A Poisson Limit for Buffer Overflow Probabilities

Jin Cao, Kavita Ramanan

Abstract— **A key criterion in the design of high-speed networks is the probability that the buffer content exceeds a given threshold. We consider n independent identical traffic sources modelled as point processes, which are fed into a link with speed proportional to n. Under fairly general assumptions on the input processes we show that the steady state probability of the buffer content exceeding a threshold $b > 0$ tends to the corresponding probability assuming Poisson input processes. We verify the assumptions for a large class of long-range dependent sources commonly used to model data traffic. Our results show that with superposition, significant multiplexing gains can be achieved for even smaller buffers than suggested by previous results, which consider $O(n)$ buffer size. Moreover, simulations show that for realistic values of the exceedance probability and moderate utilisations, convergence to the Poisson limit takes place at reasonable values of the number of sources superposed. This is particularly relevant for high-speed networks in which the cost of high-speed memory is significant.**

Keywords—**Long-range dependence, overflow probability, Poisson limit, heavy tails, point processes, multiplexing.**

I. Introduction

Empirical studies of high-speed data networks (e.g. Ethernet LAN's [1] and wide-area networks [2]) have shown that network traffic displays long-range dependence or self-similarity, thus bringing into question the use of traditional traffic source models such as on-off processes with exponential on-times, which are short-range dependent. Since on-off processes with heavy-tailed on-time distributions have been shown to exhibit long-range dependent behaviour [3], they have been proposed as more realistic data network source models. The implications of this discovery on network performance has been examined by several authors (see [4] for a survey). In particular, in [5] it is shown that when the queue is fed by on-off heavy-tailed sources, under the appropriate scaling the tail of the queue distribution is sub-exponential, rather than exponential as would be predicted by an exponential source model. An important question that arises in this context is whether this discrepancy in queueing behaviour between short-range and long-range dependent traffic models persists even when a large number of sources are multiplexed together. Some authors contend that even high levels of aggregation would not mitigate the burstiness of long-range dependent traffic [1], [2]. On the other hand, others have shown that in the presence of multiplexing some smoothing does indeed take place [6], [7], [8], [9], [10], [11], [12]. For instance in the context of traffic engineering for ATM multiplexers of VBR video sources it is shown in [9] that long-term correlations do not have a significant impact on the cell loss rate when ATM buffers are of realistic dimensions. In [6], [7], [8], [10], [11], [12] a system with n independent identical sources feeding into a link with processing rate $O(n)$ is considered and (under various assumptions on the input processes) a large deviations analysis is used to show that the steady state probability of the buffer content exceeding a threshold equal to nb has an exponential tail (rather than a polynomial one) in the asymptotic limit as n tends to infinity. For the class of semi-Markov modulated fluid arrival processes, e.g. on-off sources with general on-time and off-time distributions, [11] also analyses the asymptotic ($n \to \infty$) decay rate of the steady state probability of the buffer content exceeding nb, in the limit as $b \to 0$, and shows that in this regime the decay rate depends on the on-time and off-time distributions only through their means. Thus the above mentioned papers provide evidence of statistical multiplexing gains when both the threshold and link speed grow proportionally to the number of sources multiplexed.

With the continual increase in link speeds in modern communication networks, the cost of high-speed memory is likely to become a non-trivial factor in the design of data networks. For example, 100 ms of buffering in a 40 Gbs system can be quite expensive, and a typical switch is likely to have many 100 ms buffers supported on different linecards. Therefore it is natural to ask whether the buffer need actually grow proportionally to the link speed in order to realise multiplexing gains, or whether smaller buffers will suffice. In this paper we prove a strong insensitivity result that addresses this question. Our result is stronger than that in [11] and uses a different approach. Firstly, we consider a scaling limit in which although the processing rate is proportional to the number n of i.i.d. sources multiplexed, the buffer size remains $O(1)$. Given fairly general assumptions on the source arrival distribution we show that the exceedance probability tends to the same value as if the source arrival distribution were Poisson with the same mean. This result supports recent statistical analysis of Internet data [13] and indicates that buffers need not scale with the link speed in order to achieve significant multiplexing gains. Secondly, we consider the actual probability rather than just the logarithmic asymptote considered in [6], [7], [8], [11], [12]. Knowledge of the prefactor usually allows one to more accurately assess actual statistical multiplexing gains, and design admission control policies that are neither too conservative nor too aggressive (see, for example, [14], [15]). Furthermore, we model the packet arrival processes as point processes rather than fluid processes. This seems to better capture the behaviour of real packet traffic under multiplexing, especially on the time scales relevant for overflows from $O(1)$ buffers (see Section V for further discussion of this issue).

The paper is organised as follows. In Section II we describe the model and state our main result (Theorem 1). In Section III we state our assumptions and prove the main limit theorem. In Section IV we present some simulation results, and we conclude in Section V with a discussion of the implications of our results for network design.

II. Description of the Model and Main Result

Consider n identical independent sources that send packets to a server that has a deterministic processing rate $n\mu$. We

J. Cao is with the Statistics and Data Mining Research Department, Room 2C-260, Bell Labs, Lucent Technologies, 700 Mountain Avenue, Murray Hill, NJ 07974-2070. E-mail: cao@research.bell-labs.com.

K. Ramanan is with the Mathematics of Networks and Systems Research Department, Room 2C-319, Bell Labs, Lucent Technologies, 700 Mountain Avenue, Murray Hill, NJ 07974-2070. E-mail: kavita@research.bell-labs.com

model the arrival of packets from source i as a point process $A^{(i)}$, where $A_t^{(i)}$ represents the total number of packets emitted by source i in the interval $[0,t]$. We assume that each $A^{(i)}$ has the same distribution as a simple stationary point process A that satisfies $\mathrm{E}[A_t] = \lambda t$ for each $t > 0$ for some $\lambda \in (0,\mu)$. Note that the usual stability condition $\lambda < \mu$ ensures that the queues do not grow unboundedly. The superposed process $A^n \doteq \sum_{i=1}^n A^{(i)}$ represents the aggregate arrival of packets from all n sources. Packets that cannot be processed immediately are stored in a buffer (which is assumed to be infinite). Moreover we assume that each packet has the same size, and without loss of generality set its processing time, assuming it were served at rate μ, to be one. The buffer content, or equivalently the unfinished work in the system, at time t is defined to be the amount of time required to complete the processing of all packets present in the system at time t. The steady state probability of the unfinished work exceeding a certain level b is used as a surrogate for the steady state buffer overflow probability in a buffer of size b. We let U_{A^n} denote the stationary unfinished work in the system when the number of sources is n, the arrival process of each source is distributed according to A and the processing rate is $n\mu$. Also, let N^λ denote the stationary Poisson point process with rate λ, and let U_{N^λ} be the resulting unfinished work process when N^λ is fed into a processor with rate μ. Under the stability condition $\lambda < \mu$, U_{A^n} and U_{N^λ} exist [16] and are explicitly given by the relations

$$U_{A^n} = \sup_{t\in[0,\infty)} \left[A_t^n - n\mu t\right], \quad U_{N^\lambda} = \sup_{t\in[0,\infty)} \left[N_t^\lambda - \mu t\right]. \quad (1)$$

We now state our main result, which holds under general assumptions on the distribution of the arrival process A of each source that are stated precisely in Section III-A. The assumptions are shown to be satisfied by a large class of processes commonly used to model network traffic in Section IV-B.

Theorem 1: Suppose A satisfies Assumption 3 and let N^λ, U_{A^n} and U_{N^λ} be as defined above. Then for any $b > 0$

$$\lim_{n\to\infty} \mathrm{P}(U_{A^n} > b) = \mathrm{P}(U_{N^\lambda} > b). \quad (2)$$

It is well known that the unfinished work U_{N^λ} for a Poisson process has a steady state distribution of the form

$$\mathrm{P}(U_{N^\lambda} > b) = 1 - (1-\rho)e^{\rho b} Q_{\lfloor b\rfloor}(b - \lfloor b\rfloor), \quad (3)$$

where $\lfloor b\rfloor$ is the greatest integer less than or equal to b, $\rho = \lambda/\mu$ is the utilisation of the queue and $Q_n(x), n \in \mathbb{N}$, are polynomials that can be calculated using numerically stable recursion relations (see, for example, [17, p. 391]). Thus Theorem 1 shows that as the number of sources multiplexed increases, the exceedance probability of the unfinished work of sources having fairly general distributions approaches the corresponding probability assuming Poisson sources, which is known explicitly.

