Towards better multi-class parametric-decomposition approximations for open queueing networks

Ward Whitt

AT&T Bell Laboratories, Murray Hill, NJ 07974-0636, USA

Methods are developed for approximately characterizing the departure process of each customer class from a multi-class single-server queue with unlimited waiting space and the first-in-first-out service discipline. The model is $\Sigma(GI_i/GI_i)/1$ with a non-Poisson renewal arrival process and a non-exponential service-time distribution for each class. The methods provide a basis for improving parametric-decomposition approximations for analyzing non-Markov open queueing networks with multiple classes. For example, parametric-decomposition approximations are used in the Queueing Network Analyzer (QNA). The specific approximations here extend ones developed by Bitran and Tirupati [5]. For example, the effect of class-dependent service times is considered here. With all procedures proposed here, the approximate variability parameter of the departure process of each class is a linear function of the variability parameters of the arrival processes of all the classes served at that queue, thus ensuring that the final arrival variability parameters in a general open network can be calculated by solving a system of linear equations.

Keywords: Open queueing networks, multi-class queueing networks, parametric-decomposition approximations, departure processes, heavy-traffic limit theorems.

1. Introduction and summary

1.1. PARAMETRIC-DECOMPOSITION APPROXIMATIONS

A useful way to analyze the steady-state performance of open queueing networks with non-Poisson external arrival processes and non-exponential service-time distributions is the parametric-decomposition approximation method, first proposed by Reiser and Kobayashi [23] and subsequently extended by the author [32, 24] and many others (see the references). The main idea is to approximately analyze the individual queues separately after approximately characterizing the arrival processes to each queue by a few parameters (usually two, one to represent the rate and another to represent the variability). The goal is to approximately represent the network dependence through these arrival-process parameters. After the congestion in each queue has been described, the total network performance is approximated by acting as if all the queues are mutually independent, i.e., the rest of the approximation is performed as if the steady-state distribution of the numbers of customers at the queues had a product form.

An attractive alternative to parametric-decomposition approximations are Brownian models, as in Harrison and Nguyen [13, 14]. Brownian models can even be used together with parametric-decomposition schemes, as in Dai et al. [8]. However, here we only consider parametric-decomposition approximations.

1.2. AGGREGATION IN MULTI-CLASS MODELS

The primary purpose of this paper is to present new methodology for extending the parametric-decomposition approximation method to treat queueing networks with several classes of customers. The procedure in [32, 24] already allows multiple customer classes, but there all the classes are aggregated to form a single class before the rest of the approximation, in the spirit of the celebrated Kleinrock independence assumption; p. 50 of [18]. With this procedure, all class identity is not lost; the expected sojourn time of a customer following a given route is the sum of the expected sojourn times at the queues on that route, with the expected service time components being the original expected service times specified for that particular customer class; the aggregation only affects the calculation of the expected delays (before beginning service) at the nodes on the route. Moreover, the aggregation procedure yields the correct traffic intensities, so that in the delay calculations the only approximation appears in the variability parameters.

In many cases this aggregation step works quite well, but in some cases it does not. Difficulties with aggregation in the parametric-decomposition approximations were noted by Bitran and Tirupati [5] and Fendick et al. [9, 10]. Bitran and Tirupati point out difficulties with multiple classes and deterministic routing, especially in the low-variability context common to manufacturing models. Fendick et al. point out difficulties with multiple classes and highly variable (e.g. batch) arrival processes together with class-dependent service times.

Other difficulties with parametric-decomposition approximations are noted by Suresh and Whitt [29] and Whitt [37]. Suresh and Whitt [29] show how exceptional variability (either high or low) in an arrival process to a queue can be reduced in the departure process in a short time scale when the queue has a moderate traffic intensity (e.g. $\rho = 0.6$) and moderately variable service times (e.g. exponential), while the exceptional variability in a larger time scale remains. This exceptional variability in a larger time scale typically has little effect upon congestion in subsequent queues with low-to-moderate traffic intensity, but it typically has a dramatic effect upon the congestion in a subsequent queue with high traffic intensity. *This phenomenon means that it can be difficult to characterize the variability of an arrival process by a single variability parameter*. For example, the arrival process might have low variability in a short time scale and high variability in a longer time scale, so that in a subsequent queue with traffic intensity ρ congestion would be predicted well by having an arrival process variability parameter (squared coefficient of variation) $c_a^2 \approx 0.5$ when $\rho = 0.5$ and $c_a^2 \approx 5.0$ when $\rho = 0.9$. This difficulty is the motivation for the use of *variability functions* instead of variability parameters, e.g. the indices of dispersion in [9, 10] and references there (which we will not discuss further here).

