THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 6, July–August 1983
Printed in U.S.A.

Asymptotic Analysis of a Queueing Model With Bursty Traffic

By D. Y. BURMAN* and D. R. SMITH*

(Manuscript received September 24, 1982)

Assuming a particular model for "bursty" traffic at a packet-switching node, we find expressions for the expected delay of packets that are valid in light and heavy traffic. Each expression consists of a "correction factor" multiplied by the expected delay experienced by packets when the arrivals are "smooth" (Poisson) and of the same average rate. Approximate values for the correction factor in arbitrary traffic can be obtained by interpolation. This provides an example of a method that often gives fast approximate solutions for bursty traffic models that are not themselves tractable but become so when the offered traffic is assumed to be Poisson.

I. INTRODUCTION

Many models of queueing systems assume that arrivals occur according to a Poisson process. Intuitively, the Poisson process may be characterized by the properties that events occur one at a time and do not depend on the past history of events. Typically, this situation arises when there are large numbers of users of a system, as in the case of arrivals of calls to a central office, since one arrival does not significantly affect the probability of another. Fortunately, these models are often mathematically tractable.

For other systems the Poisson assumptions are not realistic. Often arrivals are indicative of overall activity and give information about the probability of future arrivals. For example, suppose that all arrivals

* Bell Laboratories, Holmdel, N.J.

are generated by a single user who is alternately active and inactive. An arrival indicates that the user is active and hence that there is a greater than average probability of another arrival shortly thereafter. A second example concerns the arrival of packets to a node in a packet-switching network. When virtual calls are employed, the route packets travel for a particular call is fixed for the duration of the call. If we make the assumption that an individual virtual call generates packets according to a Poisson process, then the instantaneous arrival intensity at any node is equal to the sum of the intensities for the calls routed through the node, which varies probabilistically with time. In these examples the arrivals are correlated and the traffic is said to be bursty.

Even the simplest models involving bursty traffic tend to be difficult to solve analytically. Several authors have given approximations. Heffes[1] matched the first three moments of the arrival process to those of an Interrupted Poisson Process. Assuming that an arriving packet requires an exponentially distributed amount of time to be served, he was then able to use the results of Kuczura[2] to analyze this system. Laue,[3] making a similar approximation for the arrival process, assumed that an arriving packet requires a constant service time. The mean waiting time of a packet could then be calculated using the numerical matrix techniques developed by Neuts[4] and Lucantoni and Neuts.[5] A third approach developed by Anick, Mitra, and Sondhi[6] treats a different but related model. The models of Heffes and Laue are based on the assumption that an individual customer generates packets according to a Poisson process. Anick, Mitra, and Sondhi assume that a customer generates packets at a constant rate for a random time. The resulting fluid model was treated numerically by calculating the eigenvalues of the resulting equations.

All of the above works derive numerical techniques to estimate the traffic statistics of interest over a wide range of traffic parameters. It is the goal of this paper to provide simple, closed-form expressions that give insight into the effect of burstiness on delays. This is done by studying queueing systems offered bursty traffic (see Section II for a complete description) in light and heavy traffic.

In light (heavy) traffic, it is obvious that the expected delay tends to zero (infinity). Surprisingly, when the expected delay is divided by the expected delay for the same system offered Poisson traffic of equal average intensity, the ratio goes to a nonzero finite limit in both light and heavy traffic. These limits may be thought of as "correction factors" by which the expected delay for the solvable Poisson system should be multiplied to obtain the expected delay for the bursty system. By interpolating between the light- and heavy-traffic results, one can obtain insight into the approximate effect of burstiness for all values

of traffic. Indeed, similar light- and heavy-traffic limits for the $M/E_k/c$ system have been used to obtain very accurate approximate values of the delay for all values of traffic (see Lauber and Smith[7]).

The remainder of this paper is organized as follows: The queueing model for bursty traffic and our results are discussed in detail in Section II. Rigorous proofs of the light-traffic results are given in Section III when the service times are of the phase type. (Appendix A presents a brief background on phase-type distributions.) Since any service-time distribution can be approximated arbitrarily closely by one of phase type, it is sufficient for practical purposes to establish the light-traffic results for the latter. Appendix B presents an intuitive approach for deriving the light-traffic results. Heavy-traffic results are presented in Section IV and concluding remarks in Section V.

II. QUEUEING MODEL AND RESULTS

The specific queueing model treated here is one in which the arrival process is a nonhomogeneous Poisson process whose rate equals λm, where m is the state of an $M/M/\infty$ queue with birth rate α and service rate β. The arrivals are offered to a single server whose successive services are assumed to be independent and identically distributed according to some general distribution with mean equal to μ^{-1}. Blocked arrivals queue up and are served on a first-in-first-out basis.

This queueing model supports either of two (essentially identical) scenarios for the queueing of packets of information at a packet switch. In both scenarios, the switch is modeled as a single server with an infinite buffer for queued packets, and the service time of a packet is its length (in bits) divided by the line speed. In the first scenario, an individual virtual call generates packets according to a Poisson process with rate λ, for an exponentially distributed length of time with mean β^{-1}. The distribution of requests for virtual calls is Poisson with rate α, and the number of simultaneous virtual calls that can be supported by the switch is unlimited. The second scenario is similar to a fluid model treated by Mitra and Anick[8] and Kosten.[9] In this case an individual customer is in one of two states, either "active" or "inactive." It is assumed that the time in each of the states is exponentially distributed with rates β and γ, respectively. While in the "active" state, the customer transmits packets according to a Poisson process with rate λ. If there are N (large) such customers with $N\gamma \sim \alpha$ and $\gamma\beta^{-1} \sim 0$, then the number of "active" customers is distributed like the number of customers in the $M/M/\infty$ queue described above.

Throughout the remainder of this paper, we will refer to the entities queued at the single server as packets and the entities in the $M/M/\infty$ system as virtual calls or calls. Our analysis will focus on the limiting form for the mean delay of a packet in a lightly or heavily loaded

system. The total number of packets generated during a virtual call is geometrically distributed with mean $\lambda\beta^{-1}$ and the mean total rate of packet generation is $\lambda(\alpha/\beta)$.