III. The Poisson Limit Theorem

In Section III-A we state the main assumption on the source distribution (Assumption 3) and derive some important consequences. In Section III-B we prove the main result, Theorem 1, and in Section III-C we contrast the scaling used in this paper with another more commonly used scaling.

A. Main Assumptions

For $x \in \mathbb{R}$ and $t \in [0,\infty)$, define

$$\Lambda(x,t) \doteq \sup_{\theta\in\mathbb{R}} [\theta x - t^{-1}\log \mathrm{E}[e^{\theta A_t}]]. \quad (4)$$

It is easy to verify (see [18, Lemma 2.2.5]) that

$$\Lambda(x,t) \doteq \sup_{\theta\in[0,\infty)} [\theta x - t^{-1}\log \mathrm{E}[e^{\theta A_t}]], \quad (5)$$

for $x \geq \lambda$. Also, for $x \in \mathbb{R}$, we define the quantities

$$\Lambda_1(x) \doteq \liminf_{t\to 0} \Lambda(x,t), \quad \Lambda_2(x) \doteq \liminf_{t\to\infty} \frac{t\Lambda(x,t)}{\log t}. \quad (6)$$

For any $E \subset \mathbb{R}$, $\mathcal{D}([0,\infty) : E)$ denotes the space of right continuous functions on $[0,\infty)$ with left limits taking values in E, endowed with the Skorokhod J_1 metric $d(\cdot,\cdot)$ (e.g. see [19, p. 73]). Given a point process on $\mathbb{R}$ for conciseness we use A_t to denote $A([0,t])$ and note that $A_t \in \mathcal{D}([0,\infty) : Z_+)$. Then $A(\{t\}) = A_t - \lim_{s\uparrow t} A_s$. Recall the definition of a simple point process [20, Definition 3.3.II].

Definition 2: The point process A is said to be simple if $\mathrm{P}(A(\{t\}) \in \{0,1\} \text{ for all } t \in [0,\infty)) = 1$.

We now state our main assumptions on the process A.

Assumption 3: A is a simple stationary point process satisfying the following three properties.

1. There exists $\lambda < \mu$ such that $\mathrm{E}[A_t] = \lambda t$ for $t \in [0,\infty)$.

2. There exists $\theta_0 > 0$ and $K < \infty$ such that

$$\lim_{t\downarrow 0} t^{-1}\mathrm{E}\left[e^{\theta_0 A_t} 1_{\{A_t > K\}}\right] = 0. \quad (7)$$

3. $\Lambda_2(\mu) > 0$.

The fact that A is simple reflects our modelling assumption that all packets are of the same size, and that multiple packets from a single source do not arrive simultaneously at the server. The latter assumption is quite realistic, while the former is applicable to ATM networks, where packet sizes are all of the same size. (Our framework allows for generalisations to the case of packets with variable sizes, but we do not consider them in this paper.) Along with the fact that A is stationary and has finite intensity λ, the fact that A is simple also implies that the probability of multiple packets arriving within an arbitrary small interval of zero is exceedingly small (see Lemma 4). The second assumption requires that the probability of many packets arriving in an arbitrary small interval $[0,t]$ decays sufficiently fast as the interval size $t \downarrow 0$.

In the next two lemmas we establish some consequences of Assumption 3 that will be used in the proof of the main result.

Lemma 4: Suppose A satisfies Assumption 3(1). Then

$$\mathrm{P}(A_t = 1) = \lambda t + o(t) \text{ and } \mathrm{P}(A_t \geq 2) = o(t). \quad (8)$$

Moreover, if in addition A satisfies Assumption 3(2), then there exists $\theta_0 > 0$ such that uniformly for $\theta \in [0,\theta_0]$

$$\lim_{t\to 0} t^{-1}\log \mathrm{E}\left[e^{\theta A_t}\right] = -\lambda + \lambda e^{\theta}. \quad (9)$$

Proof: The first property follows directly from [20, Propositions 3.3.I, 3.3.IV and 3.3.V]. For the second property, first let

$p_t(k)$ denote $\mathrm{P}(A_t = k)$, $k \in Z_+$, and let $\theta_0 > 0$ and $K < \infty$ be such that (7) is satisfied. Then note that $\mathrm{E}[e^{\theta A_t}] - 1$ is equal to

$$p_t(0) - 1 + p_t(1)e^{\theta} + \sum_{k=2}^{K} p_t(k)e^{k\theta} + \mathrm{E}\left[e^{\theta A_t} 1_{\{A_t > K\}}\right].$$

Using (8) this implies

$$\mathrm{E}[e^{\theta A_t}] - 1 = -\lambda t + \lambda t e^{\theta} + o(t) + \mathrm{E}\left[e^{\theta A_t} 1_{\{A_t \geq K\}}\right].$$

Combining (7) with the last display one obtains uniformly for $\theta \in [0, \theta_0]$

$$\lim_{t \to 0} t^{-1} \left[\mathrm{E}\left[e^{\theta A_t}\right] - 1\right] = -\lambda + \lambda e^{\theta}$$

and

$$\lim_{t \to 0} t^{-1} \left[\mathrm{E}\left[e^{\theta A_t}\right] - 1\right]^2 = 0.$$

The fact that $x - x^2/2 \leq \log(1 + x) \leq x$ for $x \geq 0$ then yields (9). ■

Lemma 5: Suppose A satisfies Assumption 3. Then
1. The functions $\Lambda(\cdot, t)$, $t \in [0, \infty)$, $\Lambda_1(\cdot)$ and $\Lambda_2(\cdot)$ are non-negative non-decreasing functions on $[\lambda, \infty)$;
2. For every $x \geq \mu$, $\Lambda_1(x) > 0$;
3. For any $m \in Z_+$, $m > 1$, $t \in [0, \infty)$ and $x \geq \lambda$, $\Lambda(x, t) \geq m^{-1}\Lambda(x, m^{-1}t)$.

Proof: The first property follows from [18, Lemma 2.2.5]. For the second property, observe that since Assumption 3 is satisfied, from Lemma 4 we know that there exists $\theta_0 > 0$ such that (9) is satisfied uniformly for $\theta \in [0, \theta_0]$. Choose $\theta_* \in (0, \min[\theta_0, \log(\mu/\lambda)])$. For $x \geq \mu$ using (9) and the fact that $\theta x - \lambda(e^{\theta} - 1)$ is strictly increasing on $[0, \theta_*]$, observe that

$$\begin{aligned} \Lambda_1(x) &\geq \liminf_{t \to 0} \left[\theta x - t^{-1} \log \mathrm{E}\left[e^{\theta A_t}\right]\right]_{\theta = \theta_*} \\ &\geq \theta_* x - \lambda(e^{\theta_*} - 1) \; > \; 0. \end{aligned}$$

The third property can be deduced using the stationarity of A and Holder's inequality. ■

Corollary 6: Suppose A satisfies Assumption 3 and let $\alpha(\cdot) \doteq \Lambda(\mu, \cdot)$. There exist $\tau > 1$, δ_1, $\delta_2 > 0$ such that
1) $\alpha(t) \geq \delta_1$ for $t \in [0, \tau]$.
2) $t\alpha(t)/\log t \geq \delta_2$ for $t > \tau$.

Proof: The two properties follow by setting $x = \mu$ in Lemma 5(2), using Assumption 3(3) and Lemma 5(3), along with elementary algebraic manipulations. ■

B. Proof of the Main Result

In this section we prove Theorem 1. The main intuition behind the result is that under the given assumptions the most probable time scale of exceedance of the threshold shrinks with increase in the number n of sources multiplexed (see Theorem 8). The fact that on the decreasing time scale the finite-dimensional distributions of the aggregated arrival processes tend to a Poisson limit as $n \to \infty$ [20, Proposition 9.2.VI] then suggests the result in Theorem 1. The issue of relevant time scales is discussed in more detail in Sections III-C and V.