Whitt [37] shows that multi-class queueing networks with class-dependent service times can exhibit relatively complex behavior. In particular, there can be unanticipated large fluctuations in the individual queue lengths due to the sudden movement of blocks of customers with very short service times. This phenomenon suggests that it may be important to focus on the transient behavior as well as the steady-state distribution. It remains to determine the implications for steady-state distributions.

1.3. PARAMETRIC-DECOMPOSITION APPROXIMATIONS WITHOUT AGGREGATION

The difficulties with aggregation into a single class suggest the need for parametric-decomposition procedures without aggregation. What we want is an extension of the algorithm in [32, 24] that produces arrival process parameters at each node *for each class*. (The resulting approximate congestion measures such as expected delays at each queue might also be class-dependent as in Holtzman [15], Albin [2] and Fischer and Stanford [11], but we do not focus on that here.) In fact, such a multi-class extension of the parametric-decomposition approximation was proposed by Bitran and Tirupati, and it provides dramatic improvements in accuracy in some cases. Their main contribution is an approximation for the variability parameter of the departure process for each class from a single-server queue when the arrival process for each class is characterized by an arrival rate and a variability parameter. As usual, the variability parameters are squared coefficients of variation (SCV, variance divided by the square of the mean) in renewal-process approximations. The Bitran–Tirupati approximation is based on the two-class case, by aggregating all classes except the one of interest into one. Their approximation results in a refinement of the splitting step in section 4.4 of [32].

Throughout this paper we consider a single-server queue with unlimited waiting space and the FIFO (first-in-first-out) discipline. (However, the results also provide a basis for treating multi-server queues; see remark 2.5.) Let c_d^2 and c_{d1}^2 be the variability parameters of the overall departure process and the departure process for class 1 alone; let p_1 be the proportion of all departures that are class 1. If the total departure process were a renewal process and if each successive departure were class 1 according to Bernoulli (independent) trials with probability p_1, then the exact relation is

$$c_{d1}^2 = p_1 c_d^2 + 1 - p_1, \tag{1}$$

as given in (36) of [32]. Formula (1) obviously makes c_{d1}^2 close to 1 when p_1 is small, but without Bernoulli routing the actual variability can be quite different. As shown in [5], deterministic routing can cause the true relation to deviate significantly from (1). As an improvement, Bitran and Tirupati propose

$$c_{d1}^2 = p_1 c_d^2 + c_{n1}^2, \tag{2}$$

where c_{n1}^2 is the squared coefficient of variation of the total number of customers that arrive during an interarrival time of class 1; see (6) of [5]. If the superposition arrival process of the complement to class 1, henceforth referred to as class 2, is a Poisson process, then

$$c_{n1}^2 = (1-p_1)(p_1 + (1-p_1)c_{a1}^2), \tag{3}$$

where c_{a1}^2 is the class-1 arrival-process variability parameter; see (7) of [5]. Bitran and Tirupati also develop numerical procedures (involving iteration) for calculating approximate values of c_{n1}^2 when the class-2 arrival process is less variable than Poisson, i.e., when $c_{a2}^2 < 1$. These numerical procedures (INT2 and INT3 in [5]) are based on the assumption that the class-1 and class-2 arrival processes are renewal processes with Erlang interarrival-time distributions. When the system is characterized by low variability, these numerical procedures perform significantly better than (3), but these procedures are somewhat cumbersome.

1.4. ENHANCEMENTS IN THE BITRAN–TIRUPATI SCHEME

In [35] we proposed enhancements to the Bitran–Tirupati [5] approximations, which we present here. (The present paper is an update of [35].) Further contributions in this direction have been made by Stanford and Fischer [27, 28] and Fischer and Stanford [11]. Some of the results here have also been exploited in [24].

We contribute to the Bitran–Tirupati approximation scheme by developing a new approximation for c_{n1}^2 in (2). In particular, we propose the formula

$$c_{n1}^2 = (1-p_1)(p_1 c_{a2}^2 + (1-p_1)c_{a1}^2), \tag{4}$$

where, as above, c_{a2}^2 is the approximating SCV for the superposition of all class j arrival processes except class 1. Note that (4) reduces to (3) in the special case $c_{a2}^2 = 1$. Formula (4) provides a simple alternative to the complex Erlang numerical procedures when $c_{ai}^2 \leq 1$, for $i = 1$ and 2. It also applies to the important case when $c_{ai}^2 > 1$ for $i = 1$ or 2, which was not treated in [5]. As with the formulas in [32], formulas (2)–(4) are appealing because they are linear in the arrival and departure variability parameters, so that the final arrival-process variability parameters for all the queues in the network can be obtained by simply solving systems of linear equations.