There is a technical problem in the analysis for this model since the packet arrival rate is unbounded. Hence, consider first the system where the number of virtual calls is limited to N, i.e., the rate of arrivals of packets is λ times the number of calls present in an M/M/N/N queueing system. The average rate of arrival of packets to the queue is

$$\lambda_N \stackrel{\text{def}}{=} \lambda(\alpha/\beta)[1 - B(N, \alpha/\beta)], \tag{1}$$

where $B(N, \alpha/\beta)$ is the Erlang Blocking formula. Let $D_B^{(N)}$ be the expected delay of a packet in this system, and let $D_M^{(N)}$ be the delay in an M/G/1 with arrival rate λ_N and the same service-time distribution.

Our key light-traffic result (valid for phase-type service distributions) is

$$\lim_{\lambda \to 0} \frac{D_B^{(N)}}{D_M^{(N)}} = k_l^{(N)}, \tag{2}$$

where $0 < k_l^{(N)} < \infty$, and

$$\lim_{N \to \infty} k_l^{(N)} = 1 + \frac{2\mu}{\alpha} \frac{[1 - \phi(\beta)]}{1 + C^2} \stackrel{\text{def}}{=} k_l, \tag{3}$$

where $\mu^{-1} = E(S)$, $C^2 = \text{var}(S)/E(S)^2$ (S is a service-time random variable), and $\phi(\cdot)$ is the Laplace transform of the stationary excess of S. The limit (2) is proved by using a simple extension of a lemma established in Burman and Smith.[10] The exact value of $k_l^{(N)}$ is difficult to compute; however, the limit k_l as $N \to \infty$ is computable and can be interpreted as the light-traffic limit of the ratio of the delay (D_B) for the bursty system described earlier (with no limit on the number of virtual calls) and the delay (D_M) for the M/G/1 queue with arrival rate $\lambda(\alpha/\beta)$. This statement can be summarized (although not explicitly proved) as

$$\lim_{\lambda \to 0} \frac{D_B}{D_M} = k_l, \tag{4}$$

where k_l is given in (3).

In the heavy-traffic case no rigorous limit is available. Nevertheless, diffusion analysis gives the following approximation in heavy traffic:

$$\frac{D_B^{(N)}}{D_M^{(N)}} \approx k_h^{(N)}, \tag{5}$$

where $k_h^{(N)}$ is known explicitly [see eq. (40)] and

$$\lim_{N\to\infty} k_h^{(N)} = 1 + \frac{2\mu}{\alpha}\frac{1}{1+C^2} \stackrel{\text{def}}{=} k_h. \tag{6}$$

This supports the conjecture that

$$\lim_{\lambda \to \frac{\beta\mu}{\alpha}} \frac{D_B}{D_M} = k_h. \tag{7}$$

It should be noted that $1 < k_l < k_h$, and while no proof is available, it is reasonable to conjecture that $k_l < D_B/D_M < k_h$ for all stable values of λ. (Indeed it is conjectured that D_B/D_M is monotone, as suggested by a similar analysis for the delay in an M/G/c queue normalized by the delay in an M/M/c queue (see Lauber and Smith[7]).

At this point it is worth noting that the manner in which the traffic intensity $\rho = (\lambda/\mu)(\alpha/\beta)$ approached 0 or 1 affects the limiting value of D_B/D_M. The previously described results are based on the variation of λ only; one can also allow α, the rate of arrival of calls, to vary. For an intuitive example of the difference, note that $\lambda \to 0$ corresponds to light traffic with one packet per call, while $\alpha \to 0$ corresponds to light traffic with a geometrically distributed (with mean λ/β) number of packets per call. It can be shown (although it is not explicitly reported here) that D_B/D_M is completely different in the two cases.

We now focus our attention on the behavior of k_l and k_h for a fixed mean number of calls in the system (α/β). As $\alpha, \beta \to \infty$, $(\alpha/\beta$ fixed) the arrival process of packets approaches a Poisson process and indeed, by examination of (3) and (6), k_l and k_h both go to 1. As $\alpha, \beta \to 0$ (α/β fixed), k_h goes to ∞, and k_l goes $1 + \beta/\alpha$. To understand these limits, note that in this case the number of calls remains constant for longer and longer periods of time and steady-state effects become significant. In heavy traffic, the process remains in states for which the packet generation rate is faster than the service rate, so that the delays become large. In light traffic, one may obtain the $1 + \beta/\alpha$ limit in an intuitive fashion by conditioning on the number of calls m at packet generation times and computing the conditional delay (using known expressions for delay) assuming a constant arrival rate λm.

In (4) and (7), the delay for the bursty system was normalized by the delay for the M/G/1 system. If instead we choose to normalize by the delay for an M/M/1 system with the same arrival and service rates (denoted by $\tilde{D}_m$), then the limits become

$$\lim_{\lambda\to 0} \frac{D_B}{\tilde{D}_M} = 1 + \frac{C^2-1}{2} + \frac{\mu}{\alpha}[1 - \phi(\beta)] \tag{8}$$

and

$$\lim_{\lambda \to \frac{\beta\mu}{\alpha}} \frac{D_B}{\tilde{D}_M} = 1 + \frac{C^2 - 1}{2} + \frac{\mu}{\alpha}. \tag{9}$$

These results are interesting in that they suggest separation of the effects of variability of the service time $(C^2 - 1)/2$ and the variability of the arrival process (the third term).

In addition to the results for the mean delay D_B, we show that the light-traffic limit of the distribution of the delay D, given $D > 0$, is

$$\lim_{\lambda \to 0} P(D > t \mid D > 0) = \frac{\int_t^\infty \int_y^\infty [\alpha + \beta e^{-\beta(x-t)}] dH(x) dy}{\frac{\alpha}{\mu} + 1 - \int_0^\infty e^{-\beta x} dH(x)}. \tag{10}$$

Again, in order for this to be rigorously stated it should be in terms of the limit of similar quantities for systems allowing only a finite number of virtual calls.