We introduce the scaled point process B^n defined by

$$B_t^n \doteq A_{t/n}^n \qquad \text{for } t \in [0, \infty). \tag{10}$$

Note that

$$U_{A^n} = \sup_{t \in [0, \infty)} [A_t^n - n\mu t] = \sup_{t \in [0, \infty)} [B_t^n - \mu t]. \tag{11}$$

Lemma 7: For $x \geq \lambda$, $n \in Z_+$ and $t \in (0, \infty)$, we have

$$\mathrm{P}\left(B_t^n > xt\right) \leq e^{-t\Lambda(x, n^{-1}t)}.$$

Moreover for $\alpha(\cdot) = \Lambda(\mu, \cdot)$ and $c \geq 0$ one has

$$\mathrm{P}\left(B_t^n > \mu t + c\right) \leq e^{-t\alpha(n^{-1}t)}.$$

Proof: For $\theta \geq 0$, $\mathrm{P}(B_t^n > xt)$ is no greater than

$$\begin{aligned} & e^{-xt\theta} \mathrm{E}\left[e^{\theta B_t^n}\right] = e^{-xt\theta} \left(\mathrm{E}\left[e^{\theta A_{t/n}}\right]\right)^n \\ = \; & e^{-t\left(x\theta - nt^{-1} \log \mathrm{E}\left[e^{\theta A_{t/n}}\right]\right)}. \end{aligned}$$

Taking the infimum of the last term over $\theta \geq 0$, using (5) and the fact that $x \geq \lambda$ we deduce that

$$\mathrm{P}\left(B_t^n > xt\right) \leq e^{-t\Lambda(x, n^{-1}t)}.$$

The second statement follows from the first by setting $x = \mu$ and noting that $\mathrm{P}\left(B_t^n > \mu t + c\right) \leq \mathrm{P}\left(B_t^n > \mu t\right)$. ■

For any process S, interval $I \subset [0, \infty)$ and $b \in I\!R_+$ for conciseness we introduce the notation

$$\widetilde{\mathrm{P}}_S(I, b) \doteq \mathrm{P}\left(\sup_{t \in I} [S_t - \mu t] > b\right). \tag{12}$$

Theorem 8: Suppose A satisfies Assumption 3. Then for any $b, \epsilon > 0$, there exist $T = T(b, \epsilon) \in I\!R_+$ and $N = N(b, \epsilon) \in Z_+$, such that for any $n > N$

$$\mathrm{P}\left(\sup_{t \in [T, \infty)} [B_t^n - \mu t] > b\right) < \epsilon.$$

Proof: Choose $\gamma \in (0, b)$ so that $s \doteq \gamma\mu^{-1} < 1$. For $l \in Z_+$ let $t_l \doteq sl$, and $I_l \doteq [t_l, t_{l+1})$. Then Lemma 7 and the fact that $b - \gamma > 0$ and B_t^n is non-decreasing in t yields

$$\begin{aligned} \widetilde{\mathrm{P}}_{B^n}([t_l, t_{l+1}), b) &\leq \mathrm{P}(B_{t_{l+1}}^n > b + \mu t_l) \\ = \mathrm{P}\left(B_{t_{l+1}}^n > b - \gamma + \mu t_{l+1}\right) &\leq e^{-t_{l+1}\alpha(n^{-1}t_{l+1})}. \end{aligned}$$

Given $T \in (0, \infty)$ let $L \in Z_+$ be such that $T \in [t_L, t_{L+1})$. From the above display we infer that for some $\tau \in (0, \infty)$ and all $n > L/\tau$, $\widetilde{\mathrm{P}}_{B^n}([T, \infty), b)$ is less than or equal to

$$\begin{aligned} & \widetilde{\mathrm{P}}_{B^n}\left([sL,\ s\lfloor n\tau \rfloor), b\right) + \widetilde{\mathrm{P}}_{B^n}\left([s\lfloor n\tau \rfloor,\ \infty), b\right) \\ \leq \; & \textstyle\sum_{l=L+1}^{\lfloor n\tau \rfloor} e^{-t_l\alpha(n^{-1}t_l)} + \sum_{l=\lfloor n\tau \rfloor + 1}^{\infty} e^{-t_l\alpha(n^{-1}t_l)}. \end{aligned}$$

By Corollary 6 there exist $\tau > 1$, $\delta_1, \delta_2 > 0$ such that $\alpha(t) \geq \delta_1$ for $t \in [0, \tau]$ and $t\alpha(t)/\log t \geq \delta_2$ for $t \in [\tau, \infty)$. Therefore

$$\sum_{l=L+1}^{\lfloor n\tau \rfloor} e^{-t_l\alpha(n^{-1}t_l)} \leq \sum_{l=L+1}^{\lfloor n\tau \rfloor} e^{-t_l\delta_1} < \sum_{l=L+1}^{\infty} e^{-t_l\delta_1},$$

which tends to zero as $L \to \infty$, and similarly

$$\begin{aligned} \textstyle\sum_{l=\lfloor n\tau \rfloor + 1}^{\infty} e^{-t_l\alpha(n^{-1}t_l)} &\leq \textstyle\sum_{l=\lfloor n\tau \rfloor + 1}^{\infty} e^{-n\delta_2 \log(n^{-1}t_l)} \\ = \textstyle\sum_{l=\lfloor n\tau \rfloor + 1}^{\infty} \left(\dfrac{ls}{n}\right)^{-\delta_2 n} &\leq s^{-\delta_2 n} n \textstyle\int_{\tau}^{\infty} x^{-\delta_2 n} dx, \end{aligned}$$

which tends to zero as $n \to \infty$. Thus we can choose $L = L(\gamma, b)$ and $N = N(\gamma, b)$ such that for $T = sL$ and $n > N$ each term in the last two displays is less than $\epsilon/2$. This establishes the lemma with $T \doteq sL$. ■

We now prove the main theorem, Theorem 1.

Proof: From the superposition theorem for point processes we know from [20, Lemmas 9.1.IV and 9.1.X and Proposition 9.2.VI] that B^n converges weakly to N^λ in $\mathcal{D}([0,\infty) : I\!R_+)$. Using the continuous mapping theorem and the fact that the projection operator F_T, that maps f to $\sup_{t\in[0,T]}[f(t) - \mu t]$, is continuous for all T that are not jump points of f [21, Theorems 5.1 and 15.1], we conclude that for a.e. $T \in [0,\infty)$ $F_T(B^n) \Rightarrow F_T(N^\lambda)$. By Portmanteau's theorem [21, Theorem 2.1] for any $b > 0$

$$\lim_{n\to\infty} \widetilde{\mathrm{P}}_{B^n}([0,T], b) = \widetilde{\mathrm{P}}_{N^\lambda}([0,T], b) \tag{13}$$

For $T \in (0,\infty)$ and $b > 0$, definitions (1), (10) and (11) yield

$$0 \le \mathrm{P}(U_{A^n} > b) - \widetilde{\mathrm{P}}_{B^n}([0,T], b) \le \widetilde{\mathrm{P}}_{B^n}([T,\infty), b),$$

and likewise

$$0 \le \mathrm{P}(U_{N^\lambda} > b) - \widetilde{\mathrm{P}}_{N^\lambda}([0,T], b) \le \widetilde{\mathrm{P}}_{N^\lambda}([T,\infty), b).$$

Theorem 8, along with the observation that N_t^λ has the same distribution as $\sum_{i=1}^n N_{t/n}^\lambda$ for any $n \in I\!N$, shows that given any $\varepsilon > 0$ for all large enough T and n the last terms in the above two displays are less than $\varepsilon/2$. Since $\varepsilon > 0$ is arbitrary, using (13) we conclude that for all $b > 0$

$$\lim_{n\to\infty} |\mathrm{P}(U_{A^n} > b) - \mathrm{P}(U_{N^\lambda} > b)| \le \lim_{\varepsilon\downarrow 0} \varepsilon = 0.$$

■

C. Comparison of Different Scalings

Recall the definition (10) of the scaled arrival processes B^n and consider the sequence of exceedance probabilities associated with a sequence of thresholds b_n, $n \in I\!N$. In this paper we have considered the case when $b_n = b$, $n \in I\!N$. Another more commonly used scaling (see [6], [10], [11]) is $b_n = nb$ for some $b > 0$. In this section we compare the mathematical analysis of the limiting probabilities in these two cases. The practical implications of these scalings is discussed in Section V. Using the abbreviation (12) note that $\widetilde{\mathrm{P}}_{B^n}([0,\infty), b_n)$ is clearly bounded by $\widetilde{\mathrm{P}}_{B^n}([0,T], b_n) + \widetilde{\mathrm{P}}_{B^n}([T,\infty), b_n)$. Under the scaling $b_n = b$, the first term is the asymptotically dominant term since, as shown in Theorem 8, the second term is arbitrarily small for sufficiently large T and n. On the other hand, we show below that under the scaling $b_n = nb$ the second term is asymptotically dominant. For n large enough such that $nT^{-1}b > \lambda$, from Lemma 7 we obtain

$$\widetilde{\mathrm{P}}_{B^n}([0,T], nb) \le \mathrm{P}(B_T^n > nb) \le e^{-T\Lambda(nT^{-1}b, n^{-1}T)},$$

and hence

$$\limsup_{n\to\infty} n^{-1} \log \widetilde{\mathrm{P}}_{B^n}([0,T], nb) \le -\liminf_{t\to 0}[t\Lambda(bt^{-1}, t)]. \tag{14}$$