Just as Bitran and Tirupati obtained (3) in [5], we obtain approximation (4) by considering specific renewal processes for which we can calculate c_{n1}^2 (exactly or approximately). For this purpose, we exploit batch-Poisson (B-P) and batch-deterministic (B-D) processes with geometric batch sizes; i.e. the interarrival times of batches is exponential (B-P) or deterministic (B-D) and the size of the batches is geometric on the positive integers. The geometric batch-size distribution makes the individual customer interarrival times i.i.d. (independent and identically distributed). The B-P and B-D processes are convenient because they are two-parameter renewal processes. There is thus a direct correspondence between these parameters and the rate and variability parameter used in the approximations. For any $c_a^2 \geq 0$ (≥ 1), there is a unique B-D (B-P) process with the given c_a^2 and arrival rate. We use B-D as well as B-P to treat the cases with $0 \leq c_a^2 < 1$. However, it turns out that both cases yield the same approximation.

Of course, the proof of the pudding is in the tasting. We show that (4) performs quite well when compared to simulation and the other approximations for the experiments considered by Bitran and Tirupati [5]. The accuracy in this step is good, but not phenomenal; it seems to be consistent with the accuracy of other approximations used in the overall procedure.

The desired approximation for c_{d1}^2 is obtained by combining (2) and (4). Note that the resulting formula

$$c_{d1}^2 = p_1 c_d^2 + p_1(1-p_1)c_{a2}^2 + (1-p_1)^2 c_{a1}^2 \tag{5}$$

is a convex combination: The weights $p_1, p_1(1-p_1)$ and $(1-p_1)^2$ on the variability parameters c_d^2, c_{a2}^2 and c_{a1}^2 sum to 1. Furthermore, a common approximation for c_d^2 is another convex combination

$$c_d^2 = \rho^2 c_s^2 + (1-\rho^2)c_a^2, \tag{6}$$

where ρ is the traffic intensity, and c_s^2 and c_a^2 are the variability parameters (squared coefficients of variation) for the service times and the total arrival process; see (38) of [32], (23) of [33] and (2) of [5]. Combining (5) and (6), we obtain

$$c_{d1}^2 = \rho^2 p_1 c_s^2 + (1-\rho^2) p_1 c_a^2 + p_1(1-p_1)c_{a2}^2 + (1-p_1)^2 c_{a1}^2. \tag{7}$$

If we continue and approximate c_a^2 by the asymptotic method, (4.14) of [23] or (1) of [5], then

$$c_a^2 = p_1 c_{a1}^2 + (1-p_1)c_{a2}^2. \tag{8}$$

Combining (7) and (8), we obtain the convex combination

$$c_{d1}^2 = \rho^2 p_1 c_s^2 + (2-\rho^2) p_1(1-p_1)c_{a2}^2 + [(1-p_1)^2 + (1-\rho^2)p_1^2]c_{a1}^2. \tag{9}$$

When there actually are k classes with approximating arrival SCVs $\tilde{c}_{aj}^2$, we can also use the asymptotic method for c_{a2}^2 to obtain

$$c_{a2}^2 = \sum_{j=2}^{k} [p_j/(1-p_1)]\tilde{c}_{aj}^2, \tag{10}$$

where p_j and $\tilde{c}_{aj}^2$ are the corresponding parameters for class j, and

$$c_{d1}^2 = \rho^2 p_1 c_s^2 + (2-\rho^2) p_1 \sum_{j=2}^{k} p_j \tilde{c}_{aj}^2 + [(1-p_1)^2 + (1-\rho^2) p_1^2] c_{a1}^2. \tag{11}$$

Formula (11) is the natural generalization of the first Bitran–Tirupati procedure (INT1) based on (2) and (3); their procedure is the same except $\tilde{c}_{aj}^2$ for $j \neq 1$ is replaced by 1 in (11).