III. DERIVATION OF THE LIGHT-TRAFFIC RESULTS

In this section, we derive the light-traffic result stated in (4). Our approach is to show that as the traffic intensity goes to 0, the probability of having i (greater than 0) packets in the system goes to 0 asymptotically as λ^i. The exact rate of convergence can be derived by a detailed study of the state equations and from there (4) follows trivially.

Consider a single-server queue whose service times are of phase type (see Appendix A). Let the arrival process be a nonhomogeneous Poisson process whose rate is λ times a function of the state of a Markov process. Then, the multidimensional process consisting of the number i of packets in queue, the state j of the arrival Markov process, and the phase k of the packet in service is itself a Markov process. A typical state will be denoted by (i, j, k) for $i > 0$ and $(0, j)$ for $i = 0$, where $j \geq 0$ and $k = 1, \cdots, m$, the number of phases. The ergodic distribution will be denoted by $\rho(\cdot, \cdot, \cdot)$ and define

$$\rho(i) = \sum_{j,k} \rho(i, j, k),$$

with the obvious definition for $i = 0$. When there is an upper bound on the arrival rate (as when the arrival Markov process is finite), then the technique used in Burman and Smith[10] to prove Theorem 3.1 therein can be employed to prove:

Lemma 1: There exists a constant $R > 0$ such that

$$\rho(i) \leq \lambda^i R^i.$$

One may define

$$\tilde{\rho}(i, j, k) = \lim_{\lambda \to 0} \lambda^{-1} \rho(i, j, k) \tag{11}$$

with an analogous definition for $i = 0$, and these limits may be recursively related and shown to exist by examination of the balance equations. (See Smith[11] and Burman and Smith[10] for examples of this technique.) Thus, when the arrival rates are bounded, the light-traffic methodology is straightforward. It is quite possible, however, that the resulting equations are difficult to solve.

This is exactly the case when the arrival Markov process is an M/M/N/N queue (finite number of virtual calls). It is difficult to explicitly solve for $\tilde{\rho}^{(N)}$ [the limiting light-traffic normalized probabilities for this system, see (11)], although it can be shown that $\tilde{\rho}^{(N)} \to \tilde{\rho}$ as $N \to \infty$, where $\tilde{\rho}$ is the solution to the equations (assuming Lemma 1) when the birth-death process is the M/M/∞ queue. At the core of this argument (not presented here in detail) are the facts that the equations involving $\tilde{\rho}^{(N)}(\cdot, j, \cdot)$ for $j < N$ are identical with those involving $\tilde{\rho}^{(N+1)}(\cdot, j, \cdot)$ and that

$$\lim_{N \to \infty} \tilde{\rho}^{(N)}(0, j) = \frac{1}{j!} (\alpha/\beta)^j e^{-\alpha/\beta}.$$

The existence of the limits [in (11)] can be shown by recursion on i and the fact that the limit was previously established for $i = 0$. Thus $\lim_{N \to \infty} \tilde{\rho}^{(N)} (= \tilde{\rho})$ may be calculated by studying the system with $N = \infty$.

We now turn our attention to calculating $\tilde{\rho}$ by studying the bursty system with arrival rate λ times the number of calls in an M/M/∞ queue. For this system it is not hard to show that ρ, the steady-state probability satisfies

$$\begin{aligned} &-(j\lambda + \alpha + j\beta)\rho(0, j) + \alpha\rho(0, j-1) \\ &+ (j+1)\beta\rho(0, j+1) + \sum_n \rho(1, j, n)\mathrm{E}_n = 0, \end{aligned} \tag{12}$$

$$\begin{aligned} &-(j\lambda + \alpha + j\beta - T_{kk})\rho(1, j, k) + \alpha\rho(1, j-1, k) \\ &+ (j+1)\beta\rho(1, j+1, k) + \sum_n \rho(1, j, n)T_{nk} \\ &+ j\lambda\omega_k\rho(0, j) + \omega_k \sum_n \rho(2, j, n)\mathrm{E}_n = 0, \end{aligned} \tag{13}$$

and

$$\begin{aligned} &-(j\lambda + \alpha + j\beta - T_{kk})\rho(i, j, k) + \alpha\rho(i, j-1, k) \\ &+ (j+1)\beta\rho(i, j+1, k) + \sum_n \rho(i, j, n)T_{nk} \\ &+ j\lambda\rho(i-1, j, k) + \omega_k \sum_n \rho(i+1, j, n)\mathrm{E}_n = 0, \quad \text{for} \quad i \geq 2, \end{aligned} \tag{14}$$

where $j \geq 0$ and $\omega\cdot$, $T\cdot\cdot$, and $\mathrm{E}\cdot$, are the initial, transition, and exit rates defining the phase-type distribution of the service-time process. (These are discussed in greater detail in Appendix A.) We also assume that Lemma 1 holds.

Next define the generating functions

$$q(i, z, k) = \sum_j z^j \tilde{\rho}(i, j, k), \qquad i > 0,$$

and

$$q(0, z) = \sum_j z^j \tilde{\rho}(0, j).$$

From (12) to (14), we see that q satisfies the following equations

$$-\alpha(1 - z)q(0, z) + \beta(1 - z)q_z(0, z) = 0, \tag{15}$$

$$q(1, z, \cdot)[T - \alpha(1 - z)I] + q_z(1, z, \cdot)\beta(1 - z)I = -\omega z q_z(0, z), \tag{16}$$

and

$$q(i, z, \cdot)[T - \alpha(1 - z)I] + q_z(i, z, \cdot)\beta(1 - z)I = -z q_z(i - 1, z, \cdot) \quad \text{for} \quad i > 1. \tag{17}$$

Equation (15) immediately gives

$$q(0, z) = e^{-\alpha/\beta(1-z)}. \tag{18}$$

The next lemma relates D_B, the expected delay in this system, to $q(1, z, \cdot)$.

Lemma 2:

$$\lim_{\lambda \to 0} \lambda^{-1} D_B = -\frac{\beta}{\alpha} q_z(1, 1, \cdot) T^{-1} e,$$

where e is the vector of ones.

