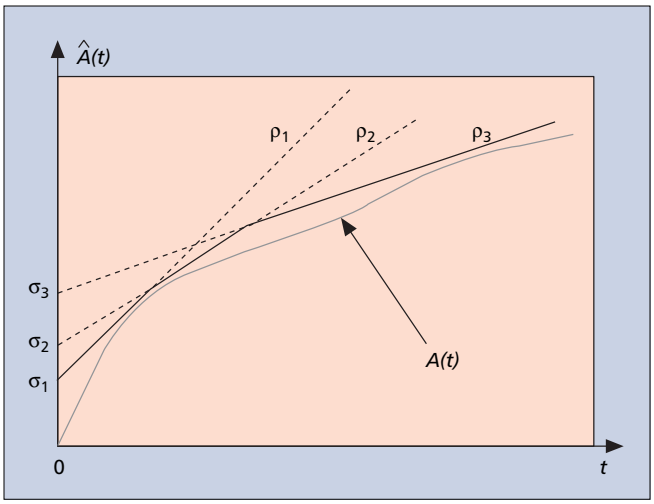

Figure 2. *A piece-wise linear* $\{\vec{\sigma}, \vec{\rho}\}$ *envelope process with three* $\{\sigma, \rho\}$ *pairs.*

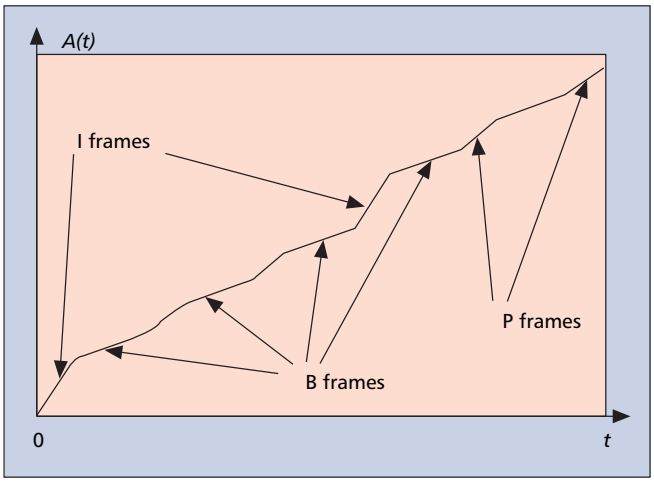

Figure 3. *Burstiness of an MPEG video trace with a frame pattern of IBBPBBPBB.*

lem, consider the example in Fig. 3, which plots the arriving process of an MPEG video trace with frame pattern IBBPBBPBB. Usually, I frames have the highest rates and B frames the lowest rates. It can be seen from this figure that using different bounds on the rate for different time intervals can provide a tighter approximation for the traffic flow than using a single worst case bound for all the intervals.

Let R_k denote a data rate and I_k denote the corresponding time interval. D-BIND was defined using multiple rate-interval pairs, $\{(R_k, I_k), k = 1, 2, \ldots, K\}$, as the following piecewise linear function:

$$\hat{A}(t) = \frac{R_k I_k - R_{k-1} I_{k-1}}{I_k - I_{k-1}}(t - I_k) + R_k I_k, I_{k-1} \le t \le I_k, \quad (5)$$

where $\hat{A}(0) = 0$ [20]. With Eq. 5, the rate R_k can be viewed as an upper bound on the session's rate over any time interval of length I_k, that is, $A(t, t + I_k)/I_k \le R_k$, $\forall t > 0$, $k = 1, 2, \ldots, K$.

It is worth noting that the (σ, ρ) model may be viewed as a special case of the D-BIND model, since both are piecewise linear functions. For both EPs, an important design parameter is K, the number of rate-interval (or $\{\vec{\sigma}, \vec{\rho}\}$) pairs to use. There is a trade-off between the tightness of the EP (which determines the bandwidth utilization at network nodes) and K (which determines the complexity in shaping and policing).

Generally, the more rate-interval pairs used, the tighter the bound and therefore the higher the utilization (see the analysis in the next section). However, considering a typical core router where tens of thousands sessions are multiplexed, it is impractical to use a large number of rate-interval pairs for each session. Knightly *et al.* suggest that a source specify a small number of rate-interval pairs (e.g., four or eight) for connection admission control and policing [20]. We examine this issue subsequently.

Applications of Deterministic EPs

Generally, the use of EPs is two-fold:

- To simplify the enforcement of user traffic flows (e.g., by adopting one or more leaky buckets) at the network boundary
- To simplify the QoS provisioning in networks

In this section we discuss how to derive performance bounds and how to perform admission control using deterministic EPs. Their performance is discussed in the next section.

Performance Bounds — Consider a network element with service rate c, modeled as a slotted-time single server $G/G/1$ queue. At time slot t, the amount of arrival traffic is $a(t)$. Under a work conserving policy,[3] the backlog process $q(t)$ of the queue is governed by the Lindley's equation [59]:

$$q(t + 1) = \max\{0, q(t) + a(t) - c\}. \quad (6)$$

That is, the queue length in the next time slot is the current queue length increased by the traffic input $a(t)$ and decreased by the traffic being served during this slot. The maximum operation is used, since the queue length should always be nonnegative. It has been shown in [59] that the distribution of $q(t)$ determined by Eq. 6 converges to a unique limit distribution as $t \to \infty$, under some mild conditions [e.g., $a(t)$ is stationary and ergodic, and the average of $a(t)$ should be less than the service capacity c].

Assuming the queue is empty at time 0, expanding Eq. 6 recursively yields

$$\begin{aligned} q(t) &= \max\{0, \max\{0, q(t-2) + a(t-2) - c\} + a(t-1) - c\} \\ &= \max\{0, a(t-1) - c, q(t-2) + a(t-1) + a(t-2) - 2c\} \\ &= \ldots \\ &= \max\{0, a(t-1) - c, \ldots, a(t-1) + a(t-2) + \ldots + a(0) - tc\}, \end{aligned}$$

that is,

$$\begin{aligned} q(t) &= \max_{0 \le k < t} \left\{0, \sum_{l=k}^{t-1} a(l) - (t-k)c\right\} \\ &= \max_{0 \le k < t} \{0, A(t-k) - (t-k)c\} \quad (7) \\ &\le \max_{0 \le k < t} \{0, \hat{A}(t-k) - (t-k)c\}. \quad (8) \end{aligned}$$

Eq. 7 conveniently relates the cumulative traffic $A(t)$ to the queue backlog.[4] An upper bound on the backlog can be derived by substituting $A(t)$ with its upper bounding approximation $\hat{A}(\tau)$ as in Eq. 8. Note that in Eq. 7, the right-hand-

[3] *That is, the server will not be idle whenever traffic is backlogged.*

[4] *For continuous time systems, we have* $q(t) = \max_{\tau \le t}\{\int_\tau^t a(t)dt - (t-\tau)c\}$ *and* $q(t) = \max_{\tau \le t}\{A(t-\tau) - (t-\tau)c\} = \max_{s \le t}\{A(s) - sc\}$. *This derivation is similar to the discrete time system (where summation is replaced with integration for continuous-time systems). In addition, this powerful equation has been used to derive probabilistic bounds on QoS metrics.*

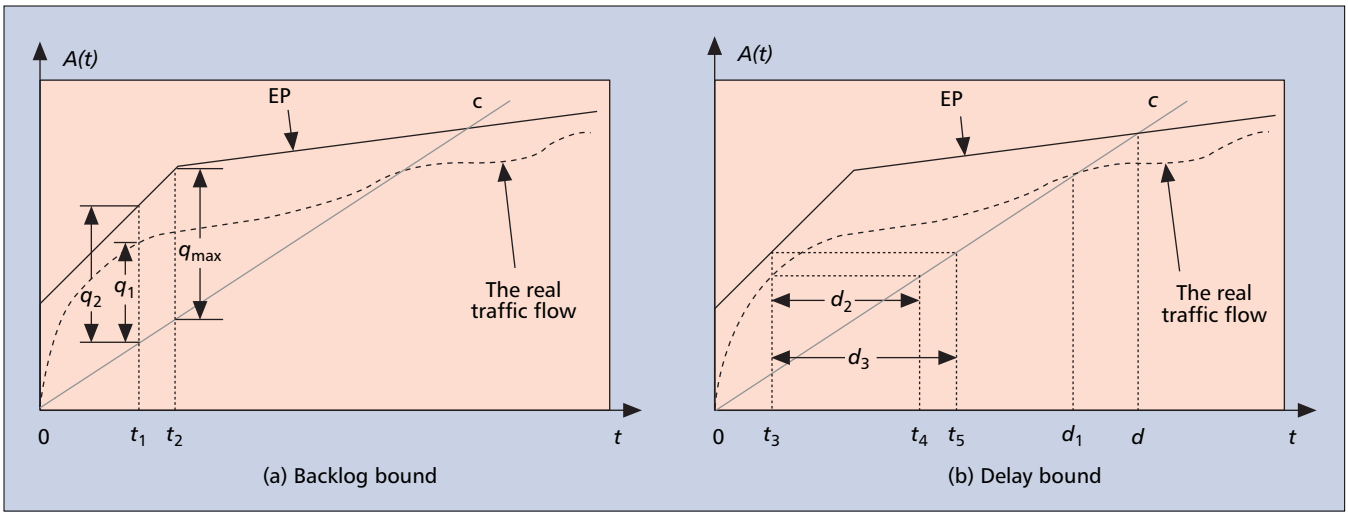

Figure 4. *A graphical interpretation of the backlog and delay bounds.*

side (RHS) consists of the random arrival process $A(t)$, while in Eq. 8, the RHS consists of a deterministic function $\hat{A}(\tau)$. The analysis is greatly simplified by this substitution.

Based on Eq. 8, we can derive performance bounds on backlog and delay for the network element as follows. Substituting $s = t - k$ into Eq. 8, we have [13]:

$$q(t+1) \le \max_{0<s\le t} \{0, \hat{A}(s) - sc\}. \tag{9}$$

This is intuitive, since backlog is actually the amount of cumulative arrival decreased by the amount that has been served.

The delay bound depends on the specific service discipline. For a work conserving queue with any service discipline (e.g., LCFS or FCFS), the delay of a traffic unit is upper bounded by the *system busy period d*. That is, since any backlog will be served within d time units, the delay experienced by any traffic unit will be no larger than d. Clearly, d is the time interval when the cumulative traffic arrival is equal to the cumulative service. An upper bound on d can be easily computed as [18, 19, 35]:

$$d \le \min \{t : t \ge 1 \text{ and } \hat{A}(t) - ct \le 0\}. \tag{10}$$

If the service discipline is FCFS, then a traffic unit will be served after all the traffic arriving earlier than itself is cleared. This fact can be exploited to tighten the delay bound. Consider a traffic unit arriving at time t. Since it sees a backlog of $q(t)$, the delay it experiences will be the time it takes to serve the backlog $q(t)$, that is, s such that $q(t) - cs \le 0$. We then have

$$d(t) \le \min \{s : s \ge 1 \text{ and } \hat{A}(t) - c(t+s) \le 0\}. \tag{11}$$

These results can be further explained via an intuitive graphical interpretation. In Fig. 4a, the queue becomes busy after time 0. At time t_1, the difference between the cumulative arrival and cumulative service is the amount of backlog in the queue (i.e., q_1). Since we use the deterministic EP to approximate the traffic flow, an upper bound on the backlog is found to be the difference between the EP and the cumulative service, (i.e., q_2). By examining the graph, we can see that a bound for the maximum backlog q_{max} occurs at time t_2.

In Fig. 4b, when the real traffic-flow curve intersects with the cumulative service curve at time d_1, the queue becomes empty again. However, since we are using the deterministic EP to approximate the traffic flow, an upper bound on the busy period is found to be the time instance when the EP intersects with the service curve (i.e., at d). In an FCFS queue, for a traffic unit arrives at time t_3, the backlog it sees will be cleared at time t_4 (i.e., when the cumulative service is equal to the cumulative arrivals at time t_3). Therefore, its actual delay is d_2, as shown in the figure. Since EP is used as an upper bound for the traffic flow, the upper bound on delay for this traffic unit is d_3.

According to the definition given by Eq. 2, there could be an arbitrary number of EPs for a given traffic flow. From the examples in Fig. 4, it is easy to see that different EPs will give different backlog and delay bounds for the same traffic flow and service capacity. As a result, tighter bounds on $A(t)$ will always be desirable for achieving more accurate performance bounds. In addition, the delay and backlog bounds are tight in the sense that these bounds are realizable, since in the worst case the cumulative arrival $A(t)$ could be identical to $\hat{A}(\tau)$, that is, the equality holds in Eq. 2. On the other hand, these bounds are loose in the sense that they are the worst-case scenarios that only occur rarely in most applications.