Substituting the definition (4) of $\Lambda(\cdot, t)$ we have

$$\liminf_{t\to 0} t\Lambda(bt^{-1}, t) \ \ge\ \sup_{\theta\in I\!R} \liminf_{t\to 0} \left[\theta b - \log \mathrm{E}[e^{\theta A_t}]\right] \tag{15}$$

Suppose that in addition to Assumption 3 the process also satisfies $\mathrm{E}[e^{\theta A_t}] < \infty$ for any $\theta > 0$ and $t \in [0, t_0]$. Then the fact that A_t is right continuous with $A_0 = 0$ and Lebesgue's bounded convergence theorem imply $\limsup_{t\to 0} \log \mathrm{E}[e^{\theta A_t}] = 0$. Together with (14) and (15), this shows that

$$\limsup_{n\to\infty} n^{-1} \log \widetilde{\mathrm{P}}_{B^n}([0,T], nb) \le -\sup_{\theta\in I\!R}[\theta b] = -\infty. \tag{16}$$

However, previous results (e.g. [8], [7], [12], [11, Theorem 3.2]) show that under the above assumptions

$$\lim_{n\to\infty} n^{-1} \log \widetilde{\mathrm{P}}_{B^n}([0,\infty), nb) = -I(b) < \infty, \tag{17}$$

where $I(b) \doteq \inf_{t>0}[t\Lambda(\mu + bt^{-1})]$ is the exponential decay rate of the exceedance probability. In contrast to the case $b_n = b$, here the asymptotic decay rate is a function of the arrival process distribution. Also, (16) and (17) imply

$$\lim_{n\to\infty} n^{-1} \log \mathrm{P}\left(\sup_{t\in[T,\infty)} [B_t^n - \mu t] > nb\right) = -I(b).$$

Thus $\widetilde{\mathrm{P}}_{B^n}([T,\infty), nb)$ is the dominant contribution to the exceedance probability under the scaling $b_n = nb$.

Our comparison of different scalings highlights the importance of a *critical time scale* [9] for performance analysis, which is loosely defined as the time scale beyond which the source behaviour does not significantly affect the exceedance probability. For the scaling $b_n = b$ considered in this paper, the critical time scale is inversely proportional to the number n of sources superposed. This follows from Theorem 8 and the fact that $\sup_{t\in[T,\infty)}[B_t^n - \mu t] = \sup_{t\in[T/n,\infty)}[A_t^n - n\mu t]$. Thus as n increases the characteristics of an individual source at the critical time scale become less important, the effect of statistical multiplexing becomes the dominant factor, and the net behaviour tends towards Poisson. However, with the scaling $b_n = nb$, the time scale of interest does not change with superposition. In this case individual source behaviour still plays an important role.

IV. Study of Convergence Rates via Simulation

We simulate the unfinished work process for a variety of representative input source models in order to gauge the qualitative dependence of the rate of convergence on parameters such as the threshold level b and utilisation ρ of the link, as well as on the source characteristics.

A. Models of the Arrival Processes

We consider five classes of arrival processes that are commonly used as models of network traffic sources. The different classes exhibit key features of observed data network traffic behaviour such as long-range dependence, asymptotic self-similarity [1] and "burstiness" to different extents. Recall that a stationary stochastic process $X(t)$ is said to be long-range dependent if there exists a constant $H \in (1/2, 1)$, referred to as the *Hurst parameter*, such that $\mathrm{Var}(X(t)) \sim O(t^{2H})$, while it is said to be self-similar if the finite-dimensional distributions of the time-changed and rescaled process $\theta^{-H}X(\theta t)$ are the same as the original process, i.e. $\theta^{-H}X(\theta\cdot) \stackrel{fidi}{=} X(\cdot)$ for $\theta > 0$. A commonly used measure for burstiness of traffic is the index of dispersion of counts (IDC) [22], [23]. Of special interest is the limiting value I_∞ of IDC associated with an arrival process A_t, which is defined by $I_\infty \doteq \lim_{t\to\infty} \mathrm{Var}[A_t]/\mathrm{E}[A_t]$. It is well known that $I_\infty = 1$ for a Poisson process and for a stationary renewal processes it is equal to the square of the coefficient of variation (COV) of the inter-arrival distribution. A process is considered more (less) bursty than Poisson if I_∞ is greater (smaller) than 1. I_∞ also has the property that it is invariant under the superposition of independent identical sources.

We now describe the five classes in detail. Classes 1), 2) and 5) are stationary renewal processes, and hence are completely characterised by their inter-arrival time distribution Z. For Class 1) Z is **Weibull** with parameter α (i.e. Z^α is exponential). The inter-arrival distribution of HTTP packet traffic has been empirically observed to be well modelled by the Weibull distribution [24], which is heavy-tailed for $\alpha \in (0,1)$. For Class 2) Z is **inverse Weibull** with parameter α (i.e. $Z^{-\alpha}$ is exponential). In this case Z has an asymptotic *Pareto* tail distribution with tail parameter α, i.e. $\mathrm{P}(Z > z) = O(z^{-\alpha})$ as $z \to \infty$. For Class 5 Z has a **Gamma** (or Erlang) distribution with parameter $\alpha \in I\!N$, (i.e. the sum of α i.i.d. exponentials). If $\alpha \in (1,2)$, the sources in Classes 2 and 5 are long-range dependent with Hurst parameter $(3-\alpha)/2$ and belong to the class of fractal point process [25], which have $I_\infty = \infty$.

Classes 3) and 4) belong to the category of what we refer to as Poisson on-off point processes, each of which consists of an underlying on-off process and a finite conditional mean arrival rate κ. Packets arrive as a Poisson point process with rate κ during an on-period, while no packets arrive during an off-period. The **ExponExpoff** Class 3 sources have exponentially distributed on and off periods. This model was used in [14] to study the queueing behaviour of superposed ATM traffic. Class 4 are **ParetoonExpoff** sources, which have an exponential off-period distribution and on-periods that are *Pareto* with tail parameter α.

B. Verification of Assumption 3 for Traffic Models

We now verify Assumption 3 for the five classes of models described in the last section.

Lemma 9: Any stationary renewal process A whose inter-arrival distribution Z satisfies $\mathrm{E}[Z] = \lambda^{-1} > 0$ has $\mathrm{E}[A_t] = \lambda t$ and satisfies Assumption 3(3). Moreover, if $\mathrm{P}(Z \le t) = O(t^\beta)$ for some $\beta > 0$, then Assumption 3(2) is also satisfied. In particular, the arrival processes from Classes 1), 2) and 5) satisfy Assumption 3.

Proof: The fact that $\mathrm{E}[Z] = \lambda^{-1} > 0$ implies automatically that for every $t \in [0,\infty)$, $\mathrm{E}[A_t] = \lambda t$.

