Dependence in Packet Queues

KERRY W. FENDICK, VIKRAM R. SAKSENA, SENIOR MEMBER, IEEE, AND WARD WHITT

Abstract—**Many packet communication networks carry several classes of traffic, each with its own service characteristics, e.g., packetized voice, short data messages, bulk data (files), facsimile, acknowledgments, and other signals. The packet arrival processes from each source are also often bursty (highly variable), which can contribute to long packet delays. The burstiness of the total arrival process has been previously characterized in packet network performance models by the dependence among successive interarrival times. Here it is shown that associated dependence among successive service times and between service times and interarrival times also can be important for packet queues involving variable packet lengths. These dependence effects are demonstrated analytically by considering a multiclass single-server queue with batch-Poisson arrival processes. For this model and more realistic models of packet queues, insight is gained from heavy-traffic limit theorems. This study indicates that all three kinds of dependence should be considered in the analysis and measurement of packet queues involving variables packet lengths. Indeed, specific measurements are proposed to apply to data from real systems and simulations. This study also indicates how to predict expected packet delays under heavy loads either from the proposed measurements or from model parameters. Finally, this study is important for understanding the limitations of procedures such as the queueing network analyzer (QNA) for approximately describing the performance of queueing networks using the techniques of aggregation and decomposition.**

I. Introduction

To properly engineer packet networks, it is necessary to identify characteristics of the network and its traffic that significantly affect performance. (The particular performance measures we consider are the average workload and the average delay per packet at individual queues in the network.) Three important characteristics that may be present are: 1) multiple classes of traffic (i.e., multiple applications: packetized voice, bulk data, signals, etc.), each with its own service requirements, 2) bursty (highly variable) arrival processes, and 3) a very large number of sources sharing the system. These characteristics interact in a complicated way, so that analyzing the performance of packet networks can be very difficult. In this paper, we show how the first two characteristics tend to cause performance problems (relatively long packet delays), even though the third characteristic tends to mitigate the problems caused by the first two.

We propose a way to analyze packet queues in order to gain insight into the basic characteristics above. In particular, *we suggest focusing on the dependence among the interarrival times and service times at the queue*. (Dependence among interarrival times and service times at *different queues* can be studied in similar ways, but we only consider a *single queue* here. It could be any queue in a network though.) We indicate how the dependence among the interarrival times and service times can be quantified and measured. Indeed, a major purpose of this paper is to identify appropriate statistics to compute from interarrival-time and service-time data in order to describe the dependence and predict system performance. The particular measurement we propose is a *three-dimensional index of dispersion for intervals*, defined in (15) below. It describes the cumulative correlations in the first n interarrival times alone, the first n service times alone, and the first n interarrival times and service times together, each as a function of n. Of course, these correlations are only a partial characterization of the dependence among the variables, but we believe a useful partial characterization. These measurements can be applied to *any queue*; they are not limited to a particular model. These measurements are different from the usual traffic measurements because they are specifically intended to *describe the variability* in the traffic.

To justify the proposed measurements, we link them to performance. A basic hypothesis underlying this work is that the correlations in the interarrival times and service times are a primary determinant of packet delays. We support this hypothesis here by proving that the three-dimensional index of dispersion actually determines the packet delay distribution under heavy loads. We obtain these results mathematically via heavy-traffic limit theorems.

We also indicate how the dependence and its effect on performance can be approximately described *analytically* in terms of model parameters. First, we propose a somewhat idealized multiclass batch-Poisson model in Section III, for which quantities of interest can be computed explicitly. Second, we propose central limit theorems and associated heavy traffic limit theorems that accurately describe both performance under heavy loads and the relevant long-term dependence for very general models.

II. Dependence Among Interarrival Times and Service Times

A. Correlated Service Times at Different Queues

Dependence between interarrival times and service times at a queue was first identified as an important issue in communication networks and was first seriously investigated by Kleinrock [1], see pp. 38 and 49. Kleinrock considered a queueing network model of a communication network, in which messages pass through a series of queues on their way from source to destination. Since the same message visits each queue on the path, the service times of each message at the successive queues it visits will typically be positively correlated and may even be identical (identical transmission rates). Moreover, if the service times at two queues in series are dependent, then the interarrival times and service times at the second queue are dependent. Evidently, if the two service times are both relatively long, then so will be the corresponding interarrival time and service time at the second queue. When this dependence is present in this form, i.e., when the interarrival times and service times are strongly positively correlated, the delays at the second queue tend to be *less* than if the dependence were not there. Under heavy loads, the delays at such a queue tend to be dramatically *less*, i.e., Conolly [2] and p. 618 and Part II of Boxma [3].

Kleinrock observed that if there is sufficient mixing of traffic, then the dependence effect may be small, which justifies ignoring it altogether (his celebrated independence

Paper approved by the Editor for Queueing/Buffer Management of the IEEE Communications Society. Manuscript received October 15, 1987; revised September 13, 1988.

K. W. Fendick and V. Saksena are with AT&T Bell Laboratories, Holmdel, NJ, 07733.

W. Whitt is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 8930924.

0090-6778/89/1100-1173$01.00 © 1989 IEEE

assumption on p. 50 of [1]). However, Kleinrock also noted that the dependence effect can be strong. Rubin [4] and Calo [5] studied ways to estimate end-to-end message delays when very strong dependence among the service times at different queues is present. Mitchell, Paulson, and Beswick [6] studied the dependence effect experimentally. Our paper is in this spirit. Our methods provide a systematic basis for estimating the effect of this dependence, both analytically and experimentally, even though we were motivated by a different problem.

B. Our Motivating Problem

This research was undertaken when extensive simulations of X.25 networks revealed packet delays at the transmission queues for both lines and trunks to be substantially *greater* than those predicted by standard *GI/G/*1 models, which assume independence among interarrival times and service times. A simple *GI/G/*1 approximation used to predict the expected waiting time is given by

$$EW = \frac{\tau\rho}{1-\rho} \frac{(c_a^2+c_s^2)}{2} \tag{1}$$

where τ is the mean packet service time, ρ is the traffic intensity, c_a^2 is the squared coefficient of variation (variance divided by the square of the mean) of an interarrival time, and c_s^2 is the squared coefficient of variation of a service time [7], [8]. Moreover, a direct simulation of a *GI/G/*1 queue using the estimated interarrival-time and service-time distributions typically yielded average delays slightly *less* than predicted by (1). (Indeed, there is an additional correction term in [7], [8] to reduce (1) somewhat.) Therefore, we tentatively concluded that the greater expected delays observed must be due to dependence in the interarrival times and service times, which is ignored in (1).

From further analysis (simulation experiments and mathematics), we conclude that indeed there is significant dependence in the packet interarrival times and packet service times of these queues. (Since space is limited and the results in this paper are very general, we will not describe the X.25 network and the various detailed models of it. A significant feature for the analysis here is variable packet lengths.)

Packet networks often serve *several classes of traffic* (e.g., packetized voice, short data messages, long data files, facsimile, acknowledgments and other signals), each with its own service characteristics (packet lengths and number of packets per message). In packet queues with several classes of traffic and variable packet lengths, we contend that one should expect to have *all* of the following:

1) successive packet interarrival times correlated,
2) successive packet service times correlated,
3) packet interarrival times and packet service times correlated.

Moreover, the successive interarrival time and the successive service times are each often positively correlated, while the interarrival times and services times are often negatively correlated. These three kinds of dependence (with these signs) each tend to make delays larger than they would be in the i.i.d case where there is no dependence. All three *together* tend to make delays *larger* than they would be with bursty arrival processes alone. In general, however, the correlations can be either positive or negative in each case. It is significant that the dependence effect on packet delays here is very different from the dependence effect on message delays in Section II-A; here the dependence makes the delays larger instead of smaller.

C. Dependence Among Successive Interarrival Times

The phenomenon of dependence in arrival processes to queues is beginning to be reasonably well understood [10]-[19]. Sriram and Whitt [11] emphasized the importance for packet queues of the dependence among successive interarrival times—and suggested the *index of dispersion for intervals* (IDI) as a reasonable way to partially characterize it. The IDI represents the cumulative correlation among successive interarrival times [18, pp. 71-72]. In [12], the limiting value of the IDI is referred to as the *asymptotic variability parameter* or *asymptotic squared coefficient of variation*, here denoted by c_A^2. It includes the effect of *all* the correlations. (See (15) for a definition.) The IDI is a sequence $\{c_k^2: 1 \leq k \leq \infty\}$ with $c_1^2 = c_a^2$ and $c_\infty^2 = c_A^2$. The full IDI is obviously more informative than the c_a^2 and c_A^2 alone, but c_a^2 and c_A^2 are useful partial characterizations that often can be obtained analytically. The full IDI shows how the variability develops over time (the arrival index).

Significant correlations among successive interarrival times in packet queues occurs primarily because of the combination of two factors. First, the superposition of independent arrival processes, each of which may be a renewal process, is not itself a renewal process unless all components processes are simple Poisson or batch Poisson with a common geometric batch-size distribution; see Section II of [9]. As shown in [10], [11], the superposition operation tends to convert the variability of the individual interarrival times in the component renewal processes into long-term correlations in the superposition process, i.e., the interarrival times in the superposition of a large number of independent renewal processes tend to be nearly exponentially distributed, and the lag-1 correlation tends to be small, but the variability reappears in many small correlations among interarrival times. When there is a very large number of sources (the third characteristic in Section I), the superposition arrival process behaves like a Poisson process, except over very large time intervals, where the dependence begins to reappear.

Second, the packet arrival process from each source tends to be bursty because of the packetization of large messages, i.e., the packets tend to arrive in bursts or clumps. The burstiness is often manifested in large correlations among interarrival times of each source (which are retained in the superposition of such sources), but it need not be: it may be reasonable to model the arrival process from each source as a renewal process, such as an interrupted Poisson process, in which there is no dependence at all. Then, the high variability is reflected by a high variance of a single interarrival time, which also appears as long-term correlations in the superposition. Thus, while the superposition of arbitrary arrival processes tends to result in some long-term correlations among interarrival times, the superposition of bursty arrival processes tends to result in *significant* long term correlations. This dependence over large time intervals becomes more relevant to queue performance as the traffic intensity increases; indeed, we show that all the correlations affect the heavy-traffic limit as $\rho \rightarrow 1$.