Admission Control Tests — The delay bound discussed in the previous section can be used in admission control tests for deterministic service. Admission control is used in network nodes to keep them from being overloaded. Usually an admission control test is invoked when a new flow request arrives in order to verify that, if the new flow is admitted, the QoS requirements (e.g., a maximum acceptable delay) of existing flows and the new flow will all be satisfied. Otherwise, the new flow request will be rejected.

The deterministic admission control test conditions for various schedulers, that is, FCFS, static priority (SP), and earliest deadline first (EDF), are presented in [18, 22, 60] and are summarized in Table 1. These tests are used to verify if the delay requirements of all the sessions (existing ones and the new one) can be satisfied, and can be used to derive the maximum number of user sessions with various deterministic delay requirements that can be accepted at a network element (called the *admissible region*).

Consider an FCFS queue fed by N source flows, each conforming to a deterministic EP $\hat{A}_i(t)$, $i = 1, \ldots, N$, and requiring a delay no larger than d. Let s_i be the maximum packet size of the i-th flow. Since the scheduler does not distinguish between packets, the delay-bound test simply verifies that the maximum delay will not be larger than the delay requirement d. Suppose the queue is idle at time 0 and is busy thereafter. Since a traffic unit arriving at time t sees a backlog of $\Sigma_{i=1}^{N} A(t)$, its delay is the time it takes to clear the backlog and the largest remaining service time of the packet that is being served at time t, $\max_{1\le i\le N} s_i/c$ (since the service is not preemptive). Replacing the backlog with the sum of EPs will give the admissible condition for the FCFS service discipline (see Eq.

Scheduler	Condition	
FCFS	$d \geq \Sigma_{i=1}^{N} \hat{A}_i(t)/c - t + \max_{1 \leq i \leq N} s_i/c$,	for all $t \geq 0$.
Static Priority	$\exists \tau \leq d_p : t + \tau \geq \Sigma_{i \in C_p} \hat{A}_i(t)/c + \Sigma_{q=1}^{p-1} \Sigma_{i \in C_q} \hat{A}_i(t+\tau)/c - t + \max_{r>p} s_r^{max}/c$,	for all p, $t \geq 0$.
EDF	$t \geq \Sigma_{i=1}^{N} \hat{A}_i(t - d_i)/c + \max_{d_i > t} s_i/c$,	for all $t \geq d_1$.

Table 1. *Delay bound tests for FCFS, EDF, and SP packet scheduler [22].*

11).

In SP systems, each traffic flow is assigned a priority level p, $1 \leq p \leq P$, according to their delay requirements d_p. Usually a higher-priority flow has a tighter delay requirement, that is, $d_p < d_q$, if $p < q$. The system maintains P FCFS queues, and traffic belonging to the same priority level is put into the same FCFS queue. When the server is available, it always serves the first packet in the nonempty FCFS queue with the highest priority.

Let the set of flows with priority p be C_p, and the maximum packet size for priority-p flows be s_p^{max}. Consider a priority-p packet arriving at time t. The packet will be served only after the following two types of traffic are served:

- The priority-p backlog at t (the first term on the RHS of the test condition in Table 1) and
- The higher-priority backlog and the higher-priority traffic that arrived after t but before the tagged packet's service time (the second term on the RHS of the test condition)

This is because such a packet arriving later than the priority-p packet can still be served earlier due to its higher priority. Similarly, the last term on the RHS of the test condition is the largest remaining service time of a lower priority packet that is being served at time t.

In EDF, each packet is assigned a deadline (e.g., its arrival time plus its delay requirement). Packets in the queue are sorted according to their deadlines and the scheduler always serves the packet with the smallest deadline [61]. EDF has been shown to be optimal with respect to schedulability, in the sense that it can provide the highest level of deterministic delay assurance among all scheduling disciplines [60, 62]. The EDF schedulability condition in Table 1 can be interpreted as follows. Consider a packet with delay requirement d_j arriving at time $t - d_j$. The packet should be served before its deadline t, which is possible if the maximum amount of traffic arrives with a tighter deadline smaller than or equal to t, that is, $\Sigma_{i=1}^{N} \hat{A}_i(t - d_i)$ has been cleared before time t. The second term on the RHS of the schedulability condition is again due to the fact that the packet currently in service cannot be preempted.

Performance of Deterministic Envelope Processes

For the class of deterministic EPs, there are two interesting questions to be answered:

- What is the best performance in utilizing network resources that can be achieved using these EPs?
- For the class of piecewise linear EPs, how many $\{\sigma, \rho\}$ pairs (or rate-interval pairs) are needed to achieve a utilization close to the best performance?

These questions have been addressed in [22], via experiments with VBR video traces. To answer the first question, the MEP (see Eq. 4) of MPEG video traces, termed *empirical envelope* in [22], is used in admission control tests. Since the empirical envelope is computed from a given video trace (*Jurassic Park*, MPEG encoded), it requires the entire trace a priori, and is hard to specify or police in practice. Nevertheless, it is the tightest envelope for the specific video trace. Using empirical envelopes in admission control tests will yield the maximum number of acceptable sessions. This can serve as a benchmark for evaluating the performance of other deterministic EPs that are more practical.

The maximum number of admissible sessions using the optimal EDF scheduler is given in Fig. 5a, as compared to peak rate allocation and average rate allocation [22]. It can be seen that the best utilization the deterministic EP can achieve (using the empirical envelope) lies between the peak rate allocation and the average rate allocation. Furthermore, the utilization increases when the delay constraint is relaxed. Although the utilization is low, compared to the average rate

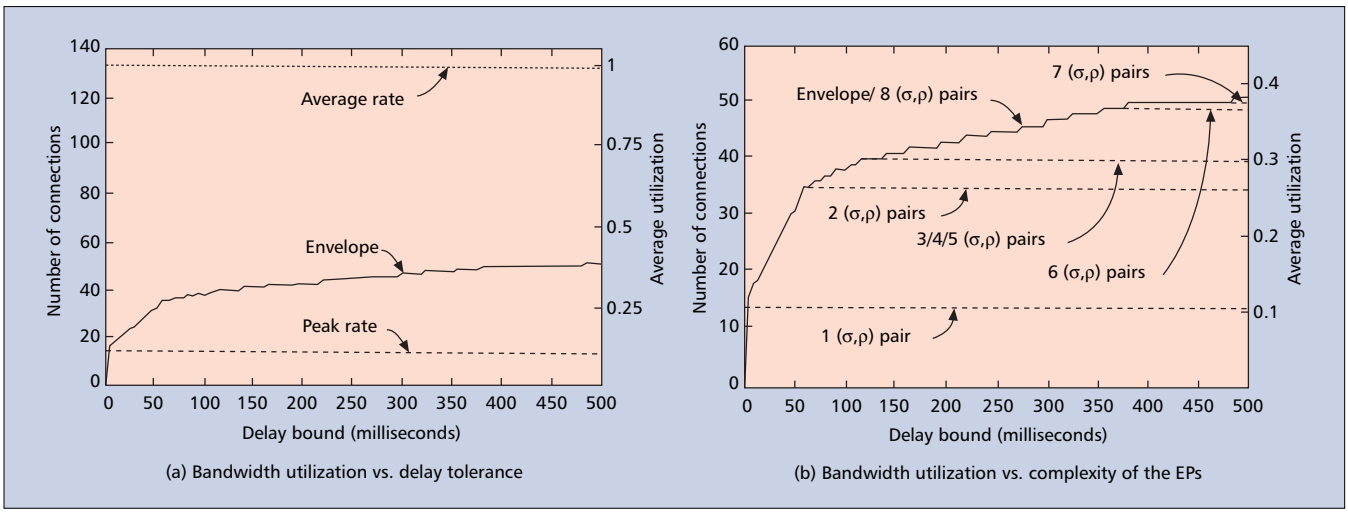

Figure 5. *Performance of deterministic envelope processes [22]. © 1996 IEEE.*

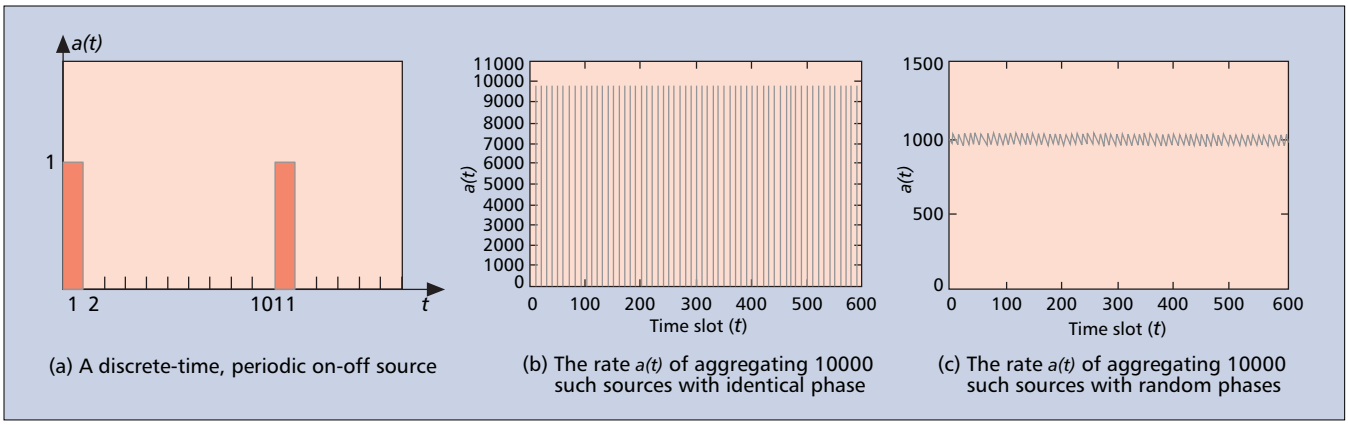

Figure 6. *An illustration of statistical multiplexing gain.*

allocation, there is still significant improvement of peak rate allocation, especially when the delay constraint is relaxed. For example, utilizations ranging from 15 to 30 percent are achievable for a delay bound of 30 ms. This is because in the worst case that all the sessions are transmitting at their peak rates, the excess traffic can be buffered briefly, as long as their delay constraints are met. The more relaxed the delay constraints are, the longer they can be buffered; hence, the smaller service capacity required.

To answer the second question, Fig. 5b in [22] presents the utilization achieved by several $\{\vec{\sigma}, \vec{\rho}\}$ EPs having different numbers of $\{\sigma, \rho\}$ pairs, where the maximum utilization achieved by the empirical envelope serves as a benchmark. The general trend in Fig. 5b is that the greater the number of $\{\sigma, \rho\}$ pairs, the higher the utilization, since a session requires less capacity when a more accurate approximation (i.e., a tighter upper bound) is used in admission control. Another interesting observation is that the additional improvement in utilization decays quickly when more $\{\sigma, \rho\}$ pairs are used. This implies that reasonably high utilization, as compared to the benchmark, can be achieved by using a relatively small number of $\{\sigma, \rho\}$ pairs.

Statistical Multiplexing of Regulated Sources

Statistical Multiplexing — The experimental studies in [22] (see the previous section) provide an idea about the best deterministic performance guarantees that can be achieved by using deterministic EPs (i.e., using the tightest envelope — MEP, and the optimal scheduling discipline — EDF). In order to achieve higher utilization, statistical multiplexing must be explored. On the other hand, deterministic EPs are still highly attractive due to the fact that they are amenable to enforcement. Therefore, a natural solution is *statistical multiplexing of deterministically regulated flows*. A statistical service can improve upon a deterministic service by taking advantage of the statistics of individual source, and the statistical independence of the source traffic flows. Note that independence of the source flows is the most fundamental assumption for the following results.

A simple example is given in Fig. 6. Figure 6a plots a discrete-time, periodic on–off source: it generates one unit of traffic in time slot 1, and then can be silent for the remaining nine time slots, and so forth. Multiplexing N such sources (with identical phases), we get an aggregate traffic flow which is on in time slot 1 and off in the remaining nine time slots, and the aggregate peak rate is N, as shown in Fig. 6b. However, if the sources have independent, random phases (e.g., source 1 is on in time slot 1, while source 2 is on in time slot 5, and so forth), the aggregate traffic is actually smoothed out, as shown in Fig. 6c. The aggregate rate $a(t)$ fluctuates slightly around the average rate $N \times 1/10$. A direct result of this observation is that the amount of resources required to support QoS for N traffic flows will be much smaller than N times the resource required to support QoS for a single traffic flow.