Proof: The mean number L of packets in the queue is given by

$$L = \sum_{i>1} (i - 1) \sum_{j,k} \rho(i, j, k).$$

By Lemma 1,

$$\lim_{\lambda \to 0} \lambda^{-2} L = \sum_{j,k} \tilde{\rho}(2, j, k) = q(2, 1, \cdot)e.$$

From (17), we get that

$$q(2, 1, \cdot) = -q_z(1, 1, \cdot)T^{-1}$$

and by Little's Law we are done. □

We now establish (3). The previous lemma shows that the key

quantity is $q_z(1, 1, \cdot)$. Differentiating (16) with respect to z and evaluating at $z = 1$ gives

$$\alpha q(1, 1, \cdot)I + q_z(1, 1, \cdot)(T - \beta I) = \omega \frac{\alpha}{\beta}\left(\frac{\alpha}{\beta} + 1\right), \tag{19}$$

where we used (18) to give us $q(0, z)$. Substituting for $z = 1$ into (16) gives that

$$q(1, 1, \cdot) = \frac{\alpha}{\beta}(-\omega T^{-1}) = \frac{\alpha}{\beta\mu}\xi,$$

where ξ is the stationary distribution of the service-time process [see (42) and (43)]. Rearranging (19) we get

$$\begin{aligned} q_z(1, 1, \cdot) &= \frac{\alpha^2}{\beta}\omega T^{-1}(\beta I - T)^{-1} + \frac{\alpha}{\beta}\left(\frac{\alpha}{\beta} + 1\right)\omega(\beta I - T)^{-1} \\ &= -\frac{\alpha}{\beta^2}[\alpha\beta\omega T^{-1}(\beta I - T)^{-1} - (\alpha + \beta)\omega(\beta I - T)^{-1}] \\ &= -\frac{\alpha}{\beta^2}[\alpha\omega T^{-1} - \beta\omega(\beta I - T)^{-1}] \\ &= \left(\frac{\alpha}{\beta}\right)^2(-\omega T^{-1}) + \left(\frac{\alpha}{\beta}\right)\omega(\beta I - T)^{-1}. \end{aligned} \tag{20}$$

Finally, from Lemma 2 and Corollary 1, we get

$$\begin{aligned} \lim_{\lambda \to 0} \lambda^{-1} D_B &= -\frac{\beta}{\alpha} q_z(1, 1, \cdot)T^{-1}e \\ &= \frac{\alpha}{\beta}\frac{1}{\mu}(-\xi T^{-1}e) - \frac{1}{\mu}\xi T^{-1}(\beta I - T)^{-1}E \\ &= \frac{\alpha}{\beta}\frac{\mu_2}{2} + \int_0^\infty e^{-\beta x}\int_x^\infty H^c(s)dsdx. \end{aligned}$$

Normalizing by the expected delay in the M/G/1 queue and integrating by parts gives us (3). □

We next calculate the Laplace Transform $E(e^{-sD} \mid D > 0)$. In light traffic, a customer who is delayed ($D > 0$) will usually see only one customer in the system and will just wait for the service completion. This is made rigorous by Lemma 1. The probability that such a customer arrives and finds the server in phase k is given by

$$\frac{\sum_m m\rho(1, m, k)}{\sum_{m,n} m\rho(1, m, n)},$$

or in terms of q this is

$$\frac{q_z(1, 1, k)}{\sum_n q_z(1, 1, n)}. \tag{21}$$

From (21) we get that

$$E(e^{-sD} \mid D > 0) = \frac{\sum_k q_z(1, 1, k) E_k(e^{-s\tau})}{\sum_n q_z(1, 1, n)}$$

$$= \frac{q_z(1, 1, \cdot)(sI - T)^{-1}\mathbf{E}}{q_z(1, 1, \cdot)e}, \tag{22}$$

where $E_k(e^{-s\tau})$ is the Laplace transform of the remaining service time given that the current phase is k, and we have used proposition (1) from Appendix A.

Evaluating the denominator first, from (20) and Corollary 1 we see that

$$[q_z(1, 1, \cdot), e] = \left(\frac{\alpha}{\beta}\right)^2 \frac{1}{\mu}(\xi, e) + \left(\frac{\alpha}{\beta}\right) \frac{1}{\mu} \xi(BI - T)^{-1}\mathbf{E}$$

$$= \left(\frac{\alpha}{\beta}\right)^2 \frac{1}{\mu} + \left(\frac{\alpha}{\beta}\right) \int_0^\infty e^{-\beta x} H^c(x) dx$$

$$= \left(\frac{\alpha}{\beta}\right)^2 \frac{1}{\mu} + \left(\frac{\alpha}{\beta}\right) \frac{1 - \tilde{H}(\beta)}{\beta},$$

where ξ is the vector of ergodic probabilities for the service-time variable and $\tilde{H}(s) = \int e^{-sx} dH(x)$ is the Laplace Transform of the service-time density. The numerator, after some algebraic reduction, becomes

$$q_z(1, 1, \cdot)(sI - T)^{-1}\mathbf{E} = \left(\frac{\alpha}{\beta}\right)^2 \omega[(-T)^{-1}(sI - T)^{-1})\mathbf{E}$$

$$+ \left(\frac{\alpha}{\beta}\right) \frac{\omega(\beta I - T)^{-1}\mathbf{E} - \omega(sI - T)^{-1}\mathbf{E}}{s - \beta}$$

$$= \left(\frac{\alpha}{\beta}\right)^2 \frac{1 - \tilde{H}(s)}{s} + \left(\frac{\alpha}{\beta}\right) \frac{\tilde{H}(\beta) - \tilde{H}(s)}{s - \beta}.$$

Combining the two calculations, we get

$$E(e^{-sD} \mid D > 0) = \frac{\alpha\left[\frac{1 - H(s)}{s}\right] + \frac{\beta}{s - \beta}[\tilde{H}(\beta) - \tilde{H}(s)]}{\alpha \frac{1}{\mu} + 1 - \tilde{H}(\beta)}. \tag{23}$$

It is not difficult to show that (23) is the transform of (10).