D. The Service Times Too

The primary purpose of this paper in relation to [10]-[19] is to point out the importance of both *dependence among successive service times* and *dependence between interarrival times and service times*, as well as dependence among successive interarrival times. We will show that all three kinds of dependence occur because of the presence of bursty arrivals and multiple sources with different mean service times (due to different packet lengths). These two additional forms of dependence are somewhat surprising, because they occur even when we assume that the service times from each source are i.i.d., and that all the service times and arrival processes from different sources are mutually independent. The dependence occurs because the mean service times from different sources are different and because several packets from the same source tend to be served consecutively (because of the burstiness of the arrival processes from individual sources).

III. A Multiclass Batch-Poisson Model

To provide insight into these three kinds of dependence, we first consider a relatively simple analytical model. In particular, we assume that the arrival processes of k customer classes are independent batch-Poisson processes. (In the context of packet networks, a customer class can be thought of as a virtual circuit carrying a particular kind of traffic, such as facsimile or bulk data. We are not modeling acknowledgments or windows. Each batch can be regarded as a message, while the customers within a batch can be regarded as packets.) As an approximation, here we assume that all customers (packets) in the same batch (message) arrive at the same instant. (We relax this assumption in Section VI.) For class i, batches arrive according to a Poisson process at rate λp_i; the successive batch sizes are i.i.d. with mean m_i and squared coefficient of variation c_{bi}^2; the packet service times are i.i.d. with mean τ_i and squared coefficient of variation c_{si}^2. (We assume that $p_1 + \cdots + p_x = 1$, so that λ is the total arrival rate of batches. The total arrival rate of packets is thus $\bar{\lambda} = \lambda m_B$ where $m_B = \Sigma_{i=1}^k p_i m_i$ is the mean batch size.) Service is provided by a single server with an unlimited waiting room and the FIFO (first-in, first-out) discipline. The model can be denoted by $(\Sigma M^{B_i}/G_i)/1$. We have not specified the batch-size distributions and the service-time distributions, which can be general. It turns out that they (beyond the given parameters) do not affect the expected equilibrium workload (or virtual waiting time), which is what we primarily focus on here. (The workload at time t is the total remaining service time of all packets waiting to be processed at time t, including the packet being served.)

This multiclass batch-Poisson model exhibits all three kinds of dependence. The overall arrival process is a batch-Poisson process having a batch-size distribution that is a mixture of the component batch-size distributions. A batch-Poisson process is a renewal process if and only if the batch-size distribution is geometrically distributed on the positive integers. Thus, the component batch-Poisson processes may well be renewal processes, but the superposition process is typically not.

Dependence between successive service times in the superposition process occurs because the different classes have different service time distributions: an arriving batch causes several service times of one type to occur successively, thus the positive correlation between service times. If large batches have short (long) service times, then the correlation between interarrival times and service times could be positive (negative) because the zero interarrival times of the customers in the batch get matched with the service times.

At this point a disclaimer is in order. We do not presume that this batch-Poisson "message" model always realistically describes packet delays in packet queues. In fact, experience indicates that this $(\Sigma M^{B_i}/G_i)1$ model usually does not describe packet delays well under light-to-moderate loads, because the packets associated with a message do not actually arrive at one instant. Instead, packet arrivals are necessarily separated because of obvious constraints imposed by packet length and transmission rates. Nevertheless, it is apparent that *this simple model captures important qualitative features of a real packet queue*. In packet queues the arrival processes are often remarkably bursty and the lengths of the packets for different classes are often quite different. Although this simple model does not describe delays well under light-to-moderate loads, we will prove that *it does describe these delays under heavy loads well and it does describe the total dependence in the interarrival times and service times well.* (Under light-to-moderate loads this dependence usually does not affect packet delays so seriously.) To prove that the batch-Poisson model describes packet delays under heavy loads well, we also determine the heavy-traffic limits for more complicated models in which the spacing between packets is included. *We prove that the heavy-traffic limits for these models and the batch-Poisson model coincide.* The $(\Sigma M^{B_i}/G_i)/1$ model thus accurately describes the limiting degradation of performance as the load increases. With the results here, it is thus possible to determine whether the dependence effects can potentially have a significant impact as the load increases in any particular application.

Our main mathematical result for the multiclass batch-Poisson queueing model concerns the expected steady-state workload (or virtual waiting time) at an arbitrary time. We conclude that the expected steady-state workload can be expressed *exactly* as

$$EL = \frac{\tau\rho}{1-\rho}\,\frac{(c_A^2 + c_S^2 - 2c_{AS}^2)}{2} \tag{2}$$

where ρ is the traffic intensity, τ is the average service time for individual packets, and λ is the total Poisson arrival rate for batches, defined by

$$\rho = \bar{\lambda}\tau = \lambda\hat{\tau}, \quad \hat{\tau} = \sum_{i=1}^{k} p_i m_i \tau_i \text{ and}$$

$$\tau = \hat{\tau}/m_B = \sum_{i=1}^{k} p_i m_i \tau_i \Big/ \sum_{i=1}^{k} p_i m_i,$$

c_A^2 and c_S^2 are the asymptotic squared coefficients of variation for the interarrival times and service times, respectively, and c_{AS}^2 is a corresponding asymptotic correlation coefficient between the interarrival times and service times to be defined in Section V; see (15)–(17) below. In other words, for this model the *asymptotic-method approximation* [12], appropriately generalized to cover the correlations between interarrival times and service times via c_{AS}^2, is *exact*. Moreover, the asymptotic parameters c_A^2, c_S^2, and c_{AS}^2 typically differ from the stationary-interval counterparts c_a^2, c_s^2, and c_{as}^2, so that the dependence (measured by the quantities $c_A^2 - c_a^2$, $c_S^2 - c_s^2$, and $c_{AS}^2 - c_{as}^2$) matters. Furthermore, $[c_A^2 + c_S^2 - 2c_{AS}^2]/2$ is typically large. Indeed, if $[c_A^2 + c_S^2 - 2c_{AS}^2]/2$ were nearly one, then a simple $M/M/1$ approximation would be fine. The analysis here is vitally important because $[c_A^2 + c_S^2 - 2c_{AS}^2]/2$ is *often much greater than one*.

It is important that for the multiclass batch-Poisson model EL can be expressed directly and simply in terms of the asymptotic variability parameters c_A^2, c_S^2, and c_{AS}^2 and that they in turn quantify the three kinds of dependence. Moreover, they can be compared with their stationary-interval counterparts c_a^2, c_s^2, and c_{as}^2 to determine the source of the variability. The six parameters c_A^2, c_S^2, c_{AS}^2, c_a^2, c_s^2, and c_{as}^2 can in turn be expressed directly in terms of the original model parameters $[\lambda, p_i, m_i, c_{bi}^2, \tau_i, c_{si}^2]$ to determine how model structure affects the different components of variability; see (17).

To appreciate the relevance of dependence in packet queues, it helps to see measurements of these variability parameters for more realistic models. One case of an X.25 link with 25 separate sources yielded *simulation estimates* of

$$c_a^2 = 1.79, \quad c_s^2 = 1.06 \text{ and } c_{as}^2 = 0.03$$

$$c_A^2 = 17.6, \quad c_S^2 = 35.1 \text{ and } c_{AS}^2 = -6.7. \tag{3}$$

(We do not describe this relatively complex model here.) The $M/M/1$, stationary-interval and asymptotic approximations for $(1 - \rho)EL/\tau\rho$ based on (1)–(3) are thus 1.0, 1.4, and 33.1, respectively. The main point is that the measurements we propose, in particular, the stationary-interval and asymptotic variability parameters, suggest a dramatic increase in delays relative to the $M/M/1$ model as ρ increases, which indeed occurs unless the dependence is controlled as ρ increases (i.e., unless $c_A^2 + c_S^2 - 2c_{AS}^2$ is reduced as ρ increases, e.g., by windows or other flow control mechanisms).

The corresponding analytic results for the multiclass batch-Poisson model without spacing or acknowledgments and a more realistic model with spacing and acknowledgments

(Section VI-C) are, respectively,

$$c_A^2=22.0,\ c_S^2=35.0 \text{ and } c_{AS}^2=-9.8$$
$$c_A^2=20.1,\ c_S^2=39.6 \text{ and } c_{AS}^2=-8.0. \tag{4}$$

The results in (3) and (4) demonstrate that the asymptotic variability parameters from the multiclass batch-Poisson model, the more general model with spacing and acknowledgments, and the simulation estimates are all very close. Thus, the relatively simple multiclass batch-Poisson model is useful to describe both the asymptotic variability parameters and the performance under heavy loads. Under light-to-moderate loads, the actual delays are not nearly as great as (2), but they are much greater than (1). Even at $\rho = 0.5$ (a low traffic intensity), the observed value of $(1-\rho)EL/\tau\rho$ was more than three times greater than predicted by (1).

We derive (2) by proving a heavy-traffic limit theorem, i.e., we establish (2) in the limit as $\rho \to 1$ (as in [19]). Then we use $M/G/1$ structure to conclude for the multiclass batch-Poisson model that (2) is actually *valid for all* ρ. Neither step is especially difficult by itself. The important idea is to combine the heavy-traffic view with the $M/G/1$ view to express EL in terms of the asymptotic variability parameters c_A^2, c_S^2, and $2c_{AS}^2$, and then express these in terms of the basic $(\Sigma M^{B_i}/G_i)/1$ parameters.

As noted above the heavy-traffic limit is valid for a much larger class of models, including quite realistic models of packet queues, as we indicate in Section VI-C. Thus, (2) is *asymptotically correct* as $\rho \to 1$ for a much larger class of models, and has strong implications for packet queue performance under heavy loads. We focus primarily on the more restrictive multiclass batch-Poisson queue because then formula (2) is *valid for all* ρ. This makes the main conclusions about dependence easy to understand.

IV. The $M/G/1$ View

It is well known that the multiclass batch-Poisson queueing model can be analyzed by aggregating all the classes and regarding batches as individual customers. This approach yields a simple $M/G/1$ queue in which the service-time cdf is the appropriate mixture of the service time cdf's for the sum of the service times in each batch.