Natural alternatives to (11) are obtained by using different approximations for the superposition variability parameters c_a^2 and c_{a2}^2 than (8) and (10). In particular, the stationary-interval method and various hybrid approximations can be used instead; see section 4.1 of [31], [1], and section 4.3 of [32]. We examine the simple alternative based on (29) and (30) of [32] for the Bitran and Tirupati experiments; i.e., $c_a^2 = wc_{AM}^2 + 1 - w$, where c_{AM}^2 is the asymptotic-method approximation and the weight w comes from (29) of [32]. For these cases, the hybrid using (29) and (30) of [32] performs better than (11), but both perform quite well. (See section 5.)

1.5. THE LOW-INTENSITY VARIABILITY-PRESERVATION PRINCIPLE

From (5) or the subsequent formulas (7), (9) and (11), we can see what the approximation predicts in limiting cases. As $p_1 \to 1$, $c_{d1}^2 \to c_d^2$ as it obviously should. As $p_1 \to 0$, $c_{d1}^2 \to c_{a1}^2$. The appropriateness of the limit as $p_1 \to 0$ is less obvious, but upon reflection it can be seen to be, as Bitran and Tirupati argue. In [36] we prove a limit theorem rigorously justifying this limiting behavior. In fact, we show that under very general conditions the entire class-1 departure process converges in distribution to the class-1 arrival process as $p_1 \to 0$ (i.e. the finite-dimensional distributions converge). In fact, with probability one, each sample path of the class-1 departure process converges to the corresponding sample path of the class-1 arrival process. (The general idea of the low-intensity variability-preservation principle is due to Bitran and Tirupati [5], but the strong forms involving the distribution of the entire stochastic process and the individual sample paths appear in [36].)

The analysis in [5, 36] and here thus supports the remarkably simple approximation

$$c_{d1}^2 \approx c_{a1}^2 \quad \text{for } p_1 \text{ small}, \tag{12}$$

which has very significant implications for queueing networks. For an open queueing network with a very large number of classes, (12) helps provide rapid back-of-the-envelope approximations. The associated delays at the queues should be calculated using superposition approximations (e.g. [1, 2, 9–11, 15, 32]) though. At queues to which many classes come, each with relatively small intensity, the delays for each class would be essentially the same as for Poisson arrivals, but if some class passes through several queues at which its proportion of the total arrival rate is very small, and then comes to a queue at which it is the only class or there are only a few classes, then the variability of the original external arrival process of this class should play a role; i.e., the appropriate variability parameter for the arrival process of this class at this last queue would be the variability parameter of the external arrival process of this class (just as in [29] discussed in section 1.2).

This important phenomenon arises in many applications. For example, in packet communication networks where messages are sent over virtual circuits (fixed routes), packets often enter the network in a highly bursty manner over a relatively slow access line where there is relatively little sharing of facilities. In contrast, in the network there is substantial sharing because the network switching and transmission are orders of magnitude faster. Finally, the packets emerge from the network and proceed to their destination over another relatively slow access line. Formula (12) and the discussion above indicate that the high variability should be substantially dissipated within the network, but should reappear at the destination. Even though the packets might pass through several queues in the network, the packet arrival process at the destination (the packet departure process from the network) should be similar to the original packet arrival process at the source. In [36] this phenomenon is substantiated by a simulation of a packet network model from [9].

1.6. CLASS-DEPENDENT SERVICE TIMES

Motivated by [9] we also want to treat models in which the different classes can have different service-time distributions, a situation not addressed by Bitran and Tirupati [5]. As shown in [9], class-dependent service times can cause strong dependence among successive service times (and thus evidently in the overall departure process). To appreciate the significance of class-dependent service-time distributions, consider the two-class case in which the class-2 service times are zero. Obviously the class-1 departure process from this queue is the same as if class 2 were not present; consequently the approximation for c_{d1}^2 should be independent of p_1. Our analysis produces approximations for the general case. It also suggests, for simple approximations, that the traffic-intensity proportion should often appear in (1)–(11) instead of the arrival-rate proportion. In fact, both play a role. To state our proposed approximation to account for class-dependent service times, let ρ_i be the contribution to the traffic intensity by class i. Instead of (11), we propose the approximation

$$c_{d1}^2 = \rho_1^2 c_{s1}^2 + p_1 \sum_{j=2}^{k} \rho_j^2 p_j^{-1} (\tilde{c}_{aj}^2 + c_{sj}^2) + (1 - 2\rho_1 \rho + \rho_1^2) c_{a1}^2. \tag{13}$$

Proper treatment of class-dependent service times is vital for treating manufacturing models in which some classes are introduced to represent occasional down times of machines. For such models, the approximations here provide significant improvements over [5], just as [5] provides significant improvements over [32].