It is worth noting that even with random phases, it is still possible that the worst-case scenario in Fig. 6b will occur, but such events only occur rarely. If we perform admission control according to the "average" or more likely case (i.e., Fig. 6c), a significant improvement in resource utilization can be achieved. When the worse-case scenario happens, there will be a violation of the QoS requirements. Through rigorous modeling and analysis, however, we can guarantee that such a violation occurs with low probability (e.g., 10^{-6}).

Related Work — The analysis work along this line generally takes two steps. First, find the worst-case traffic flow, generated by the so-called *adversarial* sources, that maximizes the resources required (in the hard QoS guarantee case) or the QoS violation probability (in the soft QoS guarantee case), which, however, still conforms to the deterministic EP. Such adversarial sources are extremal, periodic on-off processes with a random phase for bufferless multiplexers [26, 63]. For buffered multiplexers, the adversarial traffic pattern is found to be periodic, with multiple "on" phases and a different rate in each "on" phase [63–65]. Second, take advantage of the random *phases* of the adversarial sources and multiplex them statistically.

In [23], admission control for a network node with finite buffer B and capacity C is studied. The joint allocation of buffer and capacity is first reduced to a single resource allocation problem by the notion of a virtual buffer/trunk system. Then the derived single resource allocation problem is solved by applying the known results on bufferless multiplexing, where the Chernoff Bound [66] is used to estimate the loss probability.

In [28], it was found that there is no benefit in resource sharing for lossless multiplexing. Furthermore, for deterministic QoS guarantees, there is a unique timescale determined by the system parameters and the input flows, which determines the optimal buffer/bandwidth trade-off. For statistical multiplexing, [28] transforms the two-resource allocation problem into two independent single-resource allocation problems, which is shown to achieve higher multiplexing gain than [26].

In [29], the multiplexer was examined under the more general $\{\vec{\sigma}, \vec{\rho}\}$ EP. An interesting result in [29] is that although the extremal, periodic on-off source is adversarial for a bufferless multiplexer, it is not adversarial for the transformed two-resource allocation problem. In [31], a bufferless multiplexing system is studied, in which each regulated flow is first smoothed by a smoother with a capacity that is determined by

the delay requirement of that source. This novel structure enables end-to-end statistical guarantees [30], since the bufferless multiplexer ensures that a flow still keeps the same description after leaving the system. There is no need to implement per-node traffic shaping [27] with this framework.

In [23], local and global effective envelopes for the aggregate regulated flows are constructed using the Chernoff Bound and the Central Limit Theorem (CLT) [67]. Applying the effective envelopes in admission control tests further improves the admissible region [23, 68]. More details on effective envelopes are provided below.

Finally, a multiplexer with general flows is studied in [24] using a CLT approach and in [25] using large deviation theory, respectively. Although these papers do not assume regulated flows, their results can be applied to the multiplexing of regulated flows. The maximum variance bounds in [24] is found to achieve the highest utilization, as compared to several other approaches in a comparison study of admission control schemes [12].

Probabilistic Envelope Processes

Probabilistic Envelope Processes and Soft QoS Guarantees

In the previous section we discussed deterministic EPs which can be used for provisioning of deterministic or "hard" performance guarantees, such as worst-case delay bounds and no packet dropped in the network. Generally, such deterministic services are quite conservative since they are based on worst-case analysis [22]. In addition, hard performance guarantees might be an overkill for many applications, say, multimedia communications, where a certain amount of loss or delay violation is tolerable.

As opposed to the deterministic service approach, a statistical service provides *probabilistic* service assurances. For example, such "soft" QoS guarantees can be in the following forms:

$$\Pr\{\text{delay} \geq d\} \leq \varepsilon \text{ or } \Pr\{\text{loss} \geq l\} \leq \varepsilon, \quad (12)$$

where ε is the probability that the delay or loss bound d or l is violated and is generally dependent on the specific application.

By allowing a fraction of traffic to violate its QoS requirement, a user can trade-off the QoS he receives with the network resources required. More importantly, such a probabilistic approach can significantly increase network resource utilization. In the previous section, we have seen significant improvements when statistical multiplexing is explored, even though each source is deterministically regulated. An alternative approach in soft QoS provisioning is to use the class of probabilistic EPs that bound traffic flows in a probabilistic manner. In the following, we will survey representative probabilistic EPs and the probabilistic QoS guarantees they provide.

Bounded Burstiness Processes

Definition — The piecewise linear EPs discussed in the previous section extend the $\{\sigma, \rho\}$ EP by using multiple $\{\sigma, \rho\}$ pairs. Both classes of EPs bound the flows deterministically, (i.e., Eq. 2 always holds true). Another dimension of extending Cruz's work is to allow the EPs to be violated (i.e., letting Eq. 2 hold true in a probabilistic sense). The cumulative traffic $A(t)$ is allowed to exceed its EP $\hat{A}(\tau)$, but with a probability decaying with the degree that the EP is exceeded. This class of EPs are known as the Bounded Burstiness Processes [32].

Let F be the set of functions $f(\sigma)$ such that for any order n, the n-fold integral

$$\underbrace{\int_{\sigma}^{\infty} \cdots \int_{\sigma}^{\infty}}_{n \text{ times}} f(u)(du)^n$$

is bounded for any $\sigma > 0$. An arrival process $A(t)$ has a stochastically bounded burstiness (SBB) with upper rate ρ and bounding function $f(\sigma)$ if

- $f(\sigma) \in \mathsf{F}$; and
- $\Pr\{A(s, t) \geq \rho(t - s) + \sigma\} \leq f(\sigma)$, for all $\sigma \geq 0$ and all $t \geq 0$.

As will become clear shortly, the first condition in the SBB definition is required for the *closure* property: a traffic flow (or its description) will not explode as it travels through one network node after another. Also note that the finiteness of the n-fold integral implies that $f(\sigma)$ is a decaying function (otherwise the integral from σ to ∞ will not exist). The second condition specifies that the probability that $A(t)$ exceeds the $\{\sigma, \rho\}$ EP decays with $f(\sigma)$: the larger the σ, the smaller the violation probability.

Furthermore, a stochastic process $q(t)$, that is, the backlog process, is stochastically bounded (SB) with bounding function $f(\sigma)$ if

- $f(\sigma) \in \mathsf{F}$; and
- $\Pr\{q(t) \geq \sigma\} \leq f(\sigma)$, for all $\sigma \geq 0$ and all $t \geq 0$.

As in the SBB definition, the first condition above is also required for the closure property. The second condition in the above definition specifies that the probability that $q(t)$ exceeds σ (i.e., the probability that the backlog is larger than σ) decays with $f(\sigma)$: the larger the σ, the smaller the violation probability.

Since $f(\sigma)$ is general in both definitions, one can choose different forms of $f(\sigma)$ for traffic flows with different characteristics. For example, an exponential bounding function can serve a good model for short range dependent (SRD) traffic flows,[5] while the sum of exponentials or a Weibullian bounding function can be used to characterize LRD traffic that exhibits burstiness over multiple timescales.

The SBB Calculus — Based on the above definitions, Starobinski and Sidi present a network calculus for SBB processes [32], which is very useful in analyzing a feedforward network[6] with SBB arrivals, such as proving the stability of such networks and deriving QoS performance measures.[7] The SBB calculus is summarized in the following.

Summation — Let $A_i(t)$ be SBB with $\{\rho_i, f_i(\sigma)\}$, $i = 1, 2$. Then the sum of these two SBB processes $A_1(t) + A_2(t)$ is also SBB with $\{\rho_1 + \rho_2, g(\sigma)\}$, where $g(\sigma) = f_1(p\sigma) + f_2((1 - p)\sigma)$ and p is any real number in (0, 1).

This result is derived via some basic properties of random variables. Consider two random variables X_1 and X_2. For a constant x, the set of events that $\{X_1 + X_2 \geq x\}$ is a subset of $\{X_1 \geq px\} \cup \{X_2 \geq (1 - p)x\}$, for all $0 < p < 1$. Therefore, the

[5] *An SRD process has an autocorrelation function that decays exponentially.*

[6] *A feedforward network is a network in which there are no cycles in the graph generated by the routes of various sessions. This concept depends on both the topology of the network and the routing scheme applied.*

[7] *A queueing network is stable if the queues do not increase without bound, that is,* $\lim_{\sigma \to \infty} Pr\{q_i \geq \sigma\} = 0$, *where* q_i *corresponds to the steady-state workload in the* i*-th network element.*

probability of the former event should be no larger than that of the latter, that is, $\Pr\{X_1 + X_2 \geq x\} \leq \Pr\{[X_1 \geq px] \cup [X_2 \geq (1-p)x]\}$. Since the probability of the union of two events is upper bounded by the sum of the probabilities of the two events, we have $\Pr\{(X_1 + X_2) \geq x\} \leq \Pr\{X_1 \geq px\} + \Pr\{X_2 \geq (1-p)x\}$. Now if both X_1 and X_2 are SBB, substituting the definition in the previous section, we get the summation result. Note that the derivation does not require the two sources be independent.

Characterization — For a work-conserving system with service rate ρ, if the backlog $q(t)$ is SB with bounding function $f(\sigma)$, that is, $\Pr\{q(t) \geq \sigma\} \leq f(\sigma)$, then the input process of the system is SBB with upper rate ρ and the same bounding function $f(\sigma)$, that is, $\Pr\{A(t-s) \geq \rho(t-s) + \sigma\} \leq f(\sigma)$.

This result shows the relation between the input process and the backlog process. It is derived from the Lindley's equation. From Eq. 7, we have $q(t) \geq A(t-s) - (t-s)c$. From the SB definition, we have $\Pr\{A(t-s) - (t-s)c \geq \sigma\} \leq f(\sigma)$, which gives the characterization result.

Calculus for an Isolated Network Element — For a work conserving queueing network element with service rate c, if the input process $A(t)$ is SBB with $\{\rho, f(\sigma)\}$, then the output process $B(t)$ is SBB with $\{\rho, g(\sigma)\}$, where $g(\sigma) = f(\sigma) + \{1/c-\rho\} \int_{\sigma}^{\infty} f(\mu)d\mu$, and the backlog process $q(t)$ is SB with the same bounding function $g(\sigma)$.

The proof of these results is more mathematically involved. We omit the proof here and refer interested readers to [42] for more details. There are several interesting implications from these results. First, if exogenous inputs to a network with work-conserving elements are SBB, then all the traffic flows within the network are SBB and all the backlogs in the network elements are SB (i.e., the closure property is conserved). Second, the SBB calculus is quite general, since there is no assumption made on flow independence and scheduling disciplines (except that the servers are work-conserving). On the other hand, if further information on scheduling and independence is available, a finer analysis may result in better performance bounds [32, 42].

Third, the last result gives the backlog distribution, that is, SB with $g(\sigma)$ if the arrival process is SBB with $\{\rho, f(\sigma)\}$. This can serve as a good approximation of loss probability for finite buffer systems [69]. Other performance measures such as delay distribution can also be derived from this result. Finally, the SBB calculus greatly simplifies the analysis of any feedforward network fed by external SBB processes, since the end-to-end analysis can be reduced to the analysis of each isolated node along the route inductively. Also note that every time when a flow traverses a network element, there will be an additional integration over $[\sigma, \infty)$ on the bounding function. For the traffic flow to remain bounded after traversing n network elements, the n-fold integral of $f(\sigma)$ should be bounded (see the definition of SBB and SB earlier in this section). More discussion on end-to-end analysis will be presented later.

Exponential Bounded Burstiness — Exponential Bounded Burstiness (EBB) processes are SBBs with a bounding function $f(\sigma) = \phi e^{-\alpha\sigma}$, where ϕ and α are source-specific constants [33]. With this definition, the probability that the input process exceeds the linear EP $\hat{A}(\tau) = \rho t + \sigma$ decays exponentially or faster than exponentially. Obviously, this is true for all SRD traffic.

Interesting results on EBB calculus have been established in [33]. In addition, since no independence assumption is required when applying the analysis in [33], both feedforward networks and general cyclic networks can be analyzed. Performance of a GPS server (and a GPS network) with EBB sessions can be found in [70].

Sum of Exponentials — This is a more general class of bounded burstiness processes than EBB. A Sum of Exponentials Bounded Burstiness process has a bounding function in the form of the summation of a number of exponentials, that is, $f(\sigma) = \Sigma_{k=1}^{K} \phi_k e^{-\alpha_k\sigma}$, where ϕ_k and α_k are constants, for $k = 1, \ldots, K$ [32]. The motivation behind this model is the existence of multiple timescales in network traffic and the observation that, on logarithmic scale, the delay distribution in a multiplexer can be roughly broken into two linear regions, called the cell region and the burst region [71]. Using multiple exponential bounds on the tail distribution of a queue can capture this multiple timescale phenomenon and provide a tighter bound than EBB.