The random sum of service times in a batch of class i has mean $m_i\tau_i$ and variance $m_i c_{si}^2\tau_i^2 + c_{bi}^2 m_i^2\tau_i^2$; p. 301 of Feller [20]. The mean $\hat{\tau}$ and second moment $[\hat{c}_s^2 + 1]\hat{\tau}^2$ of the total service time of an average batch are then the associated mixtures, i.e.,

$$\hat{\tau}=\sum_{i=1}^{k} p_i m_i \tau_i \text{ and}$$
$$(\hat{c}_s^2+1)\hat{\tau}^2=\sum_{i=1}^{k} p_i[\tau_i^2 m_i c_{si}^2+\tau_i^2 m_i^2(c_{bi}^2+1)].$$

The expected steady-state workload in the queue at an arbitrary time can thus be expressed as

$$\begin{aligned} EL &= \left(\frac{\rho}{1-\rho}\right)\frac{\hat{\tau}(\hat{c}_s^2+1)}{2} \\ &= \left(\frac{\rho}{1-\rho}\right)\left[\frac{\sum_{i=1}^{k} p_i[\tau_i^2 m_i c_{si}^2+\tau_i^2 m_i^2(c_{bi}^2+1)]}{2\sum_{i=1}^{k} p_i\tau_i m_i}\right] \\ &= \left(\frac{\rho\tau}{1-\rho}\right)\left(\frac{\sum_{i=1}^{k} p_i r_i^2[m_i c_{si}^2+m_i^2(c_{bi}^2+1)]}{2m_B}\right) \end{aligned} \tag{5}$$

where $r_i = \tau_i/\tau$.

The final expression in (5) has the cleanest interpretation because three effects have been represented as separable and multiplicative. First, the expected workload is the total amount of service time to be processed, so that it is a time which depends on the measuring units. It is thus useful to think of the dimensionless ratio EL/τ. In other words, EL can be written as τK where K is dimensionless. Second, for this model the effect of the arrival rate or traffic intensity is also captured by the multiplicative factor $\rho/(1-\rho)$. Finally, the remaining factor is the second term in the final expression of (5). It is independent of ρ and τ, i.e., it is independent of the measuring units for time and the total Poisson arrival rate. We use the ratio $r_i = \tau_i/\tau$ to separate the measuring unit effect captured by the overall mean service time τ from the effect of different classes having different mean service times. In the rest of this paper we gain additional insight by finding new interpretations for this final term.

The expected waiting time for the first packet in each batch for any class is also given by (5) since Poisson arrivals see time averages [21]. The expected waiting time for an arbitrary packet of class i, say EW_i, is thus

$$EW_i=EL+\frac{(m_i-1)\tau_i}{2}+\frac{c_{bi}^2 m_i\tau_i}{2} \tag{6}$$

where EL is given in (5); see Section 5.10 of [22] or [23]. The expected waiting time for an arbitrary packet is thus $EW = \Sigma_{i=1}^k q_i EW_i$ where $q_i = \rho_i M_i/\Sigma_{i=1}^k \rho_i M_i$.

From (6) it is clear that the differences between EL, EW_i, and EW are asymptotically negligible as $\rho \to 1$, i.e., all three have the same heavy-traffic limit when multiplied by $(1-\rho)$. Thus, in the following analysis we change back and forth between focusing on EL and EW, depending on what is convenient. However, as an exact result, (5) applies only to the expected workload EL.

V. The General $G/G/1$ View

Our object now is to relate (5) and (2). For this purpose, we focus on individual interarrival times and service times instead of the batches considered in Section IV. Let the individual packets be ordered in the usual way according to their arrival epoch, i.e., ordered lexicographically first according to their arrival epoch and then second according to their position in their batch. Let u_n be the interarrival time between the $(n-1)$st and nth packets and let v_n be the service time of the nth packet. Let $U_n = u_1 + \cdots + u_n$, $n \geq 1$, $U_0 = 0$, $V_n = v_1 + \cdots + v_n$, $n \geq 1$, and $V_0 = 0$. Of course, the difficulty with this view is that the independence in the original $(\Sigma M^{B_i}/G_i)/1$ model specification typically does not provide independence among the random variables in the sequence $\{(u_n, v_n)\}$. We will thus have to resort to heavy-traffic limit theorems to say something useful about this general $G/G/1$ view.

We begin by defining the *stationary-interval* and *asymptotic* variability parameters for an arbitrary sequence of ordered pairs of nonnegative random variables $\{(u_n, v_n): n \geq 1\}$. These are the obvious generalizations of the definitions in [12] for the arrival process alone. We then display the parameter values for the multiclass batch-Poisson queue. The connection between (5) and (2) is expressed by (15) and (17) below (plus the heavy-traffic limit theorems in Section VI).

A. Stationary-Interval Variability Parameters

For the stationary-interval variability parameters, we assume that the sequence $\{(u_n, v_n)\}$ is stationary. Under this assumption, we define the three stationary-interval variability parameters as

$$c_a^2=\frac{\mathrm{Var}\,(u_n)}{(Eu_n)^2},\quad c_s^2=\frac{\mathrm{Var}\,(v_n)}{(Ev_n)^2} \text{ and } c_{as}^2=\frac{\mathrm{Cov}\,(u_n, v_n)}{(Eu_n)(Ev_n)}. \tag{7}$$

Note that the correlation between u_n and v_n, say $\gamma(u_n, v_n)$, is

$$\gamma(u_n, v_n) = \frac{\operatorname{Cov}(u_n, v_n)}{(\operatorname{Var}(u_n)\operatorname{Var}(v_n))^{1/2}} = \frac{c_{as}^2}{\sqrt{c_a^2 c_s^2}}. \tag{8}$$

For our batch-Poisson arrival process, a stationary interarrival time distribution is the mixture of a mass at zero with probability $(m_B - 1)/m_B$ and an exponential cdf having mean λ^{-1} with probability $1/m_B$. Thus, $Eu_n = (\lambda m_B)^{-1} = [\lambda \Sigma_{i=1}^k p_i m_i]^{-1} = \bar{\lambda}^{-1}$, the reciprocal of the total packet arrival rate. The second moment is the corresponding mixture

$$E[(u_n)^2] = \frac{(m_B - 1)}{m_B} 0 + \left(\frac{1}{m_B}\right) \frac{2}{\lambda^2} = \frac{2}{m_B \lambda^2},$$

so that

$$c_a^2 = \frac{\operatorname{Var}(u_n)}{(Eu_n)^2} = 2m_B - 1 \geq 1. \tag{9}$$

A general service time v_n has a cdf which is the mixture of the k individual service cdf's weighted proportionally to the arrival rate of each class, i.e.,

$$\tau = E(v_n) = \frac{\hat{\tau}}{m_B} = \frac{\sum_{i=1}^k p_i m_i \tau_i}{\sum_{i=1}^k p_i m_i} \text{ and}$$

$$(c_s^2 + 1)\tau^2 = \frac{\sum_{i=1}^k p_i m_i (c_{si}^2 + 1)\tau_i^2}{\sum_{i=1}^k p_i m_i}, \tag{10}$$

so that

$$\tau = \sum_{i=1}^k q_i \tau_i \text{ and } c_s^2 = \sum_{i=1}^k q_i (c_{si}^2 + 1) r_i^2 - 1 \geq 0 \tag{11}$$

where $q_i = p_i m_i / \Sigma_{i=1}^k p_i m_i$ is the proportion of the total arrival rate belonging to class i. (Nonnegativity of c_s^2 holds because $\Sigma_{i=1}^k q_i \tau_i^2 \geq [\Sigma_{i=1}^k q_i \tau_i]^2$.)

Finally, we consider the covariance between u_n and v_n (which is the same as the covariance between u_{n+1} and v_n). Recall that the interarrival time refers to the interval between the arrival of the $(n - 1)$st and nth packet. With probability $q_i = (p_i m_i)/\Sigma_{i=1}^k p_i m_i$, v_n is of class i. Conditioned on v_n being of class i, u_n, and u_{n+1} are each (separately) exponential with mean λ^{-1} with probability $1/m_i$ and zero otherwise. (To calculate the probability that the $(n - 1)$st packet is of class i, given that the nth packet is of class i, we use the discrete stationary-excess distribution [23].) Thus,

$$E(u_n v_n) = \sum_{i=1}^k \left(\frac{p_i m_i}{m_B}\right) \frac{1}{m_i} \frac{1}{\lambda} \tau_i = \frac{\sum_{i=1}^k p_i \tau_i}{\lambda m_B}, \tag{12}$$

$$\operatorname{Cov}(u_n, v_n) = E(u_n, v_n) - (Eu_n)(Ev_n) = \frac{\sum_{i=1}^k p_i \tau_i}{\lambda m_B} - \frac{\tau}{\lambda m_B} \tag{13}$$

and

$$c_{as}^2 = \frac{\operatorname{Cov}(u_n, v_n)}{(Eu_n)(Ev_n)} = \sum_{i=1}^k p_i r_i - 1. \tag{14}$$

B. Asymptotic Variability Parameters

Next we define the *asymptotic variability parameters* (which do not require stationarity) as

$$\begin{aligned} c_A^2 &= \lim_{n\to\infty} \bar{c}_{An}^2 = \lim_{n\to\infty} n \frac{\operatorname{Var}(U_n)}{(EU_n)^2} = \lim_{n\to\infty} \frac{\operatorname{Var}(U_n)}{n(Eu_n)^2} \\ c_S^2 &= \lim_{n\to\infty} \bar{c}_{Sn}^2 = \lim_{n\to\infty} n \frac{\operatorname{Var}(V_n)}{(EV_n)^2} = \lim_{n\to\infty} \frac{\operatorname{Var}(V_n)}{n(Ev_n)^2} \\ c_{AS}^2 &= \lim_{n\to\infty} \bar{c}_{ASn}^2 = \lim_{n\to\infty} \frac{n \operatorname{Cov}(U_n, V_n)}{(EU_n)(EV_n)} \\ &= \lim_{n\to\infty} \frac{\operatorname{Cov}(U_n, V_n)}{n(Eu_n)(Ev_n)} = \lim_{n\to\infty} \sqrt{c_A^2 c_S^2}\, \gamma(U_n, V_n) \end{aligned} \tag{15}$$

where U_n and V_n are the partial sums defined in the beginning of Section V. The generalized three-dimensional IDI is the sequence $\{[\bar{c}_{An}^2, \bar{c}_{Sn}^2, \bar{c}_{ASn}^2]: 1 \leq n \leq \infty\}$. (The bars are used in the notation to avoid confusion with the asymptotic variability parameters associated with individual sources; see (17) below.) The stationary-interval variability parameters in (7) appear as the case $n = 1$; the asymptotic variability parameters in (15) appear in the limit as $n \to \infty$.

Paralleling [11], we suggest estimating this IDI in measurements from interarrival-time and service-time data. For example, it is natural to estimate $\operatorname{Var}(U_n)$ via $\operatorname{Var}(U_n) = E(U_n^2) - E(U_n)^2$, using direct sample mean estimates for the moments $E(U_n^2)$ and (EU_n). If $\{(u_n, v_n)\}$ is ergodic as well as stationary then these sample mean estimates are consistent (asymptotically correct as the sample size increases). Confidence intervals for these estimators are harder to obtain because of the dependence. We remark that good estimates typically required very large sample sizes; e.g., see Section III-B of [11]. For an additional suggestion about measurements, see Section VII-B.