Let L' be the time to the first arrival, and suppose L_n is the time interval between the nth and $(n+1)$th arrivals. Since A_t is a stationary renewal process, $\{L', L_n, n \in I\!N\}$ are independent and each L_n has the same distribution as Z. Thus for any $\phi > 0$,

$$\mathrm{P}(A_t \ge n) \le \mathrm{P}\left(\sum_{i=1}^{n-1} L_i \le t\right) \le e^{\phi t}\left(\mathrm{E}[e^{-\phi Z}]\right)^{n-1}. \quad (18)$$

Since $\mathrm{E}[Z] = \lambda^{-1} > 0$, Z is positive on a set of positive probability, and hence $\mathrm{E}[e^{-\phi Z}] \in (0,1)$. So given any $\phi > 0$ there exists $\tilde\theta > 0$ such that $c \doteq e^{\tilde\theta}\mathrm{E}[e^{-\phi Z}] < 1$. Hence for $\theta \le \tilde\theta$ and $t \in [0,\infty)$ using (18) we obtain

$$\mathrm{E}[e^{\theta A_t}] \;\le\; \sum_{n=0}^{\infty} \mathrm{P}(A_t \ge n) e^{\tilde\theta n} \;\le\; \frac{e^{\phi t}\sum_{n=0}^\infty c^n}{\mathrm{E}[e^{-\phi Z}]} < \infty. \quad (19)$$

Now we use methods similar to those used in [10, Proposition 3.1] to prove Assumption 3(3). Pick $\beta > 0$ and for $t \in (0,\infty)$ define $\theta_t \doteq \beta \log t/t$. Then note that

$$\Lambda_2(\mu) \ge t\theta_t\mu/\log t - \limsup_{t\to\infty} \log \mathrm{E}[e^{\theta_t A_t}]/\log t, \quad (20)$$

which is equal to $\beta\mu - \limsup_{t\to\infty}\log\mathrm{E}[e^{\theta_t A_t}]/\log t$. Since $\theta_t \downarrow 0$ as $t \uparrow \infty$, by (19) $\mathrm{E}[e^{\theta_t A_t}]$ is finite for all sufficiently large t. Choose $C \in (\lambda,\mu)$, recall that $\lceil Ct\rceil$ is the least integer greater than or equal to Ct and note that

$$\begin{aligned}\mathrm{E}[e^{\theta_t A_t}] &= \textstyle\sum_{n=0}^\infty \mathrm{P}(A_t = n)e^{\theta_t n}\\ &\le t^{C\beta} + \textstyle\sum_{n=\lceil Ct\rceil}^\infty \mathrm{P}(A_t \ge n)e^{\theta_t n}.\end{aligned} \quad (21)$$

For $\phi \in [0,\infty)$ define $g(\phi) \doteq e^{\phi C^{-1}}\mathrm{E}[e^{-\phi Z}]$, and note that g is differentiable with derivative $g'(\phi) = e^{\phi C^{-1}}\left(C^{-1}\mathrm{E}[e^{-\phi Z}] - \mathrm{E}[Ze^{-\phi Z}]\right)$. Therefore $g(0) = 1$ and $g'(0) = C^{-1} - \lambda^{-1} < 0$, and so there exists $\phi_0 > 0$ such that $g(\phi_0) < 1$. To bound the summation on the right side of (21), now let $d \doteq \mathrm{E}[e^{-\phi_0 Z}]$ and $d_t \doteq e^{\theta_t}\mathrm{E}[e^{-\phi_0 Z}]$, and observe that $d_t \to d$ as $t \uparrow \infty$, and $d < 1$. Using this observation along with (18) and the fact that $\theta_t = \beta\log t/t$, we conclude that for all sufficiently large t,

$$\begin{aligned}&\textstyle\sum_{n=\lceil Ct\rceil}^\infty \mathrm{P}(A_t \ge n)e^{\theta_t n}\\ &\quad\le e^{\phi_0 t}\textstyle\sum_{n=\lceil Ct\rceil}^\infty (\mathrm{E}[e^{-\phi_0 Z}])^{n-1}e^{\theta_t n}\\ &\quad= e^{\phi_0 t}d^{-1}(1-d_t)^{-1}d_t^{\lceil Ct\rceil}\\ &\quad= d^{-1}(1-d_t)^{-1}g(\phi_0)^{\lceil Ct\rceil}e^{\phi_0(t - C^{-1}\lceil Ct\rceil)}e^{\lceil Ct\rceil\beta t^{-1}\log t}\\ &\quad\le d^{-1}(1-d_t)^{-1}t^{C\beta}e^{\beta t^{-1}\log t}.\end{aligned}$$

Substituting this into (21) we infer that the quantity $\limsup_{t\to\infty}\log\mathrm{E}[e^{\theta_t A_t}]/\log t$ is less than or equal to

$$\limsup_{t\to\infty} \frac{C\beta\log t + \log(1 + d^{-1}(1-d_t)^{-1}e^{\beta t^{-1}\log t})}{\log t},$$

which is equal to $C\beta$. Together with (20) this establishes the required inequality $\Lambda_2(\mu) \ge (\mu - C)\beta > 0$.

Finally, if $\mathrm{P}(Z \le t) \le Mt^\beta$ for some $M < \infty$, $\beta > 0$, then elementary calculations lead to the inequality $\mathrm{P}(A_t = k) \le (Mt^\beta)^{k-1}$. Thus given any $\theta > 0$ and $K \ge 2/\beta + 1$ for $t \in [0, (e^{-\theta}/M)^{1/\beta}]$,

$$\mathrm{E}\left[e^{\theta A_t}1_{\{A_t \ge K\}}\right] \le \sum_{k=K}^\infty e^{k\theta}(Mt^\beta)^{k-1} \le \frac{e^{K\theta}M^{K-1}t^2}{1 - e^\theta Mt^\beta},$$

which shows that Assumption 3(2) is satisfied. The last statement follows because the Weibull distribution satisfies $\mathrm{P}(Z \le t) = O(t^\beta)$ with $\beta = \alpha$ and the inverse Weibull and Gamma distributions satisfy it with $\beta = 1$. ■

TABLE I

DESCRIPTION OF FIVE SOURCE ARRIVAL MODELS IN THE SIMULATION (MEAN ARRIVAL RATE = 5.45).

Class	Name	Description and Parameters	I_∞
1	Weibull	inter-arrival time $\sim$ Weibull($\alpha = 0.5$)	5
2	InvWeibull	inter-arrival time $\sim$ Inverse-Weibull($\alpha = 1.5$)	∞
3	ExponExpoff	on period $\sim \mathrm{Exp}(1)$, off period $\sim \mathrm{Exp}(10)$, rate $\kappa = 60$	100.2
4	ParetoonExpoff	on period $\sim$ Pareto($\alpha = 1.5$), off period $\sim \mathrm{Exp}(10)$, rate $\kappa = 60$	∞
5	Gamma	inter-arrival time $\sim$ Gamma($\alpha = 5$)	0.2

Lemma 10: Any Poisson on-off process with a finite conditional mean rate κ satisfies Assumptions 3(2). Moreover, if the on-time and off-time distributions T_{on} and T_{off} satisfy $\mathrm{E}(T_{on}^{1+\zeta}) < \infty$ for some $\zeta > 0$, and $\mathrm{E}(T_{off}) < \infty$, then Assumption 3(3) is also satisfied. In particular, Classes 3) and 4) satisfy Assumption 3.

Proof: Let Y be the underlying on-off process associated with the Poisson on-off process, so that $Y_t \in \{0,1\}$ for $t \in [0,\infty)$ and $I_t \doteq \int_0^t Y_t \, dt$ represents the cumulative on-time in the interval $[0,t]$. Note that for every $t \in [0,\infty)$ $I_t \le t$ and $A_t \mid I_t$ is a Poisson random variable with mean κI_t. Thus for any $t \in [0,\infty)$ and $\theta \in I\!R$

$$\mathrm{E}\left[e^{\theta A_t} 1_{\{A_t \ge 2\}} \mid I_t\right] = e^{(e^\theta - 1)\kappa I_t} - (e^\theta - 1)\kappa I_t - 1.$$

Since $I_t \le t$ and $e^x - x - 1$ is increasing for $x > 0$, this implies

$$\mathrm{E}\left[e^{\theta A_t} 1_{\{A_t \ge 2\}} \mid I_t\right] \le e^{(e^\theta - 1)\kappa t} - (e^\theta - 1)\kappa t - 1,$$

which is $o(t)$ as $t \downarrow 0$. Taking the expectation of both sides of the above display with respect to I_t and observing that the right hand side is independent of I_t, yields Assumption 3(2). Note that a similar argument could be used to show that $\mathrm{E}[e^{\theta A_t}] < \infty$ for $t \in [0,\infty)$, $\theta \in I\!R$.