One problem with the calculus of sums of exponentials is that the number of exponentials required to bound the burstiness of a process within the network grows each time the process merges with another process (i.e., summation). As a result, this traffic model does not have the closure property. In [32], Starobinski and Sidi present an ad hoc procedure, which provides a simple way to derive bounding functions using the sum of two exponentials, but at the cost of lower accuracy.

Another interesting example of SBB is the Weibull Bounded Burstiness Processes (WBB), which has a Weibullian bounding function [34, 42]. Since WBB also belongs to the the class of EPs for self-similar traffic, we present its discussion later in this article.

Chang's Log-Moment Generating Function Bounds

In addition to bounding the burstiness of a traffic flow, probabilistic QoS guarantees can also be achieved by deterministically bounding the moment generating function of $A(t)$ [21].[8] Consider a random variable X. If its moment-generating function is bounded by a finite constant ψ as $E(e^{\theta X}) \leq \psi^{\theta}$, then from the Chernoff bound [66], its distribution is bounded exponentially with respect to θ as:

$$\Pr\{X \geq x\} \leq \psi^{\theta} e^{-\theta x}, \text{ for all } x > 0. \tag{13}$$

Since the cumulative arrival $A(t)$ is a random variable, its moment generating function can thus be bounded by a deterministic function $\hat{A}(\theta, t)$ as follows:

$$\frac{1}{\theta}\log\, Ee^{\theta A(t_1,t_2)} \leq \hat{A}(\theta, t_2 - t_1), \forall t_1 \leq t_2, \tag{14}$$

where $\hat{A}(\theta, t)$ is called an EP of $A(t)$ with respect to θ [21]. As in the deterministic case, the MEP of $A(t)$ is $A^*(\theta, t) = \sup_{s \geq 0}\{(1/\theta)\log E \exp[\theta A(s, s+t)]\}$, and the MER of $A(t)$ is $a^*(\theta) = \lim \sup_{t\to\infty} A^*(\theta, t)/t$. It is shown in [21] that the MER of $A(t)$ is an increasing function of θ: from the mean rate to the peak rate, as θ increases from 0 to ∞.

Applying the Chernoff bound Eq. 13 and the Lindley's equation (Eq. 6), the log-moment generating function EP can be used to derive QoS performance measures at a network element. Consider a queue fed by a flow with a linear EP, $\hat{A}(\tau) = \rho(\theta)t + \sigma(\theta)$. Assume $\rho(\theta) < c$ to make a stable system. We have the following interesting results from [21]:

Backlog Distribution — The total backlog $q(t)$ is bounded exponentially with respect to θ, that is, $\Pr\{q(t) \geq x\} \leq \beta(\theta)e^{-\theta x}$, where $\beta(\theta)$ is a constant.

[8] *The moment-generating function of a random variable* X *is defined to be the expectation of* $e^{\theta X}$*, that is,* $E\{e^{\theta X}\}$.

Delay Distribution — If the FCFS scheduling policy is used, the delay distribution is also bounded exponentially as $\Pr\{d(t) \geq d\} \leq e^{\theta\rho(\theta)}\,\beta(\theta)e^{-\theta c(d-1)}$.

Input–Output Relationship — The departure process $b(t)$ has an MEP $\hat{B}(\theta, t) \leq \rho(\theta)t + 1/\theta \log\beta(\theta)$. The departure process still has the same long-term average rate as the input process, but with a possibly different burst factor.

Again, there is the closure property: if the input processes to such a network have their log-moment generating functions bounded, then every process in the network (e.g., delay, backlog) and the arrival processes at internal nodes has an exponentially bounded distribution. These are useful results for soft QoS guarantees, from which one can estimate how much network resources (e.g., buffer or bandwidth) are required to achieve a desired buffer overflow probability (or a delay distribution), and conveniently trade-off the QoS violation probability and the network resource required.

Kurose's Bounds

In [35], Kurose presented a framework for soft QoS provisioning, which is based on EPs with a family of bounding random variables. With this framework, a traffic flow is characterized by a family of random variables $\{B(t_1), B(t_2), \ldots\}$ that stochastically bounds the source over the respective interval lengths t_k, $k = 1, 2, \ldots$.

Such a bounding approach is somewhat similar to D-BIND, where multiple rate-interval pairs are used to bound the cumulative arrival $A(t)$. The important difference here is that the D-BIND EP will not be violated, while Kurose's EP bounds the traffic flow in a probabilistic manner. More specifically, a random variable X is said to be *stochastically larger* than a random variable Y (denoted as $X \succ_{st} Y$) if and only if $\Pr\{X > x\} \geq \Pr\{Y > x\}$ for all x [72]. Therefore, in the definition of Kurose's bounds, the cumulative traffic in any time interval t_k is stochastically bounded by the corresponding random variable $B(t_k)$, that is, $\Pr\{B(t_k) > x\} \geq \Pr\{A(\tau, \tau + t_k\} > x)$, or $B(t_k) \succ_{st} A(\tau, \tau + t_k)$, $\forall\tau$.

Consider an FCFS multiplexer with a link speed of c and serving N flows. Each flow is characterized by its respective family of bounding random variables. A stochastic bound on the delay in this system is shown to be

$$\Pr\{D > d\} \leq \max_{0\leq t_k\leq\beta} \Pr\left\{\sum_{i=1}^{N} B_i(t_k) - ct_k \geq cd\right\}, \tag{15}$$

where β is an upper bound on the busy period and is the smallest nonnegative value such that $\Sigma_{i=1}^{N}|B_i(\beta)| \leq c\beta$ [35]. An intuitive explanation of Eq. 15 is that if the backlog seen by a tagged packet [see Lindley's equation (Eq. 8)] is larger than cd, then it will take the FCFS server more than d time units to clear the backlog before serving the tagged packet, and the delay bound of the tagged packet will be thus violated in this case.

When the input is bounded by a set of random variable-time interval pairs, the output traffic of a work-conserving network element is also bounded by the same set of random variables, but over a different (smaller) interval of time. With this result, we can derive a flow's characterizations along the hops from its ingress point to the network node being studied. Heuristic algorithms are presented in [35] to compute end-to-end delay guarantees on a per-session basis, which, however, usually give loose results as compared with simulation. Furthermore, to use Eq. 15, we need to compute the convolution of N random variables for each time interval t_k, which results in a high computation complexity [36].

H-BIND

The Hybrid Bounding Interval Dependent (H-BIND) approach provides statistical QoS guarantees by exploiting the random phases of deterministically constrained flows [20, 36]. In H-BIND, each session is regulated by a D-BIND EP and a large number of flows are multiplexed at a network element. When computing the delay distribution using Eq. 15, the bounding random variables, $B_i(t_k)$, are assumed to be Gaussian (reasonable when the number of flows is large), which is fully determined by its mean and variance [67]. For $B_i(t_k)$, the mean is calculated from the source's D-BIND EP. Since there are many arrival processes that conform to a D-BIND EP, the variance $\sigma^2(t_k)$ is computed using the arriving process that maximizes it (i.e., an adversarial source).

In [20], Knightly also extended the FCFS stochastic bound in Eq. 15 to the static priority case. Let there be P priorities, C_p be the set of flows with priority level p, and the delay requirement for priority p connections be d_p. Then the delay violation probability for a priority-p packet is bounded by [20]:

$$\Pr\{D_p > d_p\} \leq \max_{0\leq t_k\leq\beta} \Pr\left\{\sum_{i\in\mathcal{C}_p} B_{p,i}(t_k) + \sum_{q=1}^{p-1}\sum_{i\in\mathcal{C}_p} B_{q,i}(t_k + d_p) - ct_k \geq cd_p\right\}, \tag{16}$$

where β_p is a bound on the priority-p busy period. This is similar to the deterministic admission control test for static priority systems discussed earlier. However, the RHS of Eq. 16 now provides a bound on the delay *distribution* of priority-p traffic.

For H-BIND, the Gaussian assumption eliminates the convolution computation required when Kurose's bound is used. Such an approximation can be quite accurate, considering the large number of flows multiplexed at a core router. Moreover, when used for admission control, H-BIND achieves bandwidth utilization of up to 86 percent in a realistic scenario [36], which is much higher than the 15 to 30 percent bandwidth utilization achievable with deterministic EPs [22]. Such sigificant gain is due to the fact that statistical multiplexing is explored in H-BIND. In addition, the D-BIND/H-BIND framework allows for the simultaneous support of hard and soft QoS guarantees. With this framework, all the sources are constrained by D-BIND EPs. Sources demanding deterministic service are served with higher priority than sources choosing statistical service.

It is also worth noting that H-BIND is essentially a single-multiplexer analysis. In [20], Knightly discusses how to extend this analysis to derive end-to-end performance bounds. Since H-BIND explores statistical multiplexing gain, flow independence is an important assumption, which may not hold true when flows share a common buffer.[9] Knightly suggests adopting delay-jitter control (or the simpler rate-jitter control) at every network node, which decouples the network nodes along an end-to-end path by reconstructing the sources' original traffic pattern at each hop [73, 74]. With this technique, if a delay bound d_h is provided at hop h with violation probability ε_h, the end-to-end delay violation probability is found to be $\Pr\{D^{e2e} > \Sigma d_h\} \leq \Pi\varepsilon_h$ [20].

Rate Variance Envelope Processes

[9] *This problem is solved in [31] by adopting bufferless multiplexing systems.*

Rather than bounding the cumulative traffic $A(t)$ itself, the Rate Variance EPs are used to bound the statistical properties of $A(s, s + t)$ as a function of the interval length t [36, 37]. The rate-variance of a flow,

$$RV(t_k) \underline{\underline{def}} Var\left\{\frac{A(s,s+t_k)}{t_k}\right\}, k=1,2,\ldots,K, \quad (17)$$

describes the variance of a stream's arrival rate over intervals of length t_k. This characterization captures the second moment correlation structure of an arrival process. Note that it makes no assumption on $A(t)$ and allows for an arbitrary autocorrelation structure of individual sessions. Using video traces as examples, Knightly [36] shows that the $RV(t_k)$ curves of the video traces can be bound or approximated using two or three piecewise linear segments on the log-log scale.

For admission control tests, the Rate Variance EP of the aggregate traffic $\Sigma_i A_i(s, s + t_k)$ can be approximated using a Gaussian Envelope with variance $\Sigma_i t_k^2 RV_i(t_k)$ over intervals of length t_k. That is, the summations in the RHS of Eqs. 15 and 16 can be approximated with Gaussian random variables and the probabilities computed using a Gaussian distribution with the corresponding mean and variance.

The Rate Variance EPs have been shown to achieve very high utilization in [12] as compared to other methods. However, it is more difficult to enforce a flow to follow such EPs. Reference [75] presents a measurement-based admission control scheme based on Rate Variance EPs, where the maximal rate envelope of the aggregate flow is adaptively measured over the currently admitted flows.

Effective Envelopes

The class of effective envelopes are functions which upper bound multiplexed traffic with high certainty [23]. As in H-BIND, the traffic flows are assumed to be deterministically regulated, e.g., by piecewise linear deterministic EPs, and possess several general properties, such as stationarity, independence, additivity and subadditivity.

Two types of effective envelopes are defined in [23], namely, a *local effective envelope* and a *global effective envelope*. Consider a set of flows $\mathcal{C}$ with arrival functions $A_i(t)$. The aggregate traffic is $A_{\mathcal{C}}(t, t + \tau) = \Sigma_{\mathcal{C}} A_i(t, t + \tau)$. The local effective envelope that upper bounds $A_{\mathcal{C}}(t, t + \tau)$ is a function $\mathcal{G}_{\mathcal{C}}(.;\varepsilon)$ that satisfies

$$\Pr\{A_{\mathcal{C}}(t, t + \tau) \le \mathcal{G}_{\mathcal{C}}(\tau; \varepsilon)\} \ge 1 - \varepsilon, \forall\tau \ge 0 \text{ and } \forall t. \quad (18)$$

That is, a local effective envelope upper bounds the aggregate traffic for any specific ("local") time interval of length τ [23]. Furthermore, it is a probabilistic bound: the aggregate traffic is allowed to exceed the local effective envelope, but with a small probability (at most, ε).

Global effective envelopes are defined for the same aggregate traffic $A_{\mathcal{C}}(t, t + \tau)$, but are bounds for the arrival in *all* subintervals $[t, t + \tau)$ of a larger interval. More precisely, a global effective envelope for an interval of length β is a subadditive function $\mathcal{H}_{\mathcal{C}}(.; \beta; \varepsilon)$ such that

$$\Pr\{\mathcal{E}_{\mathcal{C}}(\tau;\beta) \le \mathcal{H}_{\mathcal{C}}(\tau;\beta;\varepsilon), \forall 0 \le \tau \le \beta\} \ge 1 - \varepsilon, \quad (19)$$

where $\mathcal{E}_{\mathcal{C}}(\tau;\beta)$ is the empirical envelope of $A_{\mathcal{C}}(t, t + \tau)$, and β could be an upper bound on the largest system busy period. $\mathcal{H}_{\mathcal{C}}(\tau; \beta; \varepsilon)$ is a bound for traffic for all subintervals of length $\tau \le \beta$ in the interval β, which is more stringent than local effective envelopes and leads to more conservative admission control [23].