We now identify c_A^2, c_S^2, and c_{AS}^2 as the normalization constants in central limit theorems (CLT's) for U_n, V_n and $V_n - U_n$. Let $\Rightarrow$ denote convergence in distribution (weak convergence [24]) and let $N(0, \sigma^2)$ denote a random variable normally distributed with mean 0 and variance σ^2. We assume CLTs hold and identify the asymptotic variability parameters via the normalizing constants as follows:

$$\begin{aligned} n^{-1/2}[U_n - \alpha n] &\Rightarrow N(0, \sigma_A^2), \quad c_A^2 = \sigma_A^2/\alpha^2 \\ n^{-1/2}[V_n - \beta n] &\Rightarrow N(0, \sigma_S^2), \quad c_S^2 = \sigma_S^2/\beta^2 \\ n^{-1/2}[(V_n - U_n) - (\beta - \alpha)n] &\Rightarrow N(0, \sigma_{AS}^2), \quad c_{AS}^2 = (\sigma_{AS}^2 - \sigma_A^2 - \sigma_S^2)/2\alpha\beta. \end{aligned} \tag{16}$$

Of course, in the stationary case $\alpha = Eu_n$ and $\beta = Ev_n$. In general, (15) and (16) are not quite equivalent, but they usually hold together. For example, under an extra uniform integrability condition [24, p. 32], (16) implies (15); see Theorem 20.1 of [24]. For practical purposes, (15) and (16) are equivalent specifications of the asymptotic variability parameters. The reason for two interpretations is that (16) is convenient for our heavy-traffic analysis, whereas (15) is natural for measurements. In fact, the asymptotic-method approximation for arrival and service processes in general queues, as well as for the batch-Poisson model in (2), is asymptotically correct in heavy traffic. For the arrival and service processes separately, this was demonstrated by Theorem 1 of [19] and was discussed for the arrival process alone in [12]. The same result for the arrival and service processes jointly will be justified in Section VI. (We shall actually work with stronger functional central limit theorems (FCLT's), from which (16) follows as an elementary corollary.)

In Section VI-B, we determine asymptotic variability parameters for the multiclass batch-Poisson queue. The following expressions are deduced from Theorem 2. Let c_{Ai}^2 represent the asymptotic variability parameter of the arrival process of class i, which turns out to be $m_i[c_{bi}^2 + 1]$ for the batch-Poisson queue. Then

$$c_A^2 = \sum_{i=1}^{k} q_i m_i (c_{bi}^2 + 1) = \sum_{i=1}^{k} q_i c_{Ai}^2 = m_B(c_B^2 + 1) \geq 1$$

$$c_S^2 = \sum_{i=1}^{k} q_i [r_i^2 c_{si}^2 + (r_i - 1)^2 m_i (c_{bi}^2 + 1)]$$

$$= \sum_{i=1}^{k} q_i [r_i^2 c_{si}^2 + (r_i - 1)^2 c_{Ai}^2] \geq 0$$

$$c_{AS}^2 = \sum_{i=1}^{k} q_i (1 - r_i) m_i (c_{bi}^2 + 1) = \sum_{i=1}^{k} q_i (1 - r_i) c_{Ai}^2. \quad (17)$$

From (17), we clearly see the importance of c_S^2 and c_{AS}^2 as well as c_A^2. In Section VI-C, we generalize these expressions to more realistic models of packet queues, e.g., for models in which packets from the same batch do not arrive at the same instant. We show that (17) remains valid for such queues if the expression for c_{Ai}^2 is appropriately modified. Further insight is provided by examples in Section VII.

VI. Heavy-Traffic Limit Theorems

We now prove heavy-traffic limit theorems yielding asymptotic expressions for the expected equilibrium workload and the expected equilibrium waiting time of an arbitrary packet in the multiclass batch-Poisson queue and more general models. For the multiclass batch-Poisson mode, we obtain these limits by letting the total Poisson arrival rate λ increase so that $\rho \to 1$, while holding the other parameters fixed. From (5) and (6), we see that the expected workload EL and the expected waiting time of an arbitrary packet in class i, EW_i, coincide in the limit as λ increases so that $\rho \to 1$, i.e., the extra terms in EW_i in (6) are asymptotically negligible as $\rho \to 1$ after multiplying by $(1 - \rho)$. Thus, the heavy-traffic limiting behavior of EL, EW_i, and $EW = \Sigma_{i=1}^{n} q_i EW_i$ are all identical. What we want to show, then, is that $(1 - \rho)EL \to \tau[c_A^2 + c_S^2 - 2c_{AS}^2]/2$ as $\rho \to 1$ via changing the arrival rate. By (5) we know that $EL = \bar{K}\rho/(1 - \rho)$ for all ρ where $\bar{K}$ is independent of ρ. This limit together with (5) thus will establish (2) for the workload, i.e., necessarily $\bar{K} = \tau[c_A^2 + c_S^2 - 2c_{AS}^2]/2$.

A. The General G/G/1 Model in Heavy Traffic

We first establish a general sufficient condition for a heavy-traffic limit theorem for the waiting times in a general $G/G/1$ model with interarrival-time and service-time sequence $\{(u_n, v_n)\}$. (No independence or common distribution assumptions will be in force here.) Paralleling Theorem 1 of [19], the condition is in terms of a functional central limit theorem (FCLT). See [24]–[26] for additional background.

The sequence of successive waiting times $\{W_n: n \geq 0\}$ can be defined in terms of the basic sequence $\{(u_n, v_n): n \geq 1\}$, assuming the initial workload (at the arrival epoch of the first packet) is $W_1 = 0$, by

$$W_{n+1} = \max\{W_n + v_n - u_{n+1}, 0\}$$
$$= D_n - \min\{D_j : 0 \leq j \leq n\}, \; n \geq 1 \quad (18)$$

where $d_n = v_n - u_{n+1}$, $D_n = d_1 + \cdots + d_n$, $n \geq 1$, and $D_0 = 0$.

We remark that if the sequence $\{(u_n, v_n)\}$ is stationary (*without* assuming any independence) with $\rho < 1$ and $W_1 = 0$ (or another technical condition is imposed), then W_n converges in distribution to a proper limiting random variable W as $n \to \infty$; [27] or Chapter 1 of [28].

For the FCLT, we consider a sequence of queueing systems indexed by the superscript n. The nth queueing system is characterized by the sequence $\{[u_j^n, v_j^n]: j \geq 1\}$. We will establishing convergence as $n \to \infty$. Let $(\hat{U}_n, \hat{V}_n)$ be the random element of the product function space $D \times D$ where D is the function space $D[0, \infty)$, defined by

$$[\hat{U}_n(t), \hat{V}_n(t)] = (n^{-1/2}[U_{[nt]}^n - \alpha_n nt],$$
$$n^{-1/2}[V_{[nt]}^n - \beta_n nt]), \; t \geq 0, \quad (19)$$

$U_j^n = u_2^n + \cdots + u_{j+1}^n$, $V_j^n = v_1^n + \cdots + v_j^n$ and $[x]$ is the greatest integer less than or equal to x. When $\{(u_j^n v_j^n): n \geq 1\}$ is stationary, $\alpha_n = Eu_j^n$ and $\beta_n = Ev_j^n$. Let $\hat{D}_n$ and $\hat{W}_n$ be the random elements induced by the differences and the waiting times in (18), defined by

$$\hat{D}_n(t) = n^{-1/2} D_{[nt]}^n \text{ and } \hat{W}_n(t) = n^{-1/2} W_{[nt]}^n, \; t \geq 0. \quad (20)$$

Let $f: D \to D$ be the function corresponding to the impenetrable barrier of the origin, i.e., $f(x)(t) = x(t) - \inf\{x(s): 0 \leq s \leq t\}$, $t \geq 0$. Let $e(t) = t$, $t \geq 0$, and let $B(t)$ be standard Brownian motion (BM) with zero drift and unit variance.

The following result is a variant of results in [19]; see Sections 3.4 and 4.3 of [26].

Theorem 1: a) If $[\hat{U}_n, \hat{V}_n] \Rightarrow (\hat{U}, \hat{V})$ in $D \times D$ where $(\hat{U}, \hat{V})$ has continuous paths w.p.1 and $n^{1/2}[\alpha_n - \beta_n] \to \alpha$, $-\infty < \alpha < \infty$, as $n \to \infty$, then $\hat{D}_n \Rightarrow \hat{D} = \hat{V} - \hat{U} - \alpha e$ in D;

b) If $\hat{D}_n \Rightarrow \hat{D}$ in D, then $\hat{W}_n \Rightarrow \hat{W} = f(\hat{D})$ in D;

c) If, in addition to a) above, $\alpha > 0$ and the limit $(\hat{U}, \hat{V})$ is two-dimensional Brownian motion with covariance matrix having elements σ_{11}^2, σ_{22}^2, and $\sigma_{12}^2 = \sigma_{21}^2$, then $\hat{W} = f(\hat{V} - \hat{U} - \alpha e)$ is regulated or reflecting Brownian motion (RBM) with drift $-\alpha$, which has an exponential equilibrium distribution with mean

$$E\hat{W}(\infty) = (\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2)/2\alpha.$$

Proof: Parts a) and b) are consequences of the continuous mapping theorem; Theorem 5.1 of [24]. Note that $\hat{W}_n = f[\hat{V}_n - \hat{U}_n - \hat{E}_n]$ where $\hat{E}_n(t) = n^{-1/2}[\alpha_n - \beta_n][nt]$. By assumption and Theorem 4.4 of [24], $[\hat{V}_n, \hat{U}_n, \hat{E}_n] \Rightarrow [\hat{V}, \hat{U}, \alpha e]$. Then apply addition, subtraction and the barrier function f, [25, Sections 4 and 6]. For part c), first it is elementary that $\hat{V} - \hat{U} - \alpha e$ is a BM with drift coefficient $-\alpha$ and diffusion coefficient $\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2$. Thus, $f(\hat{V} - \hat{U} - \alpha e)$ is RBM with these parameters. If RBM has negative drift, then it has an exponential equilibrium distribution. ■

Remark: Note that it is theoretically possible to have a heavy-traffic limit in Theorem 1 under b) without a).

B. The Heavy-Traffic Limit for the Multiclass Batch-Poisson Queue

Our object now is to show that the conditions of Theorem 1 are indeed satisfied for our multiclass batch-Poisson queue and that the mean of the equilibrium distribution satisfies $E\hat{W}(\infty) = \tau[c_A^2 + c_S^2 - 2c_{AS}^2]/2$, as needed to justify (2). (Recall that the workload process has the same heavy-traffic limit as the waiting times.)