IV. DERIVATION OF THE HEAVY-TRAFFIC RESULTS

In this section we describe the technique that allows us to arrive at (5) and (6). This approximation is derived by first proving that the number of packets in the system, when appropriately scaled, converges to a diffusion process $X(t)$ as $\rho^{(N)}$ converges to one. Given the drift and infinitesimal variance of the diffusion process, one can obtain the steady-state mean of $X(t)$, and by using Little's Theorem, we get the approximation given in (5).

The main theorem of this section, which can be proved rigorously using the techniques in Burman,[12] forms the theoretical basis for (6). We will outline its proof and show in detail how to calculate the mean and variance of the resulting diffusion, which from an application point of view is the difficult part of developing these approximations.

Consider a sequence of processes $X_n(t)$, where the nth process represents the number of packets at time t in a single-server system to which packets arrive at an instantaneous rate λ_n times the state of an M/M/N/N queueing system (representing the number of virtual calls). The heavy-traffic limits are found by defining λ_n as follows:

$$\lambda_n E^{(N)}V = \mu - \frac{\delta}{\sqrt{n}}, \tag{24}$$

where δ is a positive constant and $E^{(N)}V$ is the expected number of calls in the M/M/N/N system. We are initially interested in studying the sequence of scaled processes $n^{-1/2}X_n(nt)$ as $n \to \infty$.

We start by investigating the limiting behavior of the infinitesimal generators of a sequence of Markov processes created by appending supplementary variables to $X_n(t)$. A limiting generator is identified and the Trotter-Kato Theorem (see Kato[13]) implies that the finite-dimensional distributions converge. The form of the limiting generator completely determines the limiting process. Weak convergence can be established via a theorem of Stroock and Varadhan.[14]

Let $V(t)$ be the number of virtual calls at time t and let $Y(t)$ be the time since the packet currently in service entered service. The multi-dimensional process

$$M_n(t) = [n^{-1/2}X_n(nt),\ Y(nt),\ V(nt)]$$

is Markov. We denote a typical element of the state-space as (x, y, j), where $x = 0, n^{-1/2}, 2n^{-1/2}, \cdots, y \in [0, \infty)$ and $j = \{0, 1, \cdots, N\}$. For the remainder of this paper, let $h(y)$ be the density of the service-time distribution,

$$H^c(y) = \int_y^\infty h(s)ds, \qquad \mu(y) = \frac{h(y)}{H^c(y)}$$

and assume that for some constants γ and Γ,

$$0 < \gamma \leq \mu(y) \leq \Gamma \leq \infty. \tag{25}$$

The implications of and conditions for relaxing (25) are discussed in Burman.[12] These conditions are required for the method of proof and do not pose any significant restrictions on the results.

Let B and Q be the infinitesimal generators of the processes $V(t)$ and $Y(t)$, respectively. The generator B is an $(N+1) \times (N+1)$ matrix given by

$$B = \begin{bmatrix} -\alpha & \alpha & & & & \\ \beta & -(\beta+\alpha) & \alpha & & & \\ & 2\beta & -(2\beta+\alpha) & \alpha & & \\ & & & \ddots & & \\ & & & & N\beta & -N\beta \end{bmatrix}$$

and Q is a first-order linear differential operator given by

$$Qf(y) = f'(y) + \mu(y)[f(0) - f(y)].$$

The following two lemmas are needed to carry out the calculations needed in the main theorem of the section (to be stated below):

Lemma 3: The equation

$$Bv = w \tag{26}$$

has a solution v if and only if

$$(e, w) = 0, \tag{27}$$

where $e_k = 1/k!(\alpha/\beta)^k = 1/k!\ \rho^k$. In this case v is unique to an additive constant and is given by

$$v(k+1) = v(0) + \frac{1}{\alpha} \sum_{j=0}^{k} \frac{j!}{\rho^j} \sum_{i=0}^{i} \frac{\rho^i}{i!} w(i). \tag{28}$$

Proof: Since B is a finite-dimensional linear operator corresponding to an ergodic Markov process, it has a one-dimensional null space spanned by the vector e. Hence, (27) is both necessary and sufficient for a unique solution of (26) to exist. This solution is easily seen to be given by (28). □

Lemma 4: The equation

$$Qf = g \tag{29}$$

has a solution f if and only if

$$\int_0^\infty g(y) H^c(y) dy = 0. \tag{30}$$

In this case, f is unique to an additive constant and is given by

$$f(y) = \frac{1}{H^c(y)} \int_0^y H^c(s)g(s)ds. \tag{31}$$

Proof: The operator Q is the generator of an ergodic Markov process with the stationary density $\mu H^c(y)$. A unique solution exists if and only if (30) holds. Solving the first-order linear differential equation gives (31). □

We are now ready to state the main theorem of this section:

Theorem 1: Under assumption (25),

$$n^{-1/2}X_n(nt) \Rightarrow X(t),$$

where $X(t)$ is a diffusion on R^+ with pure reflection at the origin, downward drift δ and infinitesimal variance s^2 given by

$$s^2 = \frac{1}{2}\left\{\lambda E^{(N)}V + \mu^3\sigma^2 + \sum_{k=0}^{N} \frac{1}{\alpha P_k}\left[\sum_{j=0}^{k-1} P_j(\mu - j)\right]^2\right\}, \tag{32}$$

where

$$P_k = \frac{1}{k!}\left(\frac{\alpha}{\beta}\right)^k \Big/ \sum_{j=0}^{N} \frac{1}{j!}\left(\frac{\alpha}{\beta}\right)^j$$

is the steady-state probability of finding k calls in the $M/M/N/N$ blocking system.