1.7. SUPPORTING METHODOLOGY: THE CASE OF A CONTINUOUSLY BUSY SERVER

In the spirit of [31, 33], we also want to provide a systematic basis for developing approximations. Thus, we describe asymptotic-method (AM) and stationary-interval (SI) characterizations that can be the basis for refined hybrid approximations. To a large extent, this paper can thus be regarded as a multi-class extension of [33]. We focus solely on departure processes, but the application to queueing networks should be clear.

In our detailed mathematical analysis, we focus on a special limiting case, the case in which the server is continuously busy. We develop detailed descriptions of the AM and SI approximations under this condition. The results are thus directly applicable only to the case $\rho \geq 1$, but more generally they can be exploited to develop hybrid approximations. The idea is to use convex combinations with weights on the continuously-busy approximations that approach 1 as $\rho \to 1$. It is significant that the final AM and SI continuously-busy approximations agree, because our approximating assumptions make the continuously-busy class-1 departure process a renewal process; see (19) and (28). The SI continuously-busy approximation also yields an approximation for c_{n1}^2 in (2) as a special case; we simply set all the service times equal to 1. Then the class-1 interdeparture time is precisely the number of customers to arrive during a class-1 interarrival time. Our generalization of (3) appears in (31). When the service times are not class-dependent, (31) reduces to (4); otherwise the arrival rate proportions in (4) should be replaced by traffic intensity proportions.

1.8. ORGANIZATION OF THE REST OF THIS PAPER

The rest of this paper is organized as follows. In section 2 we develop the AM approximation under the continuously-busy assumption; the final AM variability parameter is (19). A fairly general interesting special case appears in (20). In section 3 we develop the SI approximation under the continuously-busy assumption. In section 3.1 we show how to approximate a general arrival process partially characterized by its arrival rate and variability parameter by a B-D or B-P renewal process. In section 3.2 (3.3) we calculate the SI approximation for c_{d1}^2 under the assumption that class 2 is a B-P (B-D) renewal process. In section 4 we discuss refined hybrid procedures. In section 5 we make comparisons with simulation and other approximations in the case of common service-time distributions using the Bitran–Tirupati experiments in [5]. There we show that (11) and the variant using (29) and (30) of

[32] for superposition instead of the AM approximation in (8) and (10) perform well. Finally, we present our conclusions in section 6.

2. Asymptotic-method approximation with a continuously busy server

Consider a single-server queue with unlimited waiting room and the FIFO discipline to which k classes of customers arrive to receive service. Let customers from class i arrive according to an arrival counting process $A_i(t)$ and have successive service times ν_{in}, $n \geq 1$. Let $D_i(t)$ be the resulting departure counting process for class i. In this section we develop an asymptotic-method (AM) approximation for the vector of departure processes $[D_1(t), \ldots, D_k(t)]$ under the heavy-traffic-type assumption that the server is continuously busy. In particular, we prove a functional central limit theorem (FCLT) for $[D_1(t), \ldots, D_k(t)]$ under general FCLT conditions. For the approximations, this means that we express the AM variability parameter for the departure process of class i, c_{Di}^2, in terms of the AM variability parameters of the arrival processes and service times $c_{A1}^2, \ldots, c_{Ak}^2$, $c_{S1}^2, \ldots, c_{Sk}^2$ and the associated means; see [4, 16, 17, 30, 31, 33] for background.

2.1. A GENERAL FUNCTIONAL CENTRAL LIMIT THEOREM

We work in the setting of [4] and [30], which means weak convergence (convergence in distribution), denoted by $\Rightarrow$. We consider random elements of $D \equiv D[0, \infty)$, the space of all real-valued functions on $[0, \infty)$ which are right continuous with left limits. Let the space D be endowed with the standard Skorohod (J_1) topology and let product spaces D^k be endowed with the usual product topology. Let $C \equiv C[0, \infty)$ be the subset of continuous functions in D. Convergence $x_n \to x$ in D reduces to uniform convergence on compact subsets when $x \in C$.

We define the following random elements of D:

$$\hat{A}_{in}(t) = n^{-1/2}[A_i(nt) - \lambda_i nt],$$

$$\hat{V}_{in}(t) = n^{-1/2}\left[\sum_{j=1}^{[nt]} \nu_{ij} - \tau_i nt\right],$$

$$\hat{D}_{in}(t) = n^{-1/2}[D_i(nt) - \delta_i nt], \tag{14}$$

for $1 \leq i \leq k$ and $t \geq 0$. Obviously λ_