For simplicity, we construct our sequence of queueing systems by changing the Poisson rate λ alone, leaving p_i, τ_i and the other variables unchanged. (The limits would also hold under other schemes.) We thus let λ depend on n. Then we let $1 - \rho_n = n^{-1/2}$, so that $1 - \lambda_n \tau m_B = n^{-1/2}$ and $\lambda_n = (m_B \tau)^{-1} (1 - n^{-1/2})$. Thus, the translation constants in (19) are $\beta_n = \tau$ and $\alpha_n = (\lambda_n m_B)^{-1} = \tau/(1 - n^{-1/2})$ for all $n \geq 1$.

Theorem 2: For the sequence of multiclass batch-Poisson queues above, $n^{1/2} (\alpha_n, \beta_n) = n^{1/2} (\alpha_n, - \tau) \rightarrow \tau$ and $[\hat{U}_n, \hat{V}_n] \Rightarrow [\hat{U}, \hat{V}]$ as $n \rightarrow \infty$ where $(\hat{U}, \hat{V})$ is two-dimensional Brownian motion without drift having covariance elements

$$\sigma_{11}^2 = \frac{\tau^3}{\hat{\tau}} \sum_{i=1}^{k} p_i m_i (c_{bi}^2 + 1) = \tau^2 c_A^2$$

$$\sigma_{22}^2 = \frac{\tau}{\hat{\tau}} \left(\sum_{i=1}^{k} p_i m_i \tau_i^2 c_{si}^2 + \sum_{i=1}^{k} p_i m_i^2 (c_{bi}^2 + 1)(\tau_i - \tau)^2 \right) = \tau^2 c_S^2$$

$$\sigma_{12}^2 = \sigma_{21}^2 = \frac{\tau^2}{\hat{\tau}} \left(\sum_{i=1}^{k} p_i m_i^2 (c_{bi}^2 + 1)(\tau - \tau_i) \right) = \tau^2 c_{AS}^2 \quad (21)$$

for c_A^2, c_S^2, and c_{AS}^2 in (17), so that $\hat{V} - \hat{U}$ is BM with variance coefficient

$$\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{12}^2 = \left(\frac{\tau}{\hat{\tau}} \right) \sum_{i=1}^{k} [p_i m_i \tau_i^2 c_{si}^2 + p_i m_i^2 (c_{bi}^2 + 1) \tau_i^2]$$

$$= \tau^2 (c_A^2 + c_S^2 - 2c_{AS}^2). \quad (22)$$

Since the proof of Theorem 2 is relatively long and detailed, it appears in Appendix A. Here we discuss its consequences. First, the variability parameters in (21) and (22) differ from those in (17) by the factor τ^2, as they should by (16). Combining Theorems 1 and 2, we obtain the following heavy-traffic limit for the waiting times in the mutliclass batch-Poisson queue.

Corollary 1: For the sequence of multiclass batch-Poisson queues above $\hat{W}_n \Rightarrow \hat{W}$ where $\hat{W}$ is RBM with drift $-\tau$, variance coefficient $\tau^2 [c_A^2 + c_S^2 - 2c_{AS}^2]$ and $E\hat{W}(\infty) = \tau[c_A^2 + c_S^2 - 2c_{AS}^2]/2$.

In Theorems 1 and 2, we have not directly shown that the sequence of normalized equilibrium waiting times $\{n^{-1/2} W_\infty^n : n \geq 1\} = \{(1 - \rho_n) W_\infty^n : n \geq 1\}$ or the associated sequence of expectations converge and, given that they do, that the limit coincides with the expected equilibrium value of RBM, $E\hat{W}(\infty)$. However, these important steps are covered by previous heavy-traffic limit theorems in the $M/G/1$ setting of Section IV. From previous heavy-traffic limit theorems in the $M/G/1$ setting [29], [30], [19], we know that these limits exist and coincide.

Corollary 2: For the sequence of multiclass batch-Poisson queues above,

$$(1 - \rho_n) W_\infty^n \Rightarrow \hat{W}(\infty) \text{ in } R \text{ as } n \rightarrow \infty$$

and

$$\lim_{n \rightarrow \infty} (1 - \rho_n) E W_\infty^n = E\hat{W}(\infty) = \tau (c_A^2 + c_S^2 - 2c_{AS}^2)/2.$$

The heavy-traffic limits so far have been for the waiting times. Corresponding results hold for the workload. Let $L^n(t)$ be the workload at time t in the nth queueing system and let $\hat{L}_n(t)$ be the associated random element of D, defined by $\hat{L}_n(t) = n^{-1/2} L_{[nt]}^n$, $t \geq 0$.

Theorem 3: For the sequence of multi-class batch-Poisson queues above, $\hat{L}_n \Rightarrow \hat{W}$, $(1 - \rho_n) L^n(\infty) \Rightarrow \hat{W}(\infty)$ and $(1 - \rho_n)\, EL^n(\infty) \rightarrow E\hat{W}(\infty)$ for $\hat{W}$ in Corollary 1.

Theorem 3 is an easy consequence of Theorem 2, but we also prove it directly together with Theorem 2 in Appendix A.

C. Heavy-Traffic Limits for More General Multiclass Queues

Theorem 1 in Section VI-A was established for very general $G/G/1$ models, whereas Theorem 2 and its corollaries in Section VI-B were established only for the special case of the multiclass batch-Poisson queue. In fact, Theorem 2 is easily extended to a much larger class of models. For these more general models, 2) is valid asymptotically as $\rho \rightarrow 1$ but not for each ρ.

1) Batch Renewal Processes: A natural generalization is to allow the arrival process of batches for class i to be a renewal process with variability parameter c_{ri}^2 instead of a Poisson process (with variability parameter $c_{ri}^2 = 1$). A further generalization is just to assume that each of these arrival processes of batches satisfies a FCLT. In particular, let $B^i(t)$ denote the arrival process of batches for class i and let $\hat{B}_n^i$ be the associated random element of D, defined by

$$\hat{B}_n^i(t) = n^{-1/2} [B^i(nt) - \lambda_n p_i nt], \; t \geq 0. \quad (23)$$

Note that $\hat{B}_n^i$ is similar to $\hat{A}^i$ in (A-1) of Appendix A; the translation term in (A-1) has an extra m_i, because $A^i(t)$ is counting packets instead of batches. If $\hat{B}_n^i \Rightarrow \hat{B}^i$ where $\hat{B}^i$ is *BM* with zero drift and variance coefficient $\hat{\tau}^{-1} p_i c_{ri}^2$, then $\hat{A}_n^i \Rightarrow \hat{A}^i$ where $\hat{A}^i$ is *BM* with zero drift and variance coefficient $\hat{\tau}^{-1} p_i m_i^2 [c_{bi}^2 + c_{ri}^2]$, by the same argument used to treat the batch-Poisson process (Section 6 of [25]). In other words, with this generalization, Theorem 2 remains valid with the simple modification that $[c_{bi}^2 + c_{ri}^2]$ replaces $[c_{bi}^2 + 1]$ in σ_{11}^2, σ_{22}^2, and σ_{12}^2 in (21). As stated above, then (2) is valid for this model asymptotically as $\rho \rightarrow 1$, with $c_{Ai}^2 = m_i [c_{bi}^2 + c_{ri}^2]$ replacing $c_{Ai}^2 = m_i [c_{bi}^2 + 1]$ in c_A^2, c_S^2, and c_{AS}^2 in (17).

2) Spacing Between Packets: For more realistic packet queue models, it is also significant that Theorem 2 remains valid if all the packets in a batch do not arrive together at one instant. As a first elementary case, suppose that packets from class i batches with parameters m_i and c_{bi}^2 occur somewhere in the interval between successive arrivals of a basic batch arrival process $B^i(t)$ as above where $B^i(t)$ satisfies the FCLT above. Then the FCLT for $A^i(t)$ is just as above with $[c_{bi}^2 + c_{ri}^2]$ replacing $[c_{bi}^2 + 1]$. In other words, the precise location of the arrivals within the interval does not affect the heavy-traffic limit.

However, both the original model and the generalization above implicitly assume that each batch size is independent of the lengths of the interval between batch arrivals. This clearly is not realistic for packet queues, where longer messages (larger batch sizes) usually entail longer intervals between message (batch) arrivals. It is not difficult to create models which represent this feature. In particular, let $\{(B_n^i, L_n^i) : n \geq 1\}$ be the sequence of pairs of successive batch sizes and interval lengths between batch arrivals for class i. Assume that successive pairs in this sequence are i.i.d., but allow B_n^i and L_n^i to be dependent for each n. As above, let m_i and c_{bi}^2 be the two parameters for B_n^i and let $(\lambda p_i)^{-1}$ and c_{ri}^2 be the two parameters for L_n^i. Let γ_{bri} be the correlation between B_n^i and L_n^i. Again Section V of [25] can be applied to establish the necessary FCLT for $\hat{A}_n^i$ cf. Appendix A. Now Theorem 2 and (17) are valid with

$$c_{Ai}^2 = m_i (c_{bi}^2 + c_{ri}^2 - 2\gamma_{bri} c_{bi} c_{ri}) \quad (24)$$

replacing $c_{Ai}^2 = m_i [c_{bi}^2 + 1]$, i.e., with $c_{ri}^2 - 2\gamma_{bri} c_{bi} c_{ri}$ replacing 1.

Of particular interest for packet queues is a more detailed model in which each batch interval L^i for class i is divided into two independent components, an idle period I^i and a busy period, where the busy period is the sum of B^i i.i.d. spaces T_j^i, so that the total interval length can be expressed as

$$L^i = I^i + \sum_{j=1}^{B^i} T_j^i \quad (25)$$

where $\{T_j^i : j \geq 1\}$ is i.i.d for each i. (Note that the arrival rate of batches of class i must satisfy $\lambda p_i = 1/(m_i E T_1^i + E I^i)$ with the associated arrival rate of packets being $\bar{\lambda} q_i = \lambda p_i m_i$.) In this more detailed model, let β_i be the proportion of busy time, i.e., $\beta_i = (EB^i)(ET^i)/EL^i$. Let c_{bi}^2, c_{Ti}^2, and c_{Ii}^2 be the squared coefficients of variation of B^i, T^i, and I^i, respectively, for the ith class. It is easy to see that in this case (24) becomes

$$c_{Ai}^2 = m_i(1-\beta_i)^2(c_{bi}^2 + c_{Ii}^2) + \beta_i^2 c_{Ti}^2. \tag{26}$$

The batch arrival case occurs in the limit as $\beta_i \to 0$. Then $c_{Ii}^2 \to c_{ri}^2$.