Proof of Theorem: As mentioned earlier, we will show how to identify the mean and variance of (32). The infinitesimal generator for $M_n(t)$ is given by

$$A_nf(x, y, j) = n(B + Q)f + \lambda j\left[f\left(x + \frac{1}{\sqrt{n}}, y, j\right) - f(x, y, j)\right]$$
$$+ \mu(y)\left[f\left(x - \frac{1}{\sqrt{n}}, 0, j\right) - f(x, 0, j)\right] \quad \text{for} \quad x \geq \frac{1}{\sqrt{n}}, \tag{33}$$

and

$$A_nf(0, 0, j) = nBf + \lambda j\left[f\left(\frac{1}{\sqrt{n}}, 0, j\right) - f(0, 0, j)\right]. \tag{34}$$

For $f(x)$ twice continuously differentiable with $f'(0) = 0$, set

$$f_n(x, y, j) = f(x) + \frac{1}{\sqrt{n}} f'(x)g(y, j) + \frac{1}{n} f''(x)h(y, j). \tag{35}$$

We construct bounded functions g and h so that

$$A_nf_n \to Af$$

uniformly, where

$$Af(x) = mf'(x) + s^2 f''(x).$$

The constants m and s^2 are determined in the process. Given (25), these constructions will complete the proof (see Burman[12]).

From (35) and (33) we see that

$$A_n f_n = n(B + Q)f + \sqrt{n} f'(x)[\lambda_n j - \mu(y) + (B + Q)g] + f''(x)\left\{\frac{1}{2}[\lambda_n j + \mu(y)] + \lambda_n j g(y, j) - \mu(y) g(0, j) + (B + Q)h(y, j)\right\} + 0\left(\frac{1}{\sqrt{n}}\right). \tag{36}$$

First note that since f is independent of y and j, $(B + Q)f = 0$. Set g equal to the solution of the equation

$$(B + Q)g = -[\lambda_n j - \lambda_n E^{(N)} V - \mu(y) + \mu]. \tag{37}$$

Lemmas 3 and 4 imply that g exists and

$$g(y, j) = \frac{1}{H^c(y)} \int_0^y H^c(s)[\mu(s) - \mu] ds - \frac{\lambda_n}{\alpha} \sum_{k=0}^{j-1} \frac{k!}{\rho^k} \sum_{i=0}^{k} \frac{\rho^i}{i!} [j - E^{(N)} V]. \tag{38}$$

Let h be the solution of the equation

$$(Q + B)h = -[w(y, j) - s^2], \tag{39}$$

where $w(y, j) = 1/2[\lambda_n j + \mu(y)] + \lambda_n j g(y, j) - \mu(y) g(0, j)$. The constant s^2 [see (26)] can be easily calculated from conditions (27) and (30), i.e.,

$$\sum_{i=0}^{N} \frac{\rho^i}{i!} \int_0^\infty H^c(y) w(y, i) dy = 0.$$

Using g and h calculated above we have

$$A_n f_n = -\delta f'(x) + s^2 f''(x) + 0\left(\frac{1}{\sqrt{n}}\right).$$

When $x = 0$ analogous calculations give that $A_n f_n$ converges provided $f'(0) = 0$, and the proof is complete. □

The limit theorem for the M/G/1 queue with arrival rate $\lambda_n E^{(N)} V$ (see Iglehart and Whitt[15]) gives the same mean δ with the variance $\lambda E^{(N)} V + \mu^3 \mu_2$ for the limiting diffusion. Calculating the steady-state expected number in the limiting systems and using Little's Theorem

(see Cooper[16]) yields the asymptotic expression for the ratio of $D_B^{(N)}$ and $D_M^{(N)}$, the delays in the two systems:

$$\frac{D_B^{(N)}}{D_M^{(N)}} \approx 1 + 2\,\frac{\sum_{k=0}^{N-1} \frac{1}{P_k} \left[\sum_{i=0}^{k} P_i(\mu - \lambda i)\right]^2}{\alpha\mu^3\mu_2}. \tag{40}$$

Letting N go to infinity, P_k goes to the Poisson distribution and (40) goes to

$$1 + 2\mu/\alpha\left(\frac{1}{1 + C^2}\right)$$

as promised. □

Remark: We are careful here not to write (40) as a limit but as an approximation. In order to strictly prove convergence in (40), two rather technical steps remain to be proved. The first is that the steady-state number in the queue (when scaled) converges to the steady-state value of the diffusion. The second is that the expected values of these steady states converge. Both are difficult issues in themselves and, aside from mathematical completeness, add little information to the approximation, which is the germaine issue of this paper.

V. CONCLUDING REMARKS

We considered here a single-server queueing system with a nonhomogeneous Poisson arrival process whose rate is some constant λ times the state of an independent M/M/∞ system. In this paper, we derived limiting values for the mean delays as the traffic intensity goes to zero and to one.

This system was used to model the delay of packets at a packet switch. In this setting, an individual data customer generates packets at a constant rate λ and the number of customers generating packets is given by the state of the M/M/∞ system. This is, of course, not the only possible model for packet arrivals. Packets generated by an individual virtual call may be "smoother" or more "bursty" than packets generated by a Poisson process. It is also possible that restrictions could be imposed on the total number of simultaneous virtual calls, thereby making the infinite-server assumption unrealistic. Work is currently under way to extend the techniques of this paper to cover these and other more general models of bursty traffic.

Our ultimate goal in studying D_B/D_M in both light and heavy traffic is to be able to derive simple, closed-form approximations for D_B for all values of the traffic intensity ρ. One candidate approximation is to

linearly interpolate between the values obtained for D_B/B_M when $\rho \to 0$ and when $\rho \to 1$ [see (4) and (5)]. This approach gives the following estimate for the mean delay:

$$D_B \sim \frac{\rho}{1-\rho}\frac{(1+C^2)}{2\mu}\left[1+\frac{\mu}{\alpha}\frac{1-(1-\rho)\phi(\beta)}{(1+C^2)}\right].$$

More sophisticated interpolations are possible (see Lauber and Smith[7] for applications of these techniques to estimating the delays in an M/G/c queue); however, in the absence of any test data, verification is difficult. One method of verification is to study the accuracy of these approximation techniques when applied to other models of bursty arrivals for which explicit expressions are known (see Yechiali and Naor[17]). This work is currently under way and will be reported on in the future.