For a given set of flows $A_i(t)$ and their corresponding deterministic EPs $\hat{A}_i(t)$, the effective envelopes can be computed as follows [23]. First, the bound on the moment generating function of an individual flow is computed, which is found to be a function of the corresponding deterministic EP $\hat{A}_i(t)$. Due to the independence assumption, the bound on the moment generating function of the aggregate traffic flow, $M_{\mathcal{C}}(s, \tau)$, can be easily derived, as the product of those of the individual flows. Second, applying Chernoff bound [see Eq. 13], we have

$$\Pr\{A_{\mathcal{C}}(0, \tau) \ge Nx\} \le e^{-Nxs} M_{\mathcal{C}}(s, \tau). \quad (20)$$

Substituting the moment generating bound derived in the first step, the local effective envelope can be obtained by solving for Nx from Eq. 20.

Once the local effective envelope is derived, Boorstyn *et al.* use a geometric argument to construct the global effective envelope $\mathcal{H}_{\mathcal{C}}$ from the local effective envelope $\mathcal{G}_{\mathcal{C}}$. Specifically, they show that $\mathcal{H}_{\mathcal{C}}$ can be bounded by $\mathcal{G}_{\mathcal{C}}$ from both sides as

$$\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) \le \mathcal{H}_{\mathcal{C}}(\tau;\beta;\varepsilon) \le \mathcal{G}_{\mathcal{C}}(\tau'; \varepsilon'), \quad (21)$$

where $\tau'/\tau > 1$ and $\varepsilon'/\varepsilon < 1$ depend on the interval β. It has been shown that for ε sufficiently small and β not too large, $\tau'/\tau \approx 1$, and the resulting global effective envelope is reasonably close to the local effective envelope.

Effective envelopes thus derived can be applied to provide statistical service assurances for various traffic scheduling algorithms. In the following we use FCFS as an example; other scheduling disciplines such as static priority or EDF can be derived similarly and we refer interested readers to [23] for these results. The schedulability condition for FCFS with a statistical delay service $\Pr\{D \ge d\} \le \varepsilon$ is

$$\Pr\left\{\sup_{\hat{\tau}}\{A_{\mathcal{C}}(t-\tau,t) - c\hat{\tau}\} \le cd\right\} \ge 1-\varepsilon, \quad (22)$$

where $t - \hat{\tau}$ is the last time before t that the queue is empty. That is, when a tagged traffic unit arrives at time t, it sees a backlog of $\sup_{\hat{\tau}}\{A_{\mathcal{C}}(t - \hat{\tau}, t) - c\hat{\tau}\}$ [see the Lindley's equation (Eq. 8)]. To provide the required statistical service, the probability that the server can clear this backlog in d time units should be no smaller than $1 - \varepsilon$. Furthermore, this schedulability condition can be approximated by[10]

$$\sup_{\hat{\tau}} \Pr\{A_{\mathcal{C}}(t-\hat{\tau},t) - c\hat{\tau} \le cd\} \ge 1-\varepsilon, \quad (23)$$

From the definition Eq. 18, $\mathcal{G}_{\mathcal{C}}(\tau;\varepsilon) < x$ implies that $\Pr\{A_{\mathcal{C}}(t, t + \tau) > x\} < \varepsilon$. Consequently, the schedulability condition can be rewritten as

$$\sup_{\hat{\tau}}\{\mathcal{G}_{\mathcal{C}}(\hat{\tau};\varepsilon) - c\hat{\tau}\} \le cd. \quad (24)$$

When the global effective envelope is used in admission control test, the schedulability condition is found to be

$$\sup_{\hat{\tau}}\{\mathcal{H}_{\mathcal{C}}(\hat{\tau};\beta;\varepsilon) - c\hat{\tau}\} \le cd, \quad (25)$$

which has a similar structure to Eq. 24.

Finally, we use Fig. 7 from [23] to illustrate the performance of several EPs discussed thus far in admission control. In this example two classes of flows are multiplexed at a network element with a service rate of 45 Mb/s. The delay bounds are 100 and 10 ms for Class 1 and 2 flows, respectively, and the delay violation tolerance is chosen to be $\varepsilon = 10^{-6}$.

[10] *This approximation is accurate if the arrivals follow a Gaussian process [12, 23]. For general cases, the left-hand side (LHS) of Eq. 23 is an upper bound for the LHS of Eq. 22.*

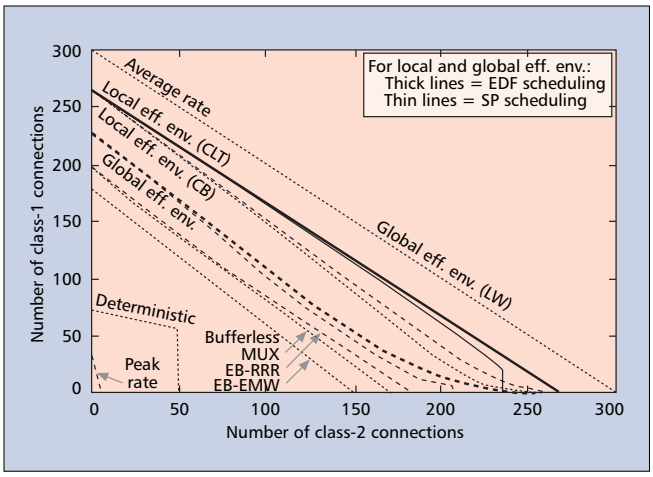

Figure 7. *Admissible region of multiplexing Class 1 and Class 2 flows with* $\varepsilon = 10^{-6}$*, delay deadlines* $d_1 = 100$ *ms, and* $d_2 = 10$ *ms [23]. © 2000 IEEE.*

The admissible regions obtained with various EPs and the static priority and EDF scheduling algorithms are plotted in Fig. 7.

In Fig. 7, the admissible regions of peak rate allocation and average rate allocation serve as benchmarks. The admissible region of any feasible algorithm should fall in between these two regions. We can see the admissible region of deterministic service is much larger than that of peak rate allocation. This is because with deterministic service, when the instantaneous aggregate rate is large than c, the extra traffic can be temporarily buffered, as long as they are served before their delay deadlines. It can also be observed that the schemes that explore statistical multiplexing gain, that is, EB-EMW [26], EB-RRR [29], Bufferless MUX [31], and local and global effective envelopes [23], achieve much larger admissible regions (indicating high bandwidth utilizations). The admissible regions for other probabilistic EPs that do not exploit statistical multiplexing gains, such as SBB, EBB, and Chang's log-moment generation bounds (although not shown in the figure), are expected to lie between the deterministic service curve and the EB-EMW curve in Fig. 7.

Envelopes for Self-Similar Traffic

Since the seminal work [40], many empirical studies have shown that network data and video traffic are LRD or self-similar in that they exhibit high burstiness over multiple timescales [38, 39, 41]. In [43], Norros introduces a framework to model the connectionless data traffic using fractal Brownian motion (fBm) models. This framework inspire the Weibull Bounded Burstiness EPs [34, 42] and is the basis for the fBm EP [15].

Weibull Bounded Burstiness Processes

In the past decade, tremendous effort has been made to build traffic models that not only model the statistical aspect of self-similar traffic, but also are manageable for traffic engineering. The sum of exponentials model tries to model the LRD characteristics within the traditional Markovian analysis paradigm, while some other traffic models have been proposed to model LRD and self-similarity in a more direct manner.

Previous work, such as [43, 76], shows that a single server FCFS queue fed by self-similar traffic, such as fBm traffic, has a Weibullian asymptotic tail. Motivated by this finding, and in a way analogous to the EBB model, the Weibull Bounded Burstiness (WBB) traffic model is proposed for fBm traffic flows [34, 42].

Weibull Bounded (WB) and WBB processes with Hurst parameter H are defined as a special type of SBB, with a Weibullian bounding function $f(\sigma) = \phi e^{-\alpha\sigma^{2(1-H)}}$, where ϕ is called the asymptotic constant and α the decay rate. This is consistent with [76] that showed that the buffer overflow probability of an FCFS queue fed with fBm traffic has the same form as $f(\sigma)$. Since WBB processes belong to the class of SBB processes, the SBB calculus can also be applied to WBB processes. Analytical results for GPS systems with WBB traffic flows can be found in [34].

The fBm EP

Consider a Brownian motion (BM) process $A(t)$ with mean ρ and variance σ^2, an EP $\hat{A}(t)$ that tightly bounds this process is found to be $\hat{A}(\tau) = \rho t + \kappa\sigma t^{\{1/2\}}$, where κ determines the probability that $A(t)$ exceeds $\hat{A}(\tau)$ at time t, that is, if $\Pr\{A(t) > \hat{A}(t)\} = \varepsilon$, then [67]:

$$\kappa = \sqrt{-2\log\varepsilon}$$

This approach can be extended to deal with fBm traffic. Consider an fBm traffic flow $A_H(t)$ with mean ρ, variance σ^2, and Hurst parameter H. Fonseca, Mayor, and Neto [15] present an fBm EP that bounds this fBm process:

$$\hat{A}_H(t) = \rho t + \kappa\sigma t^H, \text{ for } 1/2 \le H \le 1. \quad (26)$$

Similarly, here κ also determines the probability that $A_H(t)$ exceeds $\hat{A}_H(t)$ at time t, that is, if $\Pr\{A_H(t) > \hat{A}_H(t)\} = \varepsilon$, then

$$\kappa = \sqrt{-2\log\varepsilon}$$

Equation 26 has a similar format as Cruz's EP $\hat{A}(\tau) = \rho t + \sigma$: a constant rate process increased by a burst factor. But the burst factor is a constant in Cruz's EP, while an increasing function of t in Eq. 26. Such a burst factor captures the bursty nature of fBm arrival processes.

As in the effective envelope case, given a small ε, we can compute an EP which is exceeded by the traffic flow with probability ε. The computation is however much simpler, since we only need to set

$$\kappa = \sqrt{-2\log\varepsilon}$$

in Eq. 26. It is also worth noting the similarity between the definition of fBm EPs and that of local effective envelopes (Eq. 18). As a result, the schedulability conditions developed for local effective envelopes (e.g., Eq. 24) can also be used for fBm EPs for statistical service assurances. Furthermore, this condition is exact (rather than an approximation), since the traffic flow is Gaussian [12, 23].

The Fractal Leaky Bucket

The Inadequacy of the Leaky-Bucket Regulator — An ideal leaky-bucket regulator should accept all conforming packets, but drop or mark nonconforming packets. It is shown in [15] that it is very hard to choose the leaky-bucket parameters regulating an fBm traffic flow. A set of chosen parameters either gives low utilization, or results in extremely large buffer and delay.

As an example, Fig. 8 plots the buffer size required, B, under various utilizations, defined as the ratio of the mean rate of the fBm traffic ρ to the long-term rate of the leaky-bucket regulator r, that is, $u = \rho/r$. It can be seen that B increases exponentially with u. To achieve an acceptable uti-

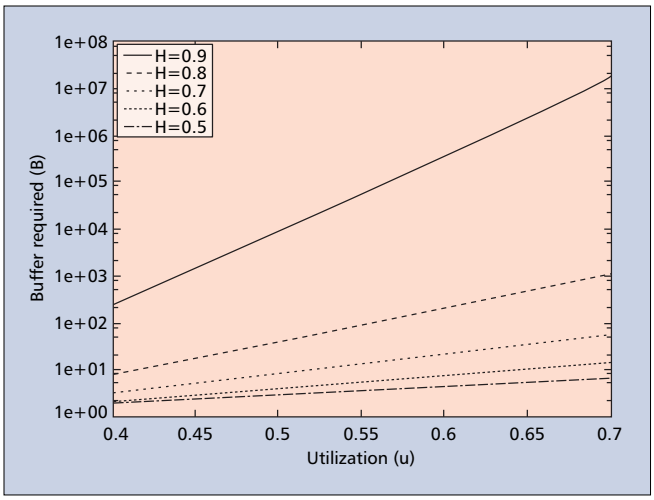

Figure 8. *The trade-off between buffer size (*B*) and utilization (*u = ρ/r*) when using a leaky bucket regulator for fBm traffic.*

lization, an impracticably large buffer is required. The buffering delay in the regulator may also be too large to be acceptable.