Formula (26) clearly depicts the effect of a batch assumption (no spacing) on the asymptotic variability parameters. As an illustration, for the X.25 model yielding the data in (3) and (4), β_i assumed values 0.01173 and 0.0484 for the two kinds of sources considered and, in both cases, $c_{Ti}^2 = 0$. Then (26) changes little when we simply set $\beta_i = 0$.

VII. Examples

In this section, we briefly examine two special cases of the multiclass batch-Poisson model. For this model, (2) and (17) are valid for all ρ.

A. Common Service Times

We first consider the special case in which all the individual service times have the same means and variances, so that $\tau_i = \tau$ and $c_{si}^2 = c_s^2$ for all i. (Consequently, $c_s^2 = c_S^2$ in this case.) In this case,

$$\hat{\tau} = \tau m_B \text{ and } (\hat{c}_s^2 + 1)\hat{\tau}^2 = \hat{\tau}^2[m_B c_s^2 + (c_B^2 + 1)m_B^2],$$

so that from (5)

$$EL = \frac{\tau\rho}{1-\rho}\frac{(c_s^2 + c_A^2)}{2} \quad \text{where } c_A^2 = m_B(c_B^2 + 1). \tag{27}$$

As a basic consistency check, note that when all batch sizes are size one, then $m_B = 1$, $c_B^2 = 0$, $c_A^2 = 1$ and (27) reduces to the familiar Pollaczek-Khintchine formula, p. 189 of [22].

In this special case, it is not difficult to relate (27) to (2). In particular, in this case (27) coincides with (2) where $c_A^2 = [c_B^2 + 1]m_B$, $c_S^2 = c_s^2$ and $c_{AS}^2 = 0$. As a consequence, we see that in order for all three dependence effects to occur it is necessary for the service characteristics of the different classes to be different. In this case, dependence only appears in the arrival process.

B. Deterministic Batches and Service Times

An interesting case occurs when there is no variability in the batches and service times for each class, i.e., $c_{bi}^2 = c_{si}^2 = 0$ for all i. (Having $c_{si}^2 = 0$ corresponding to deterministic service times is often realistic for packet queues, but typically $c_{bi}^2 > 0$.) Then the formulas simplify to

$$c_A^2 = \sum_{i=1}^{k} q_i m_i, \quad c_S^2 = \sum_{i=1}^{k} q_i m_i (1-r_i)^2, \quad c_{AS}^2 = \sum_{i=1}^{k} q_i(1-r_i)m_i$$

and

$$(c_A^2 + c_S^2 - 2c_{AS}^2) = \sum_{i=1}^{k} q_i m_i r_i^2. \tag{28}$$

To show some extreme dependence effects, we now consider limits for (28) when there are only two classes. Let $m_2 = 1$. Let $p_1 \to 0$ and $m_1 \to \infty$ so that $q_1 m_1 = p_1 m_1^2/(p_1 m_1 + 1 - p_1) \to x$. Then

$$c_A^2 = q_1 m_1 + (1-q_1) \to x + 1$$

$$c_S^2 = q_1 m_1 (1-r_1)^2 + (1-q_1)(1-r_2)^2 \to x(1-r_1)^2 + (1-r_2)^2$$

$$c_{AS}^2 = q_1 m_1 (1-r_1) + (1-q_1)(1-r_2) \to x(1-r_1) + (1-r_2)$$

$$(c_A^2 + c_S^2 - 2c_{AS}^2) = q_1 m_1 r_1^2 + (1-q_1) r_2^2 \to x r_1^2 + r_2^2. \tag{29}$$

Note that we still have not specified τ_1 and τ_2 or, equivalently, r_1 and r_2. Choose τ_1 and τ_2, so that $r_1 \to 0$, and $r_2 \to 1$. Then

$$c_A^2 \to x + 1, \; c_S^2 \to x, \; c_{AS}^2 \to x \text{ and } (c_A^2 + c_S^2 - 2c_{AS}^2) \to 1. \tag{30}$$

With this limiting scheme, the queue thus behaves like an $M/D/1$ queue even though c_A^2, c_S^2, and c_{AS}^2 can be arbitrarily large. A numerical example is $m_1 = 1000$, $\tau_1 = 0.001$, $p_1 = 0.0001$, $m_2 = 1$, and $\tau_2 = 1000$. Then $\tau = 909$, $c_A^2 = 91.8$, $c_S^2 = 90.9$, $c_{AS}^2 = 90.8$, $[c_A^2 + c_S^2 - 2c_{AS}^2] = 1.1$. The first class seriously affects the asymptotic variability parameters, but does not seriously affect the queue. The three separate components c_A^2, c_S^2, and c_{AS}^2 can be very large, without the combined asymptotic variability parameter $c_A^2 + c_S^2 - 2c_{AS}^2$ being large.

For the general $G/G/1$ queue as well as for the batch-Poisson special case, we can avoid difficulties in combining estimates for c_A^2, c_S^2, and c_{AS}^2 by measuring the sequence $\{d_n\} = \{v_n - u_{n+1}\}$, i.e., the sequence of differences between service times and interarrival times, instead of $\{(u_{n+1}, v_n)\}$, which was defined at the beginning of Section V. This approach is suggested by the fact that the sequence of waiting times $\{W_n\}$ defined in (18) depends on the basic sequence $\{(u_n, v_n)\}$ only through the sequence of partial sums of the differences $\{D_n\}$. In particular, we propose estimating the variance-time curve $\{\sigma_k^2 : k \geq 1\}$ associated with the partial sums D_n of (18), defined by

$$\sigma_k^2 = k^{-1} \operatorname{Var}(D_k), \quad k \geq 1. \tag{31}$$

Paralleling the IDI's in (15) and [18], σ_1^2 is the stationary-interval variance and $\sigma_\infty^2 = \lim_{k\to\infty} k^{-1} \operatorname{Var}(D_k)$ is the asymptotic variance. The kth term σ_k^2 includes the covariances among the first k differences.

For describing the workload in our batch-Poisson setting, it is convenient to measure $\hat{d}_k = [[v_k/Ev_k] - [u_{k+1}/Eu_{k+1}]]$ instead of d_k and estimate

$$\hat{\sigma}_k^2 = k^{-1} \operatorname{Var}(\hat{D}_k), \quad k \geq 1 \tag{32}$$

where $\hat{D}_k = \hat{d}_1 + \cdots + \hat{d}_k$, $k \geq 1$. Note that

$$\sigma_1^2 = (Eu_1)^2 c_a^2 + (Ev_1)^2 c_s^2 - 2(Eu_1)(Ev_1)c_{as}^2,$$

$$\sigma_\infty^2 = (Eu_1)^2 c_A^2 + (Ev_1)^2 c_S^2 - 2(Eu_1)(Ev_1)c_{AS}^2 \tag{33}$$

while

$$\hat{\sigma}_1^2 = c_a^2 + c_s^2 - 2c_{as}^2$$

$$\hat{\sigma}_\infty^2 = c_A^2 + c_S^2 - 2c_{AS}^2; \tag{34}$$

see (7) and (15). Hence, an estimate for (32) with large k directly yields an estimate for the variability parameter $[c_A^2 + c_S^2 + 2c_{AS}^2]/2$ in (2), whereas (31) does not. The distinction between σ_k^2 and $\hat{\sigma}_k^2$ disappears as $\rho \to 1$, because then $Eu_1 - Ev_1 \to 0$, so that Eu_1 can be factored out. However, $\hat{\sigma}_\infty^2$ directly yields the key composite parameter for the workload for all ρ in the batch-Poisson special case. On the other hand, σ_k^2 is apparently more appropriate for waiting times, in general $G/G/1$ models as well as in the batch-Poisson special case.

VIII. Conclusions

In order to understand complex queueing systems and develop useful approximations for performance measures, it is helpful to consider special limiting cases such as heavy traffic ($\rho \to 1$) for which the behavior can be analyzed exactly. In this paper we have consider *two limiting cases* of relevance to packet queues. First, we considered the multiclass batch-Poisson model, which differs from a realistic model of a packet queue primarily because the burstiness is exaggerated by assuming that all the packets associated with each message arrive at the transmission queue at one instant. This special multiclass batch-Poisson model has appeal because we can analyze it quite completely. Moreover, the analysis clearly reveals the important dependence effects. The analysis obviously has relevance for packet queues when the actual epochs of packets associated with each message are not too spread out.

Second, in Section VI-C, we considered more appropriate models for packet queues which include the spacing between the packets of a message. From heavy traffic analysis, we see that the serious effect of burstiness in multiclass batch-Poisson models will also occur in these models of packet queues under heavy loads. In particular, as is illustrated by the data in (3) and (4), the heavy-traffic limits for more appropriate packet queue models (without flow control mechanisms) and corresponding multiclass batch-Poisson models often do not differ greatly. [This is not difficult to understand because any fixed interarrival times for packets associated with the same message become smaller relative to the relevant time scale for the queue as the traffic intensity ρ approaches its upper limit for stability ($\rho = 1$)]. Moreover, in all these models there is a dramatic increase in packet delays relative to the $M/M/1$ model as ρ increases. For the example in (3) and (4), ignoring the burstiness (acting as if $c_A^2 + c_S^2 - 2c_{AS}^2 = 1$) leads to an error by a factor of 30. The results here have important implications for measurements of real or simulated packet queues.

In (15), (31), and (32), we have suggested useful quantities to estimate from interarrival-time and service-time data in order to gain insight into variability. For example, we can gain important insight into the way flow control mechanisms alter variability by seeing how the variability parameters change with ρ. We can compare the variability of different arrival processes in a network (e.g., at different queues) and see how system structure alters the variability. Thus, the analysis here provides a useful perspective for understanding packet queues and other complex queueing systems. (Also see [39].)

As observed in [11]–[16], heavy-traffic approximations can be quite inaccurate under light-to-moderate loads. The value of direct applications of heavy-traffic approximations such as [19], [26], [29]–[32] therefore has limitations. Heavy-traffic approximations and related special cases can nevertheless play an important role in approximating performance measures by serving as part of a more complicated and comprehensive heuristic. This is the approach underlying [8], [11]–[15], [33], [34] and forthcoming papers by M. Reiman and B. Simon. These heuristics typically interpolate between formulas that are good in light and heavy traffic.