To prove Assumption 3(3), first note that for any $\varepsilon > 0$, there exists $\gamma_\varepsilon > 0$ such that $e^\theta < 1 + (1+\varepsilon)\theta$ for $\theta \in (0, \gamma_\varepsilon)$. Also note that if $\theta_t \doteq \beta \log t / t$ for some $\beta > 0$, $\theta_t \to 0$ as $t \to \infty$. Thus for all t sufficiently large so that $\theta_t \in (0, \gamma_\varepsilon)$, by the above inequality

$$\begin{array}{rcccl} \mathrm{E}[e^{\theta_t A_t}] & = & \mathrm{E}[e^{(e^{\theta_t} - 1)\kappa I_t}] & \le & \mathrm{E}[e^{(1+\varepsilon)\kappa\theta_t I_t}], \\ & \le & e^{\varepsilon\kappa\theta_t t}\mathrm{E}[e^{\kappa\theta_t I_t}] & = & t^{\varepsilon\kappa\beta}\mathrm{E}[e^{\kappa\theta_t I_t}]. \end{array}$$

Therefore

$$\begin{array}{rcl} \Lambda_2(\mu) & \ge & \beta\mu - \limsup_{t\to\infty} \log \mathrm{E}[e^{\theta_t A_t}] / \log t \\ & \ge & \beta\mu - \limsup_{t\to\infty} \log \mathrm{E}[e^{\theta_t \kappa I_t}] / \log t - \varepsilon\kappa\beta. \end{array}$$

Sending $\varepsilon \downarrow 0$ we conclude that

$$\Lambda_2(\mu) \ge \beta\mu - \limsup_{t\to\infty} \frac{\log \mathrm{E}[e^{\theta_t \kappa I_t}]}{\log t}. \tag{22}$$

However from the assumptions on the on-time and off-time distributions, the proof of Proposition 3.3 of [26] shows that for any $C \in (\lambda, \mu)$ there exist positive K and α such that $\mathrm{P}(I_t > Ct) \le Kt^{-\alpha}$. Then by Proposition 3.1 of [10], it follows that there exists $\beta > 0$ such that the right hand side of (22) is strictly greater than 0. This establishes Assumption 3(3). The last statement holds since the exponential and Pareto distributions satisfy the assumptions on the on-time and off-time distributions. ∎

C. Simulation Results

We simulate representative examples from each of the five classes described in Section IV-A with specific parameter values as summarised in Table I. All arrival processes have the same mean packet arrival rate $\lambda = 5.45$ and equal packet sizes. In order to provide an intuitive feel of the relative burstiness of the the five arrival point processes that were simulated, in Figure 1 we have used Trellis graphics [27] to illustrate a sample realisation of each example in the time period [0, 91.5], which contains approximately 500 packets. The illustration suggests that the Poisson arrival process (top panel) is more regular than Classes 1-4 (bottom 4 panels), which appear to have more clustered arrivals.

Figure 2 plots the buffer exceedance probability (in the $\log_{10}$ scale) against buffer size $b = 0, \ldots, 100$ for processes from the five models (see Table I), and compares them with the unfinished work distribution U_{N^λ} for the Poisson process. We consider three different degrees of superposition, namely $n = 1, 100, 1000$, and two utilisation levels $\rho = 0.3$ and $\rho = 0.7$. (Note that the processing time per packet is given by ρ/λ.) Each panel has six curves – the five shaded lines represent the five models, and the thicker black line represents the Poisson arrival process. When $n = 1$, as expected the steady state unfinished work distributions corresponding to the different arrival processes are very different. The two Poisson on-off models show a greater exceedance probability in comparison with the others, and all curves with the exception of the **Gamma** process have an exceedance probability larger than that of Poisson. This is not surprising since Classes 1-4 are burstier than Poisson. However, as n increases, the difference between these curves decreases ($n = 100$) and eventually all the curves converge to that of Poisson ($n = 1000$). At lower utilisations, convergence seems to be achieved at lower values of n. For example, when $\rho = 0.3$ almost complete convergence is achieved for all classes for $n = 100$, while for $\rho = 0.7$ it takes $n = 1000$ for convergence to occur. Notice that although the commonly used burstiness measure I_∞ of the arrival processes does not change, the queueing behaviour changes drastically with superposition. Also notice that even though I_∞ is infinite for both **InvWeibull** and **ParetoonExpoff**, at $n = 100$ they both show smaller buffer exceedance probabilities (at least for buffer sizes between 0 and 100) than **ExponExpoff**, which has a finite I_∞. In fact, the buffer exceedance probability for the **InvWeibull** is quite close to Poisson even for $n = 1$, particularly for low utilisations. This indicates that the limiting IDC I_∞ is a rather poor metric for measuring the impact of the burstiness of superposed processes on queueing behaviour for small buffers.

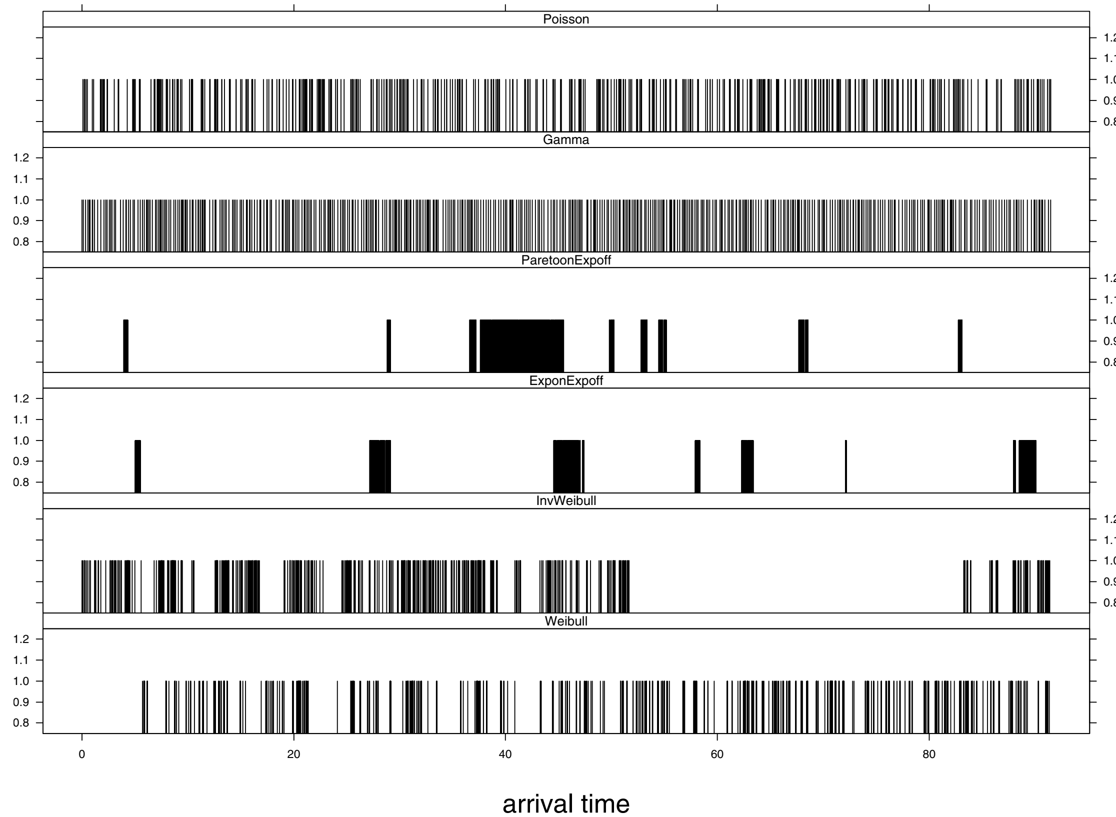

Fig. 1. A realisation of the five arrival processes in the simulation.