REFERENCES

1. H. Heffes, "A Class of Data Traffic Processes—Covariance Function Characterization and Related Queueing Results," B.S.T.J., *59*, No. 6 (July–August 1980), pp. 897–929.
2. A. Kuczura, "Queues with Mixed Renewal and Poisson Input," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1305–26.
3. R. V. Laue, unpublished work.
4. M. F. Neuts, "A Versatile Markovian Point Process," J. Appl. Prob., *16* (December 1979), pp. 764–79.
5. D. M. Lucantoni and M. F. Neuts, "Numerical Methods for a Class of Markov Chains Arising in Queueing Theory," Technical Report No. 78/10 (May 1978), Department of Statistics and Computer Science, University of Delaware.
6. D. Anick, D. Mitra, and M. M. Sondhi, unpublished work.
7. P. J. Lauber and D. R. Smith, unpublished work.
8. D. Mitra and D. Anick, unpublished work.
9. L. Kosten, "Stochastic Theory of a Multi-Entry Buffer (1)," Delft Progress Report, *1*, 10–18, 1974.
10. D. Y. Burman and D. R. Smith, "A Light Traffic Theorem for Multi-Server Queues," Math. Oper. Res., *8*, No. 1 (February 1983), pp. 15–25.
11. D. R. Smith, "Optimal Repairman Allocation—Asymptotic Results," Management Science, *24*, No. 6 (February 1978), pp. 665–74.
12. D. Y. Burman, unpublished work.
13. T. Kato, *Perturbation Theory for Linear Operators*, Berlin: Springer-Verlag, 1976.
14. D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Berlin: Springer-Verlag, 1979.
15. D. L. Igelhart and W. Whitt, "Multiple Channel Queues in Heavy Traffic," I. *Advances in Applied Probability*, 2, No. 1 (Spring 1970), pp. 150–77.
16. R. Cooper, *Introduction to Queueing Theory*, New York: North Holland, 1981.
17. V. Yechiali and P. Naor, "Queueing Problems with Heterogeneous Arrivals and Service," Oper. Res., *19*, No. 3 (May–June 1971), pp. 722–34.
18. J. Keilson, *Markov Chain Models—Rarity and Exponentiality*, New York: Springer-Verlag, 1979.

APPENDIX A

Background on Phase-Type Distributions

Appendix A summarizes results for phase-type distributions that are used in Section III. A service time is said to have phase-type distribution if it is distributed like the first exit time from a continu-

ous-time, finite-state Markov chain. Any distribution can be arbitrarily approximated in the sense of weak convergence by one of phase type. We assume that the states (also known as phases) are the nonnegative integers i, $i = 1, \cdots, m$. Let the rate of transition from phase i to phase j be T_{ij} $(i \neq j)$, and the rate of exiting from phase i be ξ_i. Define $T_{ii} = -\sum_{j \neq i} T_{ij} - \mathrm{E}_i$, i.e., minus the rate of leaving state i. We assume that T, the matrix of T_{ij}, is invertible; this is sufficient to imply that the real part of the spectrum of T is strictly negative. A customer starts service in phase i with probability ω_i. Following Neuts,[4] we use the notation (ω, T) to describe this distribution, where ω is the vector of initial probabilities and T is the m-dimensional matrix of rates.

The vector ξ of ergodic probabilities for the service phase of the renewal process is determined by normalizing the solution to the equations

$$-\xi_i T_{ii} = \sum_j \xi_j T_{ji} + (\Sigma \xi_j \mathrm{E}_j)\omega_i, \tag{41}$$

or in matrix notation,

$$\xi T = -(\xi, \mathrm{E})\omega. \tag{42}$$

The service rate μ is defined as

$$\mu = (\xi, \mathrm{E}). \tag{43}$$

It is not difficult to show that μ is the reciprocal of the mean service time.

Let τ be the first exit time from the chain; let $\mathrm{M}_i(s) = E_i(e^{-s\tau})$, the Laplace Transform of τ given the initial state is i; and $M_i^n = E_i\tau^n$. The results of Proposition 1 are well known (see Kielson,[18] page 82, and Burman and Smith[12]).

Proposition 1. For τ, $\mathrm{M}(s)$, and μ^n defined above,

$$\mathrm{M(s)} = (\mathrm{sI} - \mathrm{T})^{-1}\mathrm{E} \tag{44}$$

and

$$M^n = n!(-1)^n T^{-n} e, \tag{45}$$

where e is the vector of all ones. In particular, the nth moment of the service-time distribution $(\tilde{M}^n)$ is

$$\tilde{M}^n = n!(-1)^n(\omega, T^{-n}e). \tag{46}$$

If $H^c(x)$ equals the tail of the service-time distribution, then the following corollary is immediate.

Corollary 1.

$$-\xi T^{-1}e = \int_0^\infty x\mu H^e(x)dx = \frac{\mu\mu_2}{2}, \tag{47}$$

and

$$-\xi T^{-1}(\beta I - T)^{-1}\mathbf{E} = \int_0^\infty e^{-\beta x}\int_x^\infty \mu H^e(s)dsdx. \tag{48}$$

Proof: To see (47), observe that $\mu H^e(x)$ is the density of the remaining service time when the process is sampled in equilibrium, and apply Proposition 1. The identity on the integral can be obtained by integration by parts. In a similar fashion,

$$\xi T^{-1}(T - \beta I)^{-1}\mathbf{E} = \frac{\xi[(T - \beta I)^{-1} - T^{-1}]\mathbf{E}}{\beta}$$

$$= \frac{\left[\int_0^\infty \mu H^e(x)dx - \int_0^\infty e^{-\beta x}\mu H^e(x)dx\right]}{\beta},$$

and integration by parts gives us (48). □

APPENDIX B

Intuitive Derivation of the Results in Light Traffic

The following argument is based on the intuitive notion that, in light traffic, almost all arriving packets arrive to an empty system and are served before the arrival of another packet. Furthermore, the times when there is exactly one packet in the system constitute almost all of the time in which an arriving packet might be subject to delay. Thus, in light traffic, it is easy to derive the proportion of time in which an arrival will be delayed, and to find the delay for such an arrival. A slight complication is introduced in the analysis by the fact that the arrival rate of packets is not constant, but this is easily overcome.