The results in this paper are also important because they reveal limitations in procedures for approximating queueing networks based on aggregation and decomposition. For example, the queueing network analyzer (QNA) [8] aggregates all customer classes and decomposes the network, treating the separate queues as mutually independent and the service times at each queue as independent of the arrival processes there. The procedure in [8] was designed to capture the effect of bursty arrival processes, but the version in [8] does not treat the other two kinds of dependence involving the service times. The analysis here show that the resulting error can be substantial. The limitations of algorithms based on aggregation has also recently ben observed by Bitran and Tirupati [35] in connection with multiclass networks having deterministic routing, which arise in manufacturing. They also propose an algorithmic improvement to address the problem in that context.

Appendix A

Proof of Theorems 2 and 3

The proof of Theorems 2 and 3 is a relatively straightforward application of [25]. (In fact, these results serve as an excellent motivation for [25].) Let $A^i(t)$ be the batch-Poisson arrival process for class i and let $\bar{A}_n^i$ and $\hat{A}_n^i$ be associated random elements of D when the Poisson arrival rate is $\hat{\tau}^{-1}(1 - n^{-1/2})$, as it has been assumed to be in the nth system, defined by

$$\bar{A}_n^i(t) = n^{-1/2}[A^i(nt) - \hat{\tau}^{-1}(1 - n^{-1/2})p_i m_i nt], \; t \geq 0$$

$$\hat{A}_n^i(t) = n^{-1/2}[A^i(nt) - \hat{\tau}^{-1}p_i m_i nt], \; t \geq 0. \qquad \text{(A-1)}$$

Since the batch-Poisson process can be viewed as a random sum, we can apply Section V of [25] to conclude that $\bar{A}_n^i \Rightarrow \bar{A}^i$ where $\bar{A}^i$ is BM with zero drift and variance $\hat{\tau}^{-1} p_i m_i^2(c_{bi}^2 + 1)$. Since $\bar{A}_n^i(t) - \hat{A}_n^i(t) = \hat{\tau}^{-1} p_i m_i t$, $\hat{A}_n^i \Rightarrow \hat{A}^i = \bar{A}_i - \alpha_i e$ where $\alpha_i = \hat{\tau}^{-1} p_i m_i$. (This step generalizes, as indicated in Section VI-C.)

Let V_n^i be the nth partial sum of the service times of class i and let $\hat{V}_n^i$ be the associated random element of D defined by

$$\hat{V}_n^i(t) = n^{-1/2}[V_{[nt]}^i - \tau_i nt], \; t \geq 0. \qquad \text{(A-2)}$$

Since the service times for each class are i.i.d., we can apply Donsker's theorem [24] to obtain $\hat{V}_n^i \Rightarrow \hat{V}^i$ where $\hat{V}^i$ is BM with zero drift and variance $\tau_i^2 c_{si}^2$.

Let $S^i(t) = V_{A^i(t)}^i$ represent the total input of work of class i in time t, and let $\hat{S}_n^i$ be the associated random element of D when the Poisson arrival rate is $\hat{\tau}(1 - n^{-1/2})$, defined by

$$\hat{S}_n^i(t) = n^{-1/2}[V_{A^i(nt)}^i - \tau_i \hat{\tau}^{-1} p_i m_i nt], \; t \geq 0. \qquad \text{(A-3)}$$

By Section V of [25] again,

$$\begin{aligned}\hat{S}_n^i \Rightarrow \hat{S}^i &= \hat{V}^i \circ \hat{\tau}^{-1} p_i m_i e + \tau_i \hat{A}^i \\ &= \hat{V}^i \circ \hat{\tau}^{-1} p_i m_i e + \tau_i \bar{A}^i - \hat{\tau}^{-1} p_i m_i \tau_i e\end{aligned}$$

where $\circ$ is the composition map, i.e., $(x \circ y)(t) = x(y(t))$, $t \geq 0$, so that $\hat{S}^i$ is BM with drift coefficient $-p_i m_i \tau_i / \hat{\tau}$ and variance coefficient $[p_i m_i \tau_i^2 c_{si}^2 + p_i m_i^2 (c_{bi}^2 + 1)\tau_i^2]/\hat{\tau}$. Next let $S(t) = S^1(t) + \cdots + S^k(t)$ and

$$\hat{S}_n(t) = n^{-1/2}[S(nt) - nt], \; t \geq 0. \qquad \text{(A-4)}$$

Since the processes $S^i(t)$ are independent, we have $[\hat{S}_n^1, \cdots, \hat{S}_n^k] \Rightarrow [\hat{S}^1, \cdots, \hat{S}^k]$ in D^k and $\hat{S}_n \Rightarrow \hat{S} = \hat{S}^1 + \cdots + \hat{S}^k$. The limit process $\hat{S}$ is BM with drift coefficient -1 and variance coefficient $\Sigma_{i=1}^k [p_i m_i \tau_i^2 c_{si}^2 + p_i m_i^2 (c_{bi}^2 + 1)\tau_i^2]/\hat{\tau}$.

Let $N(t) = S(t) - t$, $t \geq 0$ represent the net input process to the queue, and let $\hat{N}_n = n^{-1/2} N^n(nt)$ for the nth system, when the Poisson arrival rate is $\hat{\tau}(1 - n^{-1/2})$. Obviously, $\hat{N}_n = \hat{S}_n$ so that $\hat{N}_n \Rightarrow \hat{S}$ too. Let $L(t)$ represent the workload at time t and let $\hat{L}_n = n^{-1/2} L^n(nt)$. Since $L(t) = f(N)(t)$ and f is a continuous map. Theorem 6.4 of [25], $\hat{L}_n \Rightarrow \hat{W} = f(\hat{S})$ where $\mathrm{f}(\hat{S})$ is RBM with drift coefficient -1 and variance coefficient $\Sigma_{i=1}^k [p_i m_i \tau_i^2 c_{si}^2 + p_i m_i^2 (c_{bi}^2 + 1)\tau_i^2]/\hat{\tau}$, and exponential limiting distribution having mean $E\hat{W}(\infty)$ as in Corollary 1. At this point we have completed a direct proof of Theorem 3.

The sequence $\{U_n\}$ of all successive arrival times is the inverse process associated with $A(t) = A^1(t) + \cdots + A^k(t)$; Section 7 of [25], [36] and Section 3 of [37]. Let $\hat{A}_n = \hat{A}_n^1 + \cdots + \hat{A}_n^k$, $\bar{A}_n = \bar{A}_n^1 + \cdots + \bar{A}_n^k$ and let $\bar{U}_n$ be the random function

$$\bar{U}_n(t) = n^{-1/2}[U_{[nt]}^n - \tau nt], \; t \geq 0, \qquad \text{(A-5)}$$

differing from $\hat{U}_n$ in (19) only by the translation term. Since

$$\hat{A}_n(t) = n^{-1/2}[A(nt) - \tau^{-1} nt], \; t \geq 0, \qquad \text{(A-6)}$$

and $\hat{A}_n \Rightarrow \hat{A} = \hat{A}^1 + \cdots + \hat{A}^k$, we can apply Section 7 of [25], especially the Corollary to Lemma 7.6, to obtain $\hat{U}_n \Rightarrow \hat{U} = -\tau \bar{A} \circ \tau e$ and $\bar{U}_n \Rightarrow \bar{U} = -\tau \bar{A} \circ \tau e + \tau e$. Hence, $\bar{U}$ is *BM* with drift coefficient $+\tau$ and variance component σ_{11}^2 as displayed in (21).

The sequence of service times $\{V_n\}$ ordered the same way as $\{U_n\}$ can be obtained by observing that $S(t) = V_{A(t)}$, so that we can obtain V_n by undoing (inverting) the composition map; p. 76 of [25] and [38]. Let

$$\hat{V}_n(t)=n^{-1/2}[V^n_{[nt]}-\tau nt],\ t\geq 0,$$

$$V^*_n(t)=\hat{S}_n(t)-\tau\hat{A}_n(t)=n^{-1/2}[S^n(nt)-\tau A(nt)]$$
$$=n^{-1/2}[V_{A(nt)}-\tau A(nt)]=\hat{V}_n(\hat{\Phi}_n(t)),\quad t\geq 0 \qquad \text{(A-7)}$$

where $\hat{\Phi}_n(t) = \tau n^{-1}A(nt),\ t \geq 0$.

To establish $V^*_n \Rightarrow V^*$ and the final joint convergence $[\hat{U}_n, \hat{V}_n] \Rightarrow (\hat{U}, \hat{V})$, we need to do all the limits jointly from the beginning. We start with

$$\hat{Z}_n=(\hat{A}^1_n, \cdots, \hat{A}^k_n, \hat{V}^1_n, \cdots, \hat{V}^k_n)$$
$$\Rightarrow \hat{Z}=(\hat{A}^1, \cdots, \hat{A}^k, \hat{V}^1, \cdots, \hat{V}^k) \text{ in } D^{2k}$$

which is valid because all the marginal processes are independent. Then we apply the previous arguments to obtain

$$\bar{Y}_n=(\hat{Z}_n, \hat{S}^1_n, \cdots, \hat{S}^k_n, \hat{S}_n, \hat{A}_n, \hat{U}_n)$$
$$\Rightarrow \hat{Y}=(\hat{Z}, \hat{S}^1, \cdots, \hat{S}^k, \hat{S}, \hat{A}, \hat{U}) \text{ in } D^{3k+3}. \qquad \text{(A-8)}$$

We now treat the inverse of composition. First by Theorems 4.4 and 5.1 of [24],

$$(\bar{Y}_n, V^*_n, \hat{\Phi}_n) \Rightarrow (\hat{Y}, \hat{S}-\tau\hat{A}, e) \text{ in } D^{3k+5}. \qquad \text{(A-9)}$$

Then

$$(\bar{Y}_n, V^*_n, \hat{\Phi}_n, \hat{V}_n) \Rightarrow (\hat{Y}, \hat{S}-\tau\hat{A}, e, \hat{V}) \text{ in } D^{3k+6}. \qquad \text{(A-10)}$$

by virtue of Theorem 3.3 of [25] where $\hat{V} = \hat{S} - \tau\hat{A}$, i.e.,

$$\hat{V}=\sum_{i=1}^{k} \hat{V}^i \circ \hat{\tau}^{-1}p_im_ie \circ \tau e+\sum_{i=1}^{k} \tau_i\bar{A}^i \circ \tau e-\tau e+\hat{U}$$
$$=\sum_{i=1}^{k} \hat{V}^i \circ p_im_i\tau\hat{\tau}^{-1}e+\sum_{i=1}^{k} \tau_i\bar{A}_i \circ \tau e-\tau e$$
$$-\tau\sum_{i=1}^{k} \bar{A}^i \circ \tau e+\tau e$$
$$=\sum_{i=1}^{k} \hat{V}^i \circ p_im_i\tau\hat{\tau}^{-1}e+\sum_{i=1}^{k} (\tau_i-\tau)\bar{A}^i \circ \tau e. \qquad \text{(A-11)}$$

Since $\hat{V}^i$ and $\bar{A}^i$ are *BM*'s, $\hat{V}$ is *BM* with 0 drift and variance coefficient σ^2_{22} as shown in (21). (Recall that $a(B\circ be)$ has the same distribution as $ab^{1/2}B$ and has variance a^2b.) (We remark that it is also not difficult to apply Theorem 3.4 of [25] to treat the inverse of composition. It is easy to show that $\{\hat{V}_n\}$ is tight as required in condition (i) there, because obviously $w_{\hat{V}_n}(\delta) \leq \Sigma^k_{i=1} w^i_{\hat{V}_n}(\delta)$ where $w_x(\delta)$ is the modulus of continuity for C-tightness on p. 54 of [24]).