The Fractal Leaky Bucket — The inadequacy of a leaky bucket stems from the inherent assumption that the traffic behaves as a linear function of time, while the fBm cumulative traffic is not a linear function of time, since its EP contains the nonlinear term t^H. To address this problem, Fonseca, Mayer, and Neto propose a fractal leaky-bucket model [15]. The amount of traffic accepted by the fractal leaky bucket for an fBm flow characterized with $\{\rho, \sigma, H\}$ is given by

$$\hat{A}(\tau) = \rho t + \kappa\sigma t^H + B. \quad (27)$$

This fractal leaky bucket works as follows. At the beginning, it monitors the cumulative traffic in a basic time window of length τ time units. If the monitored amount exceeds the average $\rho\tau$, the monitored amount will be compared with that allowed by Eq. 27. If the monitored amount also exceeds the allowed amount, the excess traffic will be marked and the length of the time window will be increased by τ. Next, the amount of cumulative traffic within this new time window (starting from when the average was violated) is measured and compared with the average $\rho \cdot 2\tau$. If again the average is violated, the measured amount is again compared with that allowed by Eq. 27. The excess traffic, if any, is decreased by the traffic that is already marked, that will be marked, and so forth. Whenever the monitored average falls below the average, the time window will be reduced to τ time units. We refer interested readers to [15] for more implementation details and a performance study of the fractal leaky bucket.

Envelope Processes for Multifractal Traffic

Norros's fBm model is accurate for connectionless, or "free" traffic, where network resources are unlimited and there is no feedback control mechanism [43]. In the Internet, however, the TCP traffic is dominant and incorporates flow and congestion control. In addition, it has been shown that at the network core, long-term correlations are dominant due to traffic aggregation, while at the network edge, variabilities at small timescales play a major role [16, 77, 78]. The multifractal traffic model is proposed to capture both long-term memory and high variability at small timescales [77, 79–81].

A stochastic process $X(t)$ is multifractal if $E|X(t+\tau) - X(t)|^2 \sim C(t)|\tau|^{2H(t)}$, where $0 < H(t) < 1$ is called the Holder function. The multifractal Brownian motion (mBm) process is a generalization of the fBm process. If in the neighborhood of time t, an mBm can be approximated by an fBm with Hurst parameter $H(t)$, an EP for mBm increments can be derived as upper bounds for such local fBm increments [16]:

$$\hat{A}(t) = \int_0^t \left\{ p + \kappa\sigma H(x) x^{H(x)-1} \right\} dx. \quad (28)$$

When $H(t)$ is a constant, Eq. 28 reduces to the fBm EP discussed earlier. The mBm EP bounds an mBm traffic flow as $\Pr\{A(t) > \hat{A}(\tau)\} = \varepsilon$ with

$$\kappa = \sqrt{-2\log\varepsilon}\ .$$

The analysis for fBm EPs, such as deriving the backlog or delay distributions and the timescale of interest, can be applied to the mBm EPs with appropriate modifications [82, 83].

Service Curves

So far we have discussed EPs, deterministic or probabilistic, that bound the cumulative arrival traffic. Such EPs are also called *arrival curves* in the literature [13]. The current Internet consists of heterogeneous network elements with diverse service capacity and algorithms. The service a flow receives, as the cumulative arrival itself, could also be a complex (or stochastic) process. Naturally, a "dual" approach to bounding cumulative arrivals is to adopt envelope processes that bound the cumulative service a flow receives, which are termed *service curves* in the literature [13, 17, 41, 44, 45, 47–56]. Such service curves can abstract complex service disciplines,[11] and when combined with arrival EPs, can greatly simplify the derivation of performance bounds at various network elements. More importantly, as we will show in this section, service curves are very useful in deriving end-to-end performance measures for QoS provisioning.

Definition

Before we introduce the concept of service curves, we first define a *convolution* operation of nondecreasing, right continuous and *causal* processes (such processes have a zero value for $t < 0$). Given two such processes $A(t)$ and $B(t)$, their convolution is defined as

$$A \otimes B(t) \underline{\underline{def}} \inf_{\tau \in R} \{A(\tau) + B(t-\tau)\}. \quad (29)$$

Recall that in linear systems theory, convolution is defined as $A \otimes B(t) \int_{\tau\in\mathbb{R}} A(\tau) \times B(t-\tau)\, d\tau$. The new definition (Eq. 29) actually replaces the integration with an *infimum* operation, and the multiplication with a *summation*. For this reason, this newly defined operation can be termed as a (min, +) convolution based on *min-plus calculus* [58]. Similarly, a deconvolution operation can be defined as

$$A \oslash B(t) \underline{\underline{def}} \sup_{\tau \in \mathbb{R}} \{A(t+\tau) - B(\tau)\}. \quad (30)$$

A graphical interpretation of Eq. 29 is given in Fig. 9. For a fixed value of τ, the graph of $A(t) + B(t-\tau)$ versus t is obtained by shifting the $B(t)$ curve from the origin to $[\tau, A(t)]$. The convolution is obtained by taking the lower bounding envelope of all such translations.

[11] *The time-varying nature of wireless links also provides a natural application of service curves [84].*

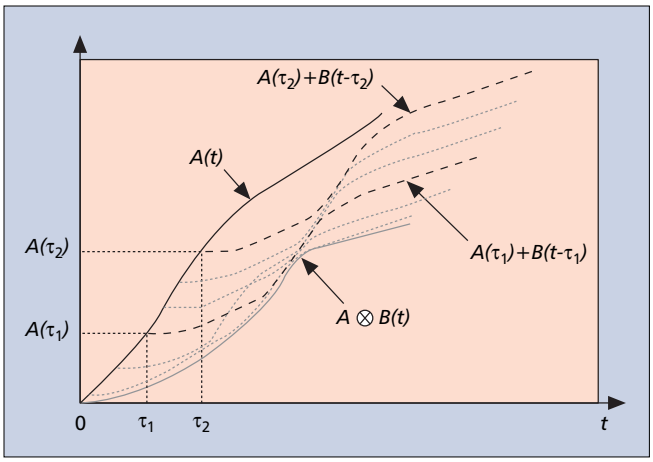

Figure 9. *A graphical interpretation of the convolution operation* A ⊗ B(t).

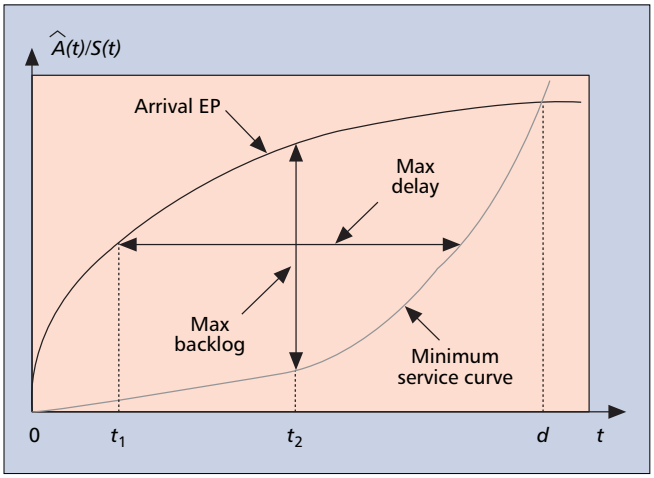

Figure 10. *A graphical interpretation of maximum delay and backlog using the arrival EP and minimum service curve.*

Now we are ready to define the service curve for a network element [45]. Consider a network element serving a cumulative traffic flow of $A(t)$ and generate an output process of $B(t)$. A causal process S is a *minimum service curve* if the departure process satisfies $B(t) \geq A \otimes S(t)$, and a causal process $\bar{S}$ is a *maximum service curve* if $B(t) \leq A \otimes \bar{S}(t)$. Thus $S(t)$ provides a lower bound on the cumulative service the traffic flow receives, while $\bar{S}(t)$ provides an upper bound on the cumulative service the traffic flow receives. This is analogous to linear systems such as a low-pass filter, where the response is the convolution of the input signal and the impulse response $h(t)$ of the system.

Application of Service Envelopes for Deterministic Service Guarantees

Among the two types of service curves, minimum service curves play a larger role in service assurance since, combined with arrival EPs, they can provide upper bounds on the QoS performance measures. Consider a network element with cumulative arrival EP $\hat{A}(d)$ and a minimum service curve $S(t)$. As illustrated in Fig. 10, the maximum delay d_{max} is equal to the maximum horizontal distance between the two curves, the maximum backlog q_{max} is equal to the maximum vertical distance between the two curves, and the time instance when the two curves intersect d provides an upper bound on the system busy period.

This collection of results, including the input/output characterization, is termed deterministic network calculus [13, 18, 19, 21, 50, 57]. We use the notation in [17] and summarize the key results in the following. Consider a network element with a minimum service curve $S(t)$. The cumulative arrival $A(t)$ is bounded by its deterministic EP $\hat{A}(t)$, and the cumulative departure process is $B(t)$. The following hold:

- *Output Envelope*: The function $\underline{B}(t) = \hat{A} \oslash S(t)$ is an envelope for the departure process, in the sense that, for all t, $\tau \geq 0$,

$$B(t + \tau) - B(\tau) \leq \underline{B}(t). \tag{31}$$

- *Backlog Bound*: An upper bound for the backlog, denoted by q_{max}, is given by

$$q_{max} = \hat{A} \oslash S(0). \tag{32}$$

- *Delay Bound*: An upper bound for the delay, denoted by d_{max}, is given by

$$d_{max} = \inf\{d \geq 0 \mid \forall t \geq 0 : \hat{A}(t - d) \leq S(t)\}. \tag{33}$$

Given the notion of service curve, these results can be easily extended to end-to-end performance bounds. For example, consider the guaranteed service adopted by the IETF Intserv working group [5]. The source describes the offered traffic in terms of an arrival EP and requests a lossless service with a fixed upper bound on end-to-end delay. Each of the routers along the path allocates network resources (bandwidth and buffers) to serve this session. Such a service can be modeled as a flow traversing a sequence of service curve elements [45].

Let there be n network elements. Each element i has a minimum service curve $S_i(t)$, a maximum service curve $\bar{S}_i(t)$, is fed by the departure process $B_{i-1}(t)$ from the upstream element, and generates a departure flow of $B_i(t)$. Consider the output of the last network element $B_n(t)$. From the definition of minimum service curves, we have

$$\begin{aligned} B_n(t) &\geq B_{n-1} \otimes S_n(t) \\ &\geq (B_{n-2} \otimes S_{n-1}) \otimes S_n(t) \\ &\geq \ldots \\ &\geq A \otimes (S_1 \otimes S_2 \otimes \ldots \otimes S_n)(t). \end{aligned}$$

Similarly, for the maximum service curves, we can obtain $B_n(t) \leq A \otimes (\bar{S}_1 \otimes \bar{S}_2 \otimes \ldots \otimes \bar{S}_n)(t)$.

Let $S^{net}(t) \stackrel{def}{=} S_1 \otimes S_2 \otimes \ldots \otimes S_n(t)$ and $\bar{S}^{net} \stackrel{def}{=} \bar{S}_1 \otimes \bar{S}_2 \otimes \ldots \otimes \bar{S}_n(t)$. With the use of service curves, not only the service process at a network element can be abstracted, but also the the entire path (or the network cloud) can be modeled with a minimum network service curve $S^{net}(t)$ and a maximum network service curve $\bar{S}^{net}(t)$. End-to-end performance bounds can be easily computed by plugging in the network service curves into Eqs. 31, 32, and 33, as in the single-network-element analysis.

Statistical Network Calculus

As in the arrival EP case, service curves can also be used to bound the service a flow receives in the probabilistic sense. This approach has been explored by several researchers. In [85], Cruz introduces a probabilistic service curve which allows violations according to a certain distribution. A statistical network calculus for the class of "dynamic F-servers" was hinted at in [14], and a family of "statistical service envelopes" is defined in [53] to lower bound the service received by an aggregated flow.