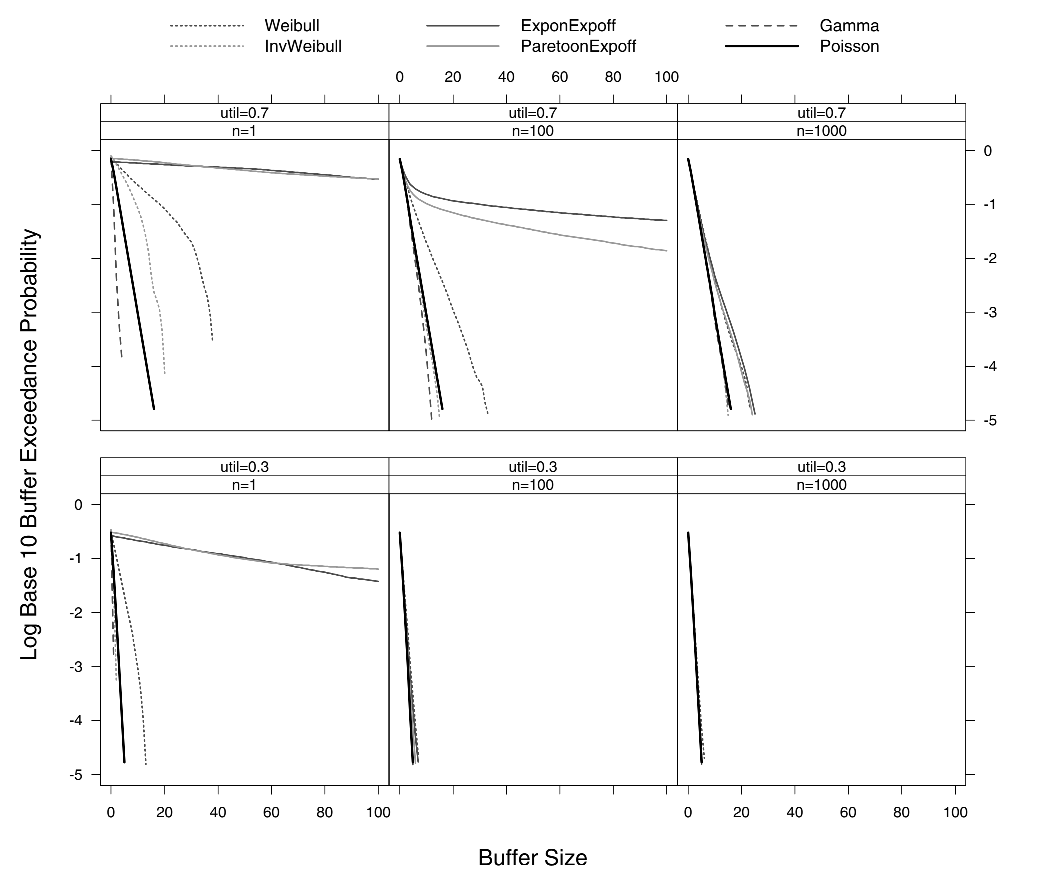

Fig. 2. Buffer exceedance probability against buffer sizes, for superposed processes from the five models when $n = 1, 100, 1000$.

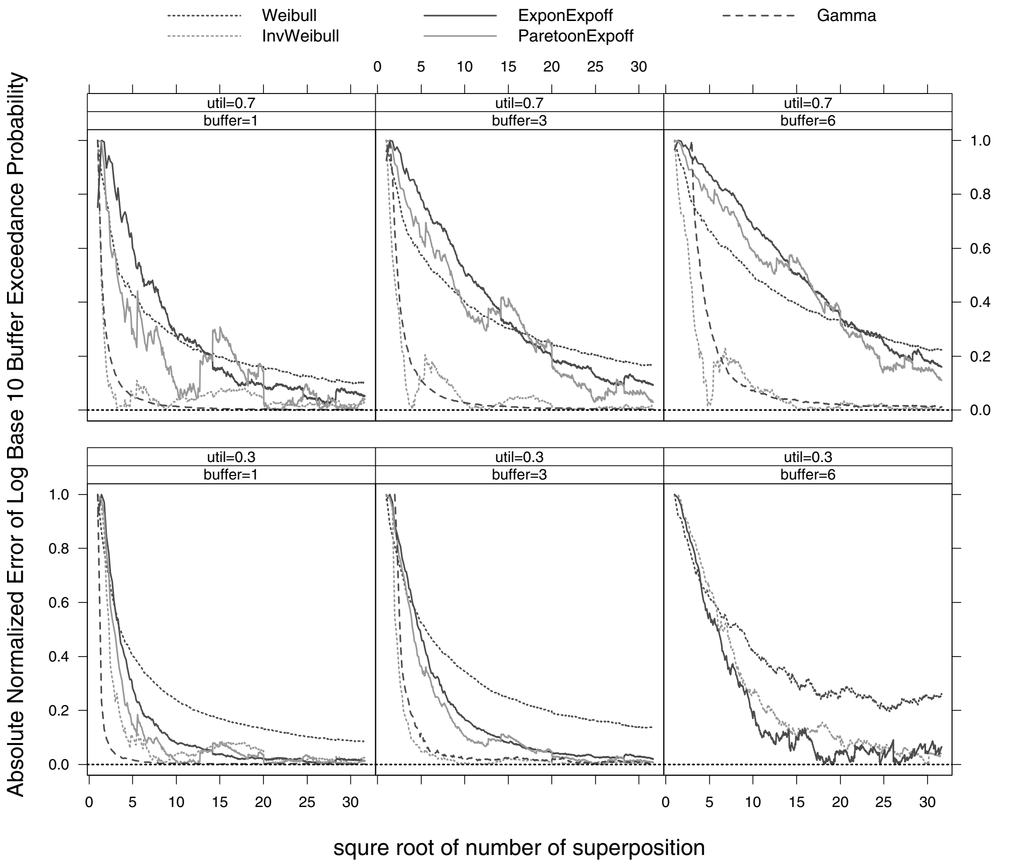

Fig. 3. Absolute normalised error of buffer exceedance probability against different superpositions, for buffer size $b = 1, 3, 6$ and for superposed processes from the fi ve models.

For each of the five representative examples, Figure 3 plots the absolute normalised difference of the logarithm of the exceedance probability from that of the Poisson process against the square root of the number of superpositions n, once again for two utilisation levels (0.3 and 0.7). Let $p_{A^n}(b)$ be the probability of exceeding a threshold size b for the superposed arrival process A^n, and let $p_{N^\lambda}(b)$ be the corresponding probability for the Poisson process. Then the normalised error is defined as

$$\frac{|\log_{10} p_{A^n}(b) - \log_{10} p_{N^\lambda}(b)|}{\sup_{n \in I\!N} |\log_{10} p_{A^n}(b) - \log_{10} p_{N^\lambda}(b)|}. \quad (23)$$

In the simulation the supremum in the denominator, i.e. the maximum of the absolute error is estimated using the observed values from $1 \leq n \leq 1000$. This maximum is mostly achieved at small values of n. The three panels represent three different buffer sizes, namely $b = 1, 3, 6$, with the five lines representing the five models. Notice that there are no curves for **InvWeibull** and **Gamma** at utilization 0.3 and buffer size 6. This is because the exceedance probability from these arrival processes are very small (on the order of 10^{-6}) and our simulation does not give enough accuracy for small probabilities. This makes the normalising quantity in (23) unstable. The normalising quantity in (23) is larger for larger buffer sizes, and is smaller for higher utilisations. Overall, convergence to the Poisson limit for large buffers and high utilisation seems to take place at higher values of n (see both Figures 2 and 3). Between different classes of arrival processes the convergence seems fastest for **InvWeibull** and **Gamma** and slowest for **Weibull**. The slow convergence for Weibull may be related to the propensity of the Weibull distribution to create short inter-arrivals for $\alpha < 1$. Indeed it can be shown that as t goes to zero, the **Weibull** arrival process with shape parameter α satisfies $\mathrm{P}(A_t \geq 2) = O(t^{1+\alpha})$. However, for the other models $\mathrm{P}(A_t \geq 2) = O(t^2)$.

V. Implications for Network Design

In this paper we have proved a strong insensitivity result for the unfinished work of superposed point processes, in the limit as the number of sources, and proportionally the processing rate, goes to infinity. Specifically, we show that the actual steady state probability (and not just the logarithmic asymptote) of the unfinished work exceeding a *fixed* positive threshold tends to the corresponding probability assuming the arrival processes were Poisson. Simulations suggest that at moderate utilisations, for a large class of traffic models convergence is achieved at reasonable levels of superposition, say when several hundreds of sources are multiplexed.

Our result supports recent statistical analysis of Internet traffic [13], [24]. In [13], [24], Internet traces collected at a corporate research site and several universities are studied. These sites typically have a few hundred to a few thousand simultaneous streams during a busy period. The empirical results in [13], [24] suggest that with increase in the number of superpositions the packet arrival process tends toward Poisson and packet sizes tend toward independence. An open-loop queueing study was also carried out in [24] using traces from the corporate research site with different degrees of superposition. It was shown that when the utilisation is kept fixed, the effect of the long-range dependence of network traffic diminishes and the steady state exceedance probabilities (assuming an infinite buffer) tend to the Poisson limit. This is to be contrasted with the study in [28], where a single Ethernet LAN trace is used to demonstrate the impact of long-range dependence on queueing performance, with almost no reference to the degree of superposition. In short, our results suggest that with increase in the amount of superposition, the effect of long-range dependence on queueing reduces even more markedly than suggested by previous results.