Finally, we obtain the desired $[\hat{U}_n, \hat{V}_n] \Rightarrow [\hat{U}, \hat{V}]$ by taking the (continuous) projection. This yields

$$(\hat{U}, \hat{V})=\Bigg(-\tau\sum_{i=1}^{k} \bar{A}^i \circ \tau e+\tau e,$$
$$\sum_{i=1}^{k} [\hat{V}^i \circ p_im_i\tau\hat{\tau}^{-1}e+(\tau_i-\tau)\bar{A}_i \circ \tau e]\Bigg) \qquad \text{(A-12)}$$

and

$$\hat{V}-\hat{U}=\left(\sum_{i=1}^{k} [\hat{V}^i \circ p_im_i\tau\hat{\tau}^{-1}e+\tau_i\bar{A}_i \circ \tau e]-\tau e\right) \qquad \text{(A-13)}$$

which is *BM* with drift $-\tau$ and variance

$$\sigma^2_{11}+\sigma^2_{22}-2\sigma^2_{12}=(\tau/\hat{\tau})\sum_{i=1}^{k} [p_im_i\tau^2_ic^2_{si}+p_im^2_i\tau^2_i(c^2_{bi}+1)]. \qquad \text{(A-14)}$$

Hence, σ^2_{12} is as in (21). ■

Acknowledgment

We thank M. Zuckerman for helpful comments on this paper.

References

[1] L. Kleinrock, *Communication Nets.* New York: McGraw-Hill, 1964; also Dover, 1972.

[2] B. W. Conolly, "The waiting time process for a certain correlated queue," *Oper. Res.*, vol. 16, pp. 1006-1015, 1968.

[3] O. J. Boxma, "On a tandem queueing model with identical service times at both counters, I and II," *Adv. Appl. Prob.*, vol. 11, pp. 616-659.

[4] I. Rubin, "An approximate time-delay analysis for packet-switching communication networks," *IEEE Trans. Commun.*, vol. COM-24, pp. 210-221, 1976.

[5] S. B. Calo, "Message delays in repeated-service tandem connections," *IEEE Trans. Commun.*, vol. COM-29, pp.670-678, 1981.

[6] C. R. Mitchell, A. S. Paulson, and C. A. Beswick. "The effect of correlated exponential service times on single server tandem queues," *Nav. Res. Log. Quart.*, vol. 24, pp. 95-112, 1977.

[7] W. Kraemer and M. Langenbach-Belz, "Approximate formulae for the delay in the queueing system *GI/G/*1," *Eighth Int. Teletraffic Cong.*, Melbourne, Australia, p. 235, 1976.

[8] W. Whitt, "The queueing network analyzer," *Bell Syst. Tech. J.*, vol. 62, no. 9, pp. 2779-2815, Nov. 1983.

[9] E. Cinlar, "Superposition of point processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. Lewis, Ed. New York: Wiley, 1972, pp. 549-606.

[10] S. L. Albin, "On Poisson approximations for superposition arrival processes in queues," *Manage. Sci.*, vol. 28, pp. 126-137, 1982.

[11] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 833-846, 1986.

[12] W. Whitt, "Approximating a point process by a renewal process, I: Two basic methods," *Oper. Res.*, vol. 30, pp. 125-147, 1982.

[13] W. Whitt, "Queues with superposition arrival processes in heavy traffic," *Stoch. Proc. Appl.*, vol. 21, pp. 81-91, 1985.

[14] S. L. Albin, "Approximating a point process by a renewal process, II: Superposition arrival processes to queues," *Oper. Res.*, vol. 32, pp. 1133-1162, 1984.

[15] D. Y. Burman and D. R. Smith, "An asymptotic analysis of a queueing system with Markov-modulated arrivals," *Oper. Res.*, vol. 34, pp. 105-119, 1986.

[16] H. Heffes and D. M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 856-868, 1986.

[17] C. L. Monma and D. D. Sheng, "Backbone network design and performance analysis: A methodology for packet-switching networks," *IEEE J. Select. Areas Commun.*, vol. SAC-4, pp. 946-965, 1986.

[18] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events.* London, England: Methuen, 1966.

[19] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic, II: Sequences, networks and batches," *Adv. Appl. Prob.*, vol. 2, pp. 355-369, 1970.

[20] W. Feller, *An Introduction to Probability Theory and Its Applications Vol. I.* New York: Wiley, 1968, 3rd ed.

[21] R. W. Wolff, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, pp. 223-231, 1982.

[22] R. B. Cooper, *Introduction to Queueing Theory.* New York: North Holland, 2nd ed.

[23] W. Whitt, "Comparing batch delays and customer delays," *Bell Syst. Tech. J.*, vol. 62, pp. 2001-2009, 1983.

[24] P. Billingsley, *Convergence of Probability Measures.* New York: Wiley, 1968.

[25] W. Whitt, "Some useful functions for functional limit theorems," *Math. Oper. Res.*, vol. 5, pp. 67-85, 1980.

[26] ——, "Heavy traffic limit theorems for queues: A survey," in *Mathematical Methods in Queueing Theory*, A. B. Clarke, Ed., Lecture Notes in Econ. and Math. Syst. 98, New York: Springer-Verlag, pp. 307-350, 1974.

[27] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Cambridge Phil. Soc.*, vol. 48, pp. 497-520, 1962.

[28] A. A. Borovkov, *Stochastic Processes in Queueing Theory.* New York: Springer-Verlag, 1976.

[29] J. F. C. Kingman, "The single server queue in heavy traffic," *Proc. Camb. Phil. Soc.*, vol. 57, pp. 902-904, 1961.

[30] ——, "On queues in heavy traffic," *J. Roy. Statist. Soc.*, vol. B24, pp. 383-392, 1962.

[31] M. I. Reiman, "Open queueing networks in heavy traffic," *Math. Oper. Res.*, vol. 9, pp. 441-458, 1984.

[32] J. M. Harrison, "Brownian models of queueing networks with heterogeneous customer populations," *Proc. IMA Workshop Stoch. Differential Syst.* New York: Springer-Verlag, 1987.

[33] D. Y. Burman and D. R. Smith, "A light traffic theorem for multi-server queues," *Math. Oper. Res.*, vol. 8, pp. 15-25, 1983.

[34] P. J. Fleming, "An approximate analysis of sojourn times in the *M/G/1* queue with round-robin service discipline," *AT&T Bell Lab. Tech. J.*, vol. 63, pp. 1521-1535, 1984.

[35] G. R. Bitran and D. Tirupati, "Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference," *Management Sci.*, vol. 34, pp. 75-100, 1988.

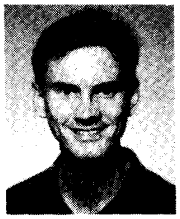
[36] D. L. Iglehart and W. Whitt, "The equivalence of functional central limit theorems for counting processes and associated partial sums," *Ann. Math. Statist.*, vol. 42, pp. 1372-1378, 1971.

[37] P. W. Glynn and W. Whitt, "Ordinary CLT and WLLN versions of $L = \lambda W$," *Math. Oper. Res.*, vol. 13, pp. 674-692, 1988.

[38] R. F. Serfozo, "Functional limit theorems for stochastic processes based on embedded processes," *Adv. Appl. Prob.*, vol. 7, pp. 123-139, 1975.

[39] K. W. Fendick and W. Whitt, "Measurements and approximations to describe the offered load and predict the average workload in a single-server queue," *Proc. IEEE*, vol. 77, 1989.

★



Kerry W. Fendick received the B.A. degree in 1982 from Colgate University, Hamilton, NY, and the M.S. degree in mathematics in 1984 from Clemson University, Clemson, SC.

Since 1984 he has worked at AT&T Bell Laboratories, Holmdel, NJ. As a current member there of the Data Network Analysis Department, he works primarily on problems involving the traffic engineering of packet networks. He is especially interested in queueing problems, as well as other applications of probability theory.

Vikram R. Saksena (S'79-M'82-SM'87) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kharagpur, in 1978, where he was awarded the President of India Gold Medal for academic excellence. Subsequently, he received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1980 and 1982, respectively.

During his stay at the University of Illinois, he served as a Research Assistant in the Decision and Control Laboratory of the Coordinated Science Laboratory where he conducted research on problems in singular perturbations, differential games, and large scale systems. Since 1982 he has been with AT&T Bell Laboratories. As a Member of Technical Staff, he was involved in developing traffic engineering, routing and network design methods for AT&T's Packet Transport Network, in fundamental investigations of packet network architectures, and in the modeling and analysis of integrated voice/data networks. As a Group Supervisor, he is currently directing projects on fault-tolerant data network architectures, traffic engineering of local area networks, wide area networks, and their interconnections, and high speed networking technologies for multiservice integration. His overall interest is in applying systems engineering techniques to problems in modeling analysis, and synthesis of modern telecommunications networks. He has several publications in the areas of control theory and communications networks.

Dr. Saksena is a member of Tau Beta Pi, Eta Kappa Nu, and Phi Kappa Phi.

★

Ward Whitt received the A.B. degree in mathematics from Dartmouth College, Hanover, NH, in 1964, and the Ph.D. degree in operations research from Cornell University, Ithaca, NY, in 1969.

He taught in the Department of Operations Research at Stanford University in 1968-1969 and in the Department of Administrative Sciences at Yale University from 1969-1977. Since 1977, he has been employed by AT&T Bell Laboratories. He is currently a Member of Technical Staff in the Mathematical Sciences Research Center, Murray Hill, NJ.

Dr. Whitt is a member of the Operations Research Society of America and the Institute of Mathematical Statistics.