While the overall derivation is not rigorous, it is convincing and gives insight into the light-traffic behavior. The results, of course, are consistent with the rigorous results derived for phase-type distributions in Section III.

We will begin with traffic-independent results for the model described previously in the beginning of Section II, in which the instantaneous arrival rate equals λ times the state of an M/M/∞ queue with birth rate α and service rate equal to β. The equilibrium distribution of the M/M/∞ queue governing the arrival process is easily seen to be Poisson with mean α/β. However, this is not its distribution immediately after an arrival, since the rate of arrivals is proportional to the

state of the M/M/∞ queue. Thus, the conditional probability that the state of the M/M/∞ queue is m (for $m \geq 1$) immediately after an arrival is

$$\frac{m(\alpha/\beta)^m \frac{1}{m!} e^{-\alpha/\beta}}{\sum_{k=0}^{\infty} k(\alpha/\beta)^k \frac{1}{k!} e^{-\alpha/\beta}} = \frac{1}{(m-1)!} (\alpha/\beta)^{m-1} e^{-\alpha/\beta}.$$

This may be thought of as the distribution of a Poisson random variable with mean (α/β) with support right-shifted one unit. This point of view is useful in computing the distribution of the state of the M/M/∞ queue t time units after an arrival of a packet (denoted N_t), since the Poisson distribution is stationary, and the additional unit is still present with probability $e^{-\beta t}$. Thus,

$$E(z^{N_t}) = e^{(\alpha/\beta)(z-1)}[1 + e^{-\beta t}(z-1)]. \tag{49}$$

Now consider the system in light traffic. Denote the state of the system by (N, M, T) where N is the number of packets in the system, M is the state of the M/M/∞ queue that modulates the arrival process, and T is the remaining service time of the packet being served.

We first obtain an expression for $E(e^{-sT}z^M\delta_{1N})$ in light traffic, where $\delta_{1N} = 0$ for $N \neq 1$ and $\delta_{11} = 1$. We then show how to use this quantity to obtain the desired limits. In light traffic, almost every packet arrives to an empty system and is served before another arrival. By Little's Law

$$P(N = 1) = E\delta_{1N} \sim \lambda\alpha E(S)/\beta.$$

Again, making the assumption of the last sentence and looking at the system at only those times for which $N = 1$, we obtain by standard renewal theory arguments that

$$E(e^{-sT}z^M N = 1) = \frac{1}{ES} E\left[\int_0^S E(z^{N_t}) e^{-s(S-t)} dt\right] \stackrel{\text{def}}{=} \frac{1}{ES} G(z, s).$$

Multiplying, we obtain that

$$E(e^{-sT}z^M\delta_{1N}) \sim \lambda(\alpha/\beta)G(z, s). \tag{50}$$

By conditioning on S and using (49), we obtain

$$G(z, s) = e^{(\alpha/\beta)(z-1)} \frac{1}{s} \left\{[1 - \tilde{H}(s)] + \left(\frac{z-1}{\beta - s}\right)[\tilde{H}(s) - \tilde{H}(\beta)\right\}, \tag{51}$$

where $\tilde{H}$ is the Laplace Transform of a service time.

Equations (50) and (51) can be used to find desired properties of the queue in light traffic. For example, since nearly all packets that

are delayed are generated when there is only one other packet in the system, and since the rate of generation of packets is λ times the number of calls, we find the rate of generation of delayed packets, λ_D, is

$$\lambda_D \sim \lambda^2(\alpha/\beta)G_z(1, 0), \tag{52}$$

or

$$\lambda_D \sim \lambda^2 \left\{(\alpha/\beta)^2 \frac{1}{\mu} + \alpha/\beta[1 - \tilde{H}(\beta)]\right\}. \tag{53}$$

The Laplace Transform of the delay, given that the delay is greater than 0, defined to be $\phi(s)$, is also found easily since the delay of a packet equals the remaining service time on arrival. Thus,

$$\phi(s) \sim \frac{G_z(1, s)}{G_z(1, 0)},$$

or

$$\phi(s) \sim \frac{\alpha/s[1 - \tilde{H}(s)] + \dfrac{\beta}{\beta - s}[\tilde{H}(s) - \tilde{H}(\beta)]}{\alpha/\mu + 1 - \tilde{H}(\beta)}. \tag{54}$$

Inversion of the Laplace Transform gives (10) and also gives

$$E(D \mid D > 0) \approx \frac{\dfrac{\alpha\mu_2}{2} + \dfrac{1}{\mu} - \dfrac{1}{\beta}[1 - \tilde{H}(\beta)]}{\alpha/\mu + 1 - \tilde{H}(\beta)}, \tag{55}$$

where $\mu_2 = E(S^2)$. Of course, $P(D > 0) = \dfrac{\lambda_D}{\lambda(\alpha/\beta)}$ so that (52) and (54) give

$$E(D) \approx \frac{\lambda}{\beta}\left\{\frac{\alpha\mu_2}{2} + \frac{1}{\mu} - \frac{1}{\beta}[1 - \tilde{H}(\beta)]\right\}. \tag{56}$$

This may be combined with the Pollaczek-Khintchine formula to give (3).

AUTHORS

Donald R. Smith, A.B. (Physics), 1969, Cornell University; M.S. (Operations Research), 1974, Columbia University; Ph.D. (Operations Research), 1975, University of California, Berkeley; Bell Laboratories, 1980—. Before joining Bell Laboratories, Mr. Smith was employed at Adaptive Technology, Inc. from 1970 to 1974, and was Assistant Professor in the Department of Industrial Engineering and Operations Research, Columbia University from 1975 to 1979. At Adaptive Technology, Mr. Smith developed mathematical models for

new techniques in statistical multiplexing. At Bell Laboratories he is in the Operations Research Department of the Network Analysis Center pursuing interests in applied stochastic processes.

David Y. Burman, B.S. (Mathematics), 1968, C.C.N.Y.; Ph.D. (Applied Mathematics), 1979, New York University; Bell Laboratories, 1969—. At Bell Laboratories, Mr. Burman has worked on problems in network flows, scheduling, and stochastic processes.