As discussed in Section III-C, our results differ from previous ones with respect to the type of scaling considered. In our analysis the buffer size is $O(1)$, so that the time scale of interest is inversely proportional to the number n of sources superposed (see Theorem 8), and hence proportional to the mean packet inter-arrival time. When n increases, this inter-arrival time decreases, while the packet inter-arrivals from an individual source remains the same. Thus on that time scale the characteristics of an individual source become less important, the packet inter-arrivals tend to look more Poisson-like, and the steady state exceedance probability also tends to that of Poisson. Our asymptotic analysis is valid for a whole range of magnitudes of the exceedance probability, and not just for small probabilities, as is often the case in large deviations analyses. Previous analyses [6], [7], [12], [8], [10], [11] consider a different limiting regime, where the buffer size is $O(n)$, in which case the time scale of interest does not change with superposition. Those results show that the logarithm of the exceedance probability decreases exponentially with n, and the decay rate depends on the distribution of the arrival process. Moreover, the nature of their analysis implies that the results are valid only in a regime where exceedance probabilities are small. (Note that [10] also provides estimates of the prefactor.)

The different scalings have important practical implications for network design. Both our analysis as well as that of [6], [7], [12], [8], [10], [11] show that statistical multiplexing gains obtained from superposing many sources are significant even when individual sources have long-range dependent behaviour. This is particularly applicable to traffic in the core of a network, where a high level of aggregation takes place. However the results of [6], [7], [12], [8], [10], [11] only show that multiplexing gain (i.e. an exponential decay rate) can be achieved when the buffer size increases proportionally to the number of sources multiplexed. Our results suggest that for high speed systems (assuming that the characteristics of the individual sources do not change) relatively small buffers may be adequate in order to achieve the same loss rate. Specifically, this implies that a system design with a linear increase in the buffer size may be overly conservative. As mentioned in the introduction, this is particularly relevant in cases where the cost of high-speed memory is high.

Another important observation is that since the critical time scale decreases with superposition for our scaling, a fluid approximation of the processes may no longer be valid to study $O(1)$ buffer overflows. We model our packet arrival processes as point processes because this perspective seems to reflect more accurately the discrete nature of packet arrivals in routers and switches. Some authors have considered models in which time is discretised and the arrival process is represented in terms of the accumulated workload over each discrete interval. This approach appears to be too crude to study $O(1)$ buffer overflows, especially if the size of the discretised interval is kept fixed with increase in the degree of superposition since in that case the accumulated workload tends to infinity.

It is well known that the superposition of n suitably scaled i.i.d. simple stationary point processes on the positive real line converges weakly to a Poisson process as $n \uparrow \infty$ [20]. Thus the main result of this paper can be viewed as a continuity result for the mapping that takes the arrival processes into the steady state buffer content process, i.e. the reflection mapping on the nonnegative real line. This is the viewpoint taken in [29], where the analysis here is generalised to the case of marked point processes to model the more realistic case when packets are of different sizes. It is also shown in [29] that the packet size process tends towards independence with increase in the number of sources multiplexed. In future work, one would like to gain more insight into the dependence of the convergence rates on various parameters of the system, in order to develop a useful heuristic for sizing buffers in a network to smooth burstiness effects. In this paper we have used the probability of exceeding a level b as a proxy for the overflow probability in a finite buffer of size b. It would be worthwhile to see how accurate this approximation is in practice. Finally, here we have considered only the steady state unfinished workload distribution. There are other quantities of interest such as the waiting time and queue length processes that may also be of interest.

ACKNOWLEDGMENTS

We would like to thank B. Cleveland, D. Sun and D. Lin for the empirical work in [13] that supports this result.

REFERENCES

[1] Will Leland, Murad Taqqu, Walter Willinger, and Daniel Wilson, "On the Self-Similar Nature of Ethernet Traffic," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, 1994.

[2] Vern Paxson and Sally Floyd, "Wide-Area Traffic: The Failure of Poisson Modeling," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 226–244, 1995.

[3] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 71–86, 1997.

[4] O. J. Boxma and V. Dumas, "Fluid queues with long-tailed activity period distributions," *Computer Communications*, vol. 21, pp. 1509–1529, 1998.

[5] T. Konstantopoulos and S-J. Lin, "High variability vs long-range dependence for network performance," in *Proc. 35th. IEEE CDC*, Kobe, Dec. 1996, pp. 1354–1359.

[6] D. D. Botvich and N. G. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing Systems*, pp. 293–320, 1995.

[7] C. Courcoubetis and R. Weber, "Buffer overflow asymptotics for a buffer handling many traffic sources," *Journal of Applied Probability*, vol. 33, pp. 886–903, 1996.

[8] N. G. Duffield, "Economies of scale in queues with sources having power-law large deviations scalings," *Journal of Applied Probability*, vol. 33, pp. 840–857, 1996.

[9] B. K. Ryu and A. Elwalid, "The Importance of Long-Range Dependence of VBR Traffic in ATM Traffic Engineering: Myths and Realities," in *Proc. ACM SIGCOMM '96*, 1996, pp. 3–14.

[10] N. Likhanov and R. Mazumdar, "Cell loss asymptotics for buffers fed with a large number of independent stationary sources," *Journal of Applied Probability*, vol. 36, no. 1, pp. 86–96, 1999.

[11] Michel Mandjes and Jeong Han Kim, "Large deviations for small buffers: an insensitivity result," *Queuing Systems*, 2001, to appear.
[12] A. Simonian and J. Guibert, "Large deviations approximation for fluid queues fed by a large number of on/off sources," *IEEE Journal of Selected Area in Communications*, vol. 13, pp. 1017–1027, 1995.
[13] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "The Effect of Statistical Multiplexing on the Long Range Dependence of Internet Packet Traffic," Tech. Rep., Bell Labs, Murray Hill, NJ, 2001.
[14] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt, "Squeezing the Most out of ATM," *IEEE transactions on communications*, vol. 44, no. 2, 1996.
[15] A. Elwalid, D. Heyman, T.V.Lakshman, D. Mitra, and A. Weiss, "Fundamental Bounds and Approximations to ATM Multiplexers with Applications to Video Teleconferencing," *IEEE J. Sel. Areas in Comm.*, vol. 13, no. 6, pp. 1004–1016, 1995.
[16] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Camb. Philos. Soc.*, pp. 58:497–520, 1962.
[17] U. Mocci J. Roberts and J. Virtamo, *Broadband Network Teletraffic, Final Report of Action COST 242*, Springer, Berlin, 1996.
[18] A. Dembo and O. Zeitouni., *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston, 1992.
[19] W. Whitt, "Some useful functions for functional limit theorems," *Mathematics of Operations Research*, vol. 5, no. 1, pp. 67–85, 1980.
[20] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 1988.
[21] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.
[22] D. R. Cox and P. A. W. Lewis, *The Statistics Analysis of Series of Events*, Methuen, London, England, 1966.
[23] Kotikalapudi Sriram and Ward Whitt, "Charactering superposition arrival processes in packet multiplexers for voice and data," *IEEE Journal on Selected Areas in Communications*, vol. SAC-4, pp. 833–846, 1986.
[24] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "On the Nonstationarity of Internet Traffic," *Proc. ACM SIGMETRICS*, 2001.
[25] B. K. Ryu and S. B. Lowen, "Point process approaches to the modeling and analysis of self-similar traffic: Part I–Model constrction," in *Proc. IEEE Infocom '96*, 1996, pp. 1468–1475.
[26] Michel Mandjes and Sem Borst, "Overflow behavior in queues with many long-tailed inputs," *Advances in Applied Probability*, vol. 32, pp. 1150–1167, 2001.
[27] W. S. Cleveland and M. Fuentes, "Trellis Display: Modeling Data from Designed Experiments," Tech. Rep., Bell Labs, 1997, black and white postscript, http://cm.bell-labs.com/stat/doc/trellis.doe.bw.ps. color postscript, http://cm.bell-labs.com/stat/doc/trellis.tour.col.ps.
[28] A. O. Erramilli, O. Narayan, and W. Willinger, "Experimental Queueing Analysis with Long-Range Dependent Packet Traffic," *IEEE/ACM Transactions on Networking*, vol. 4, pp. 209–223, 1996.
[29] J. Cao and K. Ramanan, "A Poisson Limit for Queueing Functionals of Superposed Marked Point Processes," *in preparation*, 2002.