In [17], Burchard, Liebeherr, and Patek define a (minimum) effective service curve, given by

$$\Pr\{B(t) \geq A \otimes S^{\varepsilon}(t)\} \geq 1 - \varepsilon. \tag{34}$$

The effective service curve $S^{\varepsilon}(t)$ bounds the service a single flow receives with high certainty. Using effective service curves, a set of statistical network calculus results are developed in [17]. We use the notation in [17] and summarize the main results in the following. Given an arrival process $A(t)$ conforming to a deterministic EP $\hat{A}(t)$, and given an effective service curve $S^{\varepsilon}(t)$, the following hold:

- *Output Envelope*: The function $\hat{A} \oslash S^{\varepsilon}(t)$ is a probabilistic bound for the departures on $[0, t]$, in the sense that, for all $t, \tau > 0$,

$$\Pr\{B(t, t+\tau) \le \hat{A} \oslash S^{\varepsilon}(\tau)\} \ge 1 - \varepsilon. \quad (35)$$

- *Backlog Bound*: A probabilistic bound for the backlog is given by $q_{max} = \hat{A} \oslash S^{\varepsilon}(0)$, in the sense that, for all $t > 0$,

$$\Pr\{q(t) \le q_{max}\} \ge 1 - \varepsilon. \quad (36)$$

- *Delay Bound*: A probabilistic bound for the delay is given by $d_{max} = \inf\{d \ge 0 \mid \forall t \ge 0: \hat{A}(t-d) \le S^{\varepsilon}(t)\}$, in the sense that, for all $t > 0$,

$$\Pr\{d(t) \le d_{max}\} \ge 1 - \varepsilon. \quad (37)$$

Furthermore, consider a flow $A(t)$ that traverses K network elements in series, each having an effective service curve $S^{k,\varepsilon(t)}$. Then, for any $t \ge 0$,

$$\Pr\left\{B(t) \ge A \otimes \left(S^{1,\varepsilon} \otimes \cdots \otimes S^{K,\varepsilon} \otimes \delta_{(K-1)a}\right)(t)\right\} \ge 1 - \varepsilon\left[1 + (K-1)\frac{t}{a}\right], \quad (38)$$

where $a > 0$ is an arbitrary parameter, and $\delta_{\tau}(t)$ is an impulse function which is ∞ for $t > \tau$ and 0 for $t \le \tau$.

It is worth noting that when $\varepsilon = 0$ (and by letting $a \to 0$ in Eq. 38), these results reduce to the deterministic network calculus discussed in the previous section. These are elegant and important results for end-to-end statistical QoS provision. We refer interested readers to [17] for proofs and other details, and to [46, 51] for the latest advances along this line of work.

Concluding Remarks

In this article we have surveyed various EPs proposed in the literature over the past 15 years, as well as their applications in QoS provisioning. The EPs we have discussed include the class of deterministic EPs for deterministic service assurance, the class of probabilistic EPs for statistical service assurance, the class of EPs for self-similar traffic flows, the class of service curves. We also reviewed the results from deterministic network calculus and statistical network calculus.

A summary and qualitative comparison of the EPs examined in this article are provided in Table 2. These EPs differ in many aspects and their efficiency and accuracy heavily depend on the traffic assumptions. Generally, deterministic EPs are appealing because they are simple in implementation, while probabilistic EPs are more efficient in utilizing network resources. Furthermore, statistical multiplexing can achieve significant improvement in resource utilization, especially when the number of flows is large. Indeed, statistical multiplexing is very useful in a number of contexts, including both wired and wireless networks. Therefore, a promising direction is to use deterministic EPs to regulate flows at the network edge, and statistically multiplex regulated flows within the network for statistical service assurance. Although it is also possible to statistically multiplex flows regulated by probabilistic EPs, the margin for further improvement may be small. As can be seen from Fig. 7, statistical multiplexing of deterministically regulated flows has already achieved a utilization very close to that of average rate allocation.

Currently there is considerable ongoing research on QoS issues in several new networking environments, such as wireless access networks (including 4G wireless networks, wireless mesh networks, mobile ad hoc networks, and wireless sensor networks), MPLS networks, and P2P networks. These networks bring about some interesting problems and unique difficulties for QoS provisioning. For example, a wireless link has a time-varying capacity subject to fading and interference, which is quite different from wired links with a fixed capacity. Wireless transmissions also suffer high loss rates due to transmission errors, which is also quite different from Internet links where buffer overflow is considered as the main cause of loss. Furthermore, user mobility also introduces more frequent topology change in such networks. As another example, the logical topology in a P2P network is quite different from the traditional "edge-core" architecture under which most QoS mechanisms/architectures are developed. How to adapt QoS mechanisms in such environments still remains open, since all these issues need to be addressed in the problem formulation.

Another promising direction for future research is cross-layer design and optimization. Most existing QoS mechanisms are developed under the layered protocol architecture. Although leading to simple independent implementations, this "layered" approach also results in suboptimal application performance. Breaking the barrier among the layers, combined with jointly optimizing the QoS mechanisms/operations across multiple layers (e.g., directly optimizing multimedia replay quality via the resource management, adaptation, control, and protection strategies available at the lower layers of the stack) has the potential of "squeezing the most" out of resource-constrained wireless networks, which are much more unreliable than their wired counterparts. We believe this survey can be useful for research efforts along these directions.

Acknowledgments

The authors are grateful to the editor, Dr. Martin Reisslein, and the seven anonymous reviewers whose comments improved the quality of this article. This work has been supported in part by the National Science Foundation under grants CNS-0520054, CNS-0435303, and CNS-0435228, and by the New York State Office of Science, Technology and Academic Research (NYSTAR) through the Center for Advanced Technology in Telecommunications (CATT) at Polytechnic University.

References

[1] K. Asatani, H. Ueda, and C. M. Lockhart, "Voice over IP and Quality of Service," Guest Editorial, *IEEE Commun. Mag.*, vol. 42, no. 7, July 2004.
[2] R. Chandramouli *et al.*, Guest Editorial, "Recent Advances in Wireless Multimedia," *IEEE JSAC*, vol. 21, no. 10, Dec. 2003.
[3] N. L. S. da Fonseca and P. Shenoy, Guest Editorial, "Proxy Support for Streaming in the Internet," *IEEE Commun. Mag.*, vol. 42, no. 8, Aug. 2004.
[4] S. Blake *et al.*, "An Architecture for Differentiated Services," IETF RFC2457, Dec. 1998.
[5] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: An overview," IETF RFC 1633, July 1994.
[6] J. Soldatos, E. Vayias, and G. Kormentzas, "On the Building Blocks of Quality of Service in Heterogeneous IP Networks," *IEEE Commun. Surveys and Tutorials*, vol. 7, no. 1, 1st Quarter 2005, pp. 70–89.
[7] J. Liebeherr, "Post-Internet QoS Research," *IWQoS 2004 Panel*

	What is bounded	Stat. multiplexing	Det. or stat. service	Considers self-sim.?	Implementation complexity	Utilization performance
{σ, ρ} EP	$A(t)$	No	Deterministic	No	Low	Low
$\{\vec{\sigma}, \vec{\rho}\}$ EPs	$A(t)$	No	Deterministic	No	Low	Low
D-BIND	$A(t)$	No	Deterministic	No	Low	Low
Stat. mux. of regulated flows	$\Sigma_i A_i(t)$	Yes	Statistical	No	High	High
SBB	$A(t)$	No	Statistical	No	Medium	Medium
EBB	$A(t)$	No	Statistical	No	Medium	Medium
Sum of exponentials	$A(t)$	No	Statistical	Yes	Medium	Medium
Chang's bounds	$1/\theta \log Ee^{\theta A(t1, t2)}$	No	Statistical	No	Medium	Medium
Kurose's bounds	$\Sigma_i A_i(t)$	Yes	Statistical	No	High	Medium
H-BIND	$\Sigma_i A_i(t)$	Yes	Statistical	No	High	High
Rate variance EPs	$Var\left\{\frac{A_i(s, s+t_k)}{t_k}\right\}$	Yes	Statistical	No	High	High
Effective envelopes	$\Sigma_i A_i(t)$	Yes	Statistical	No	High	High
WBB	$A(t)$	No	Statistical	Yes	Medium	Medium
fBm EPs	$A(t)$	No	Statistical	Yes	Medium	High
mBm EPs	$A(t)$	No	Statistical	Yes	Medium	High
Det. service curves	$B(t)$	N/A	Det./Stat.	N/A	Low	Low
Prob. service curves	$B(t)$	N/A	Statistical	N/A	High	High

Table 2. *A summary of the EPs discussed in this article.*

Speech, Montreal, Canada, June 2004.
[8] Cisco IOS Documentation, "QC: Cisco IOS Release 12.0, Quality of Service Solutions Configuration Guide," http://www.cisco.com/
[9] Juniper Networks, Inc., http://www.juniper.net/
[10] W. Almesberger, "Traffic Control–Next Generation: Reference Manual," http://linux-ip.net/ gl/tcng/
[11] D. K. Y. Yau and X. Chen, "Resource Management in Software Programmable Router Operating Systems," *IEEE JSAC*, vol. 19, no. 3, Mar. 2001 pp. 488–500.
[12] E. Knightly and N. Shroff, "Admission Control for Statistical QoS: Theory and practice," *IEEE Network*, vol. 13, no. 2, Mar. 1999, pp. 20–29.
[13] K. Kumaran *et al.*, "Novel Techniques for the Design and Control of Generalized Processor Sharing Schedulers for Multiple QoS classes," *Proc. IEEE INFOCOM'00*, Tel-Aviv, Israel, Mar. 2000, pp. 932–41.
[14] C. S. Chang, *Performance Guarantees in Communication Networks*, London: Springer Verlag, 2000.
[15] N. L. S. da Fonseca, G. S. Mayor, and C. A. V. Neto, "On the Equivalent Bandwidth of Self-Similar Sources," *ACM Trans. Modeling and Computer Simulation*, vol. 10, no. 2, Apr. 2000, pp. 104–24.
[16] C. A. V. Melo and N.L.S. da Fonseca, "An Envelope Process for Multifractal Traffic Modeling," Tech. Rep. IC-03-21, Universidade Estadual de Campinas, Nov. 2003.
[17] A. Burchard, J. Liebeherr, and S. D. Patek, "A Calculus for End-to-End Statistical Service Guarantees (2nd revised version)," Tech. Rep. CS-2001-19, University of Virginia, May 2002.
[18] R.L. Cruz, "A Calculus for Network Delay, Part I: Network Elements in Isolation," *IEEE Trans. Info. Theory*, vol. 37, no. 1, Jan. 1991, pp. 114–31.
[19] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *IEEE Trans. Info. Theory*, vol. 37, no. 1, Jan. 1991, pp. 132–41.
[20] E. W. Knightly, "H-BIND: A New Approach to Providing Statistical Performance Guarantees to VBR Traffic," *Proc. IEEE INFOCOM'96*, San Francisco, CA, Mar. 1996, pp. 1091–99.
[21] C.-S. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Trans. Automatic Control*, vol. 39, no. 5, May 1994, pp. 913–31.
[22] D. E. Wrege *et al.*, "Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and practical Trade-Offs," *IEEE/ACM Trans. Net.*, vol. 4, no. 3, June 1996, pp. 352–62.
[23] R. R. Boorstyn *et al.*, "Statistic Service Assurances for Traffic Scheduling Algorithms," *IEEE JSAC*, vol. 18, no. 12, Dec. 2000, pp. 2651–64.
[24] J. Choe and N. B. Shroff, "A Central Limit Theorem based Approach to Analyze Queue Behavior in ATM Networks," *IEEE/ACM Trans. Net.*, vol. 6, no. 5, Oct. 1998, pp. 659–71.
[25] M. Montgomery and G. De Veciana, "On the Relevance of Timescales in Performance Oriented Traffic Characterization,"

Proc. IEEE INFOCOM'96, San Francisco, CA, Apr. 1996, pp. 513–20.
[26] A. Elwalid, D. Mitra, and R. H. Wentworth, "A New Approach for Allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node," *IEEE JSAC*, vol. 13, no. 6, Aug. 1995, pp. 1115–27.
[27] L. Georgiadis *et al.*, "Efficient Network QoS Provisioning based on Per Node Traffic Shaping," *IEEE/ACM Trans. Net.*, vol. 4, no. 4, Aug. 1996, pp. 482–501.
[28] F. Lo Presti *et al.*, "Source Time Scale and Optimal Buffer/Bandwidth Trade-off for 39 Heterogeneous Regulated Traffic in an Network Node," *IEEE/ACM Trans. Net.*, vol. 7, no. 4, Aug. 1999, pp. 490–501.
[29] S. Rajagopal, M. Reisslein, and K.W. Ross, "Packet Multiplexers with Adversarial Regulated Traffic," *Proc. IEEE INFOCOM'98*, San Francisco, CA, Mar./Apr. 1998, pp. 347–55.
[30] M. Reisslein, K. W. Ross, and S. Rajagopal, "Guaranteeing Statistical QoS to Regulated Traffic: The Multiple Node Case," *Proc. IEEE CDC'98*, Tempa, FL, Dec. 1998, pp. 531–38.
[31] M. Reisslein, K. W. Ross, and S. Rajagopal, "A Framework for Guaranteeing statistical QoS," *IEEE/ACM Trans. Net.*, vol. 10, no. 1, Feb. 2002, pp. 27–42.
[32] D. Starobinski and M. Sidi, "Stochastically Bounded Burstiness for Communication Networks," *IEEE Trans. Info. Theory*, vol. 46, no. 1, Jan. 2000, pp. 206–121.
[33] O. Yaron and M. Sidi, "Performance and Stability of Communication Networks via Robust Exponential Bounds," *IEEE/ACM Trans. Net.*, vol. 1, no. 3, June 1993, pp. 372–85.
[34] X. Yu *et al.*, "Queueing Processes in GPS and PGPS with LRD Traffic Inputs," *IEEE/ACM Trans. Net.*, vol. 13, no. 3, June 2005, pp. 676–89.
[35] J. Kurose, "On Computing Per-Session Performance Bounds in High-Speed Multi-hop Computer Networks," *Proc. ACM SIGMETRICS'92*, New Port, RI, June 1992, pp. 128–139.
[36] E. W. Knightly, "Enforceable Quality of Service Guarantees for Bursty Traffic Streams," *Proc. IEEE INFOCOM'98*, San Francisco, CA, Mar./Apr. 1998, pp. 635–42.
[37] E. W. Knightly, "Second Moment Resource Allocation in Multi-Service Networks," *Proc. ACM SIGMETRICS'97*, Seattle, WA, June 1997, pp. 181–91.
[38] J. Beran *et al.*, "Long-Range Dependence in Variable-Bit-Rate video traffic," *IEEE Trans. Commun.*, vol. 43, no. 2, Feb. 1995, pp. 1566–79.
[39] M. Garrett and W. Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," *Proc. ACM SIGCOMM'94*, London, U.K., Aug. 1994, pp. 265–79.
[40] W. E. Leland *et al.*, "On the Self-Similar Nature of Ethernet Traffic," *IEEE/ACM Trans. Net.*, vol. 2, no. 1, Feb. 1994, pp. 1–15.
[41] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Trans. Net.*, vol. 3, no. 3, June 1995, pp. 226–44.
[42] D. Starobinski, Quality of Service in High Speed Networks with Multiple Time-Scale Traffic, Ph.D. dissertation, Technion–Israel Institute of Technology, May 1999.
[43] I. Norros, "On the Use of Fractional Brownian Motion in the Theory of Connectionless Networks," *IEEE JSAC*, vol. 13, no. 6, Aug. 1995, pp. 953–62.
[44] R. Agrawal and R. Rajan, "Performance Bounds for Guaranteed and Adaptive Services," Tech. Rep. RC20649, IBM Research Division, Yorktown Heights, NY, 1996.
[45] R. Agrawal *et al.*, "Performance Bounds for Flow Control Protocols," *IEEE/ACM Trans. Net.*, vol. 7, no. 3, June 1999, pp. 310–23.
[46] F. Ciucu, A. Burchard, and J. Liebeherr, "A Network Service Curve Approach for the Stochastic Analysis of Networks," *Proc. ACM SIGMETRICS'05*, Banff, Alberta, Canada, June 2005, pp. 279–90.
[47] R. L. Cruz, "Quality of Service Guarantees in Virtual Circuit Switched Networks," *IEEE JSAC*, vol. 13, no. 6, Aug. 1995, pp. 1048–56.
[48] P. Goyal, S. S. Lam, and H. M. Vin, "Determining End-to-End Delay Bounds in Heterogeneous Networks," *Multimedia Systems*, vol. 5, no. 3, May 1997, pp. 157–163.
[49] A. Hung and G. Kesidis, "Bandwidth Scheduling for Wide-Area ATM Networks using Virtual Finishing Times," *IEEE/ACM Trans. Net.*, vol. 4, no. 1, Feb. 1996, pp. 49–54.
[50] J.-Y Le Boudec, "Application of Network Calculus to Guaranteed Service Networks," *IEEE Trans. Info. Theory*, vol. 44, no. 3, May 1998, pp. 1087–96.
[51] C. Li, A. Burchard, and J. Liebeher, "A Network Calculus with Effective Bandwidth," Tech. Rep. CS-2003-20, University of Virginia, Nov. 2003.
[52] 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. Net.*, vol. 2, no. 2, Apr. 1994, pp. 137–50.
[53] J.-Y. Qiu and E. W. Knightly, "Inter-Class Resource Sharing using Statistical Service Envelopes," *Proc. IEEE INFOCOM'99*, New York, NY, Apr. 1999, pp. 1404–11.
[54] H. Sariowan, R. L. Cruz, and G. C. Polyzos, "Scheduling for Quality of Service Guarantees Via Service Curves," *Proc. IEEE ICCCN'95*, Las Vegas, NV, Sept. 1995, pp. 512–20.
[55] H. Sariowan, A Service-Curve Approach to Performance Guarantee in Integrated-Service Networks, Ph.D. dissertation, University of California at San Diego, June 1996.
[56] D. Stiliadis and A. Varma, "Latency-Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithms," *IEEE/ACM Trans. Net.*, vol. 6, no. 5, Oct. 1998, pp. 611–24.
[57] C. S. Chang, "On Deterministic Traffic Regulation and Service Guarantees: A Systematic Approach by Filtering," *IEEE Trans. Info. Theory*, vol. 44, no. 3, May 1998, pp. 1097–110.
[58] F. Baccelli *et al.*, *Synchronization and Linearity: An Algebra for Discrete Event Systems*, New York: Wiley, 1992.
[59] R. M. Loynes, "The Stability of a Queue with Non-Independent Inter-Arrival and Service Times," *Proc. Cambridge Philos. Soc.*, vol. 58, no. 3, 1962, pp. 497–520.
[60] J. Liebeherr, D. Wrege, and D. Ferrari, "Exact Admission Control for Networks with Bounded Delay Services," *IEEE/ACM Trans. Net.*, vol. 4, no. 6, Dec. 1996, pp. 885–901.
[61] D. Ferrari and D. C. Verma, "A Scheme for Real-Time Channel Establishment in Wide-Area Networks," *IEEE JSAC*, vol. 8, no. 3, Apr. 1990, pp. 368–79.
[62] L. Georgiadis, R. Guerin, and A. Parekh, "Optimal Multiplexing on a Single Link: Delay and Buffer Requirements," *IEEE Trans. Info. Theory*, vol. 43, no. 6, Sept. 1997, pp. 1518–35.
[63] B. T. Doshi, "Deterministic Rule based Traffic Descriptors for Broadband ISDN: Worst Case Behavior and Connection Acceptance Control," *Proc. ITC'94*, Antibes–Juan-les-Pins, France, June 1994, pp. 591–600.
[64] G. Kesidis and T. Konstantopoulos, "Extremal Traffic and Worst-Case Performance for Queues with Shaped Arrivals," *Proc. Wksp. Analysis Simulation Commun. Networks*, Toronto, Canada, Nov. 1998.
[65] G. Kesidis and T. Konstantopoulos, "Extremal Shape-Controlled Traffic Patterns in High-Speed Networks," *IEEE Trans. Commun.*, vol. 48, no. 5, May 2000, pp. 813–19.
[66] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis based on the Sum of Observations," *Annals Math. Statist.*, vol. 23, 1952, pp. 493–507.
[67] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes,* 4th ed., New York: McGraw-Hill, 2002.
[68] C. Oottamakorn, S. Mao, and S. S. Panwar, "On Generalized Processor Sharing with Regulated Multimedia Traffic Flows," *IEEE Trans. Multimedia*, vol. 8, no. 6, Dec. 2006.
[69] J. Song and R. Boorstyn, "Efficient Loss Estimation in High Speed Networks," *Proc. IEEE ATM Wksp.'98*, Fairfax, VA, May 1998, pp. 360–67.
[70] Z.-L. Zhang, End-to-End Support for Statistical Quality-of-Service Guarantees in Multimedia Networks, Ph.D. dissertation, University of Massachusetts, Amherst, Feb. 1997.
[71] N.B. Shroff and M. Schwartz, "Improved Loss Calculations at an ATM Multiplexer," *IEEE/ACM Trans. Net.*, vol. 6, no. 4, Aug. 1998, pp. 411–21.
[72] S. M. Ross, *Stochastic Processes, 2nd ed.*, New York: Wiley, 1996.
[73] D. Ferrari, "Design and Application of a Delay Jitter Control Scheme for Packet-Switching Internetworks," *Comp. Commun.*, vol. 15, no. 6, July 1992, pp. 367–73.

[74] H. Zhang and E. Knightly, "Providing End-to-End Statistical Performance Guarantees with Bounding Interval Dependent Stochastic Models," *Proc. ACM SIGMETRICS'94*, Nashville, TN, May 1994, pp. 211–20.
[75] J. Qiu and E. W. Knightly, "Measurement-based Admission Control with Aggregate Traffic Envelopes," *IEEE/ACM Trans. Net.*, vol. 9, no. 2, , Apr. 2001 pp. 199–210. [81] R. Riedi and J. Levy-Vehel, "TCP Traffic is Multifractal: A Numerical Study," Tech. Rep. 3129, INRIA Rocquencourt, Mar. 1997.
[76] I. Norros, "A Storage Model with Self-Similar Input," *Queueing Systems*, vol. 16, no. 3–4, 1994, pp. 387–96.
[77] P. Abry *et al.*, "The Multiscale Nature of Network Traffic: Discovery, Analysis and Modeling," *IEEE Sig. Process. Mag.*, vol. 19, no. 3, May 2002, pp. 28–46.
[78] A. Feldmann *et al.*, "Dynamics of IP Traffic: A Study of the Role of Variability and the Impact of Control," *Proc. ACM SIGCOMM'99*, Cambridge, MA, Aug. 1999, pp. 301–13.
[79] A. Erramilli *et al.*, "Performance Impacts of Multi-Scaling in Wide Area TCP/IP Traffic," *Proc. IEEE INFOCOM'00*, Tel-Aviv, Israel, Mar. 2000, pp. 352–59.
[80] A. Erramilli *et al.*, "Multi-Scaling Models of TCP/IP and Sub-Frame VBR Video Traffic," *J. Commun. and Networks*, vol. 4, no. 4, Dec. 2001, pp. 383–95.
[82] C. A. V. Melo and N. L. S. da Fonseca, "Statistical Multiplexing of Multifractal Flows," *Proc. IEEE ICC'04*, Paris, June 2004, pp. 1135–40.
[83] C. A. V. Melo and N. L. S. da Fonseca, "Envelope Process and Computation of the Equivalent Bandwidth of Multifractal Flow," *Computer Networks*, vol. 48, no. 3, June 2005, pp. 351–75.
[84] D. Wu and R. Negi, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, July 2003, pp. 630–43.
[85] R. L. Cruz, "Quality of Service Management in Integrated Service Networks," *Proc. 1st Semi-Annual Research Review*, CWC, UCSD, June 1996.

BIOGRAPHIES

SHIWEN MAO [S'99, M'04] (smao@ieee.org) received B.S. and M.S. degrees from Tsinghua University, Beijing, P.R. China in 1994 and 1997, respectively, both in electrical engineering. He received an M.S. degree in system engineering and a Ph.D. degree in electrical and computer engineering from Polytechnic University, Brooklyn, in 2000 and 2004, respectively. He was a Research Member at the IBM China Research Lab, Beijing from 1997 to 1998. He spent the summer of 2001 as a research intern at the Avaya Labs-Research, Holmdel, NJ. He was a Research Scientist in the Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA from 2004 to 2006. Currently, he is an Assistant Professor in the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL. His research interests include cross-layer design, optimization, and cooperative networking in multihop wireless networks, and multimedia communications in wired and wireless networks. He is the co-author of the textbook, *TCP/IP Essentials: A Lab-Based Approach* (Cambridge University Press, 2004), and a corecipient of the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.

SHIVENDRA S. PANWAR [S'82, M'85, SM'00] (panwar@catt.poly.edu) is a Professor in the Electrical and Computer Engineering Department at Polytechnic University. He received a B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1981, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts, Amherst, in 1983 and 1986, respectively. He joined the Department of Electrical Engineering at the Polytechnic Institute of New York, Brooklyn (now Polytechnic University). He is currently the Director of the New York State Center for Advanced Technology in Telecommunications (CATT). He spent the summer of 1987 as a Visiting Scientist at the IBM T.J. Watson Research Center, Yorktown Heights, NY, and has been a consultant to AT&T Bell Laboratories, Holmdel, NJ. His research interests include the performance analysis and design of networks. Current work includes video systems over peer-to-peer networks, switch performance and wireless networks. He has served as the Secretary of the Technical Affairs Council of the IEEE Communications Society ('92, '93) and is a member of the Technical Committee on Computer Communications. He is a coeditor of two books, *Network Management and Control*, vol. II, and *Multimedia Communications and Video Coding* (published by Plenum in 1994 and 1996, respectively), and coauthored the textbook, *TCP/IP Essentials: A Lab-Based Approach* (Cambridge University Press, 2004). He is a co-recipient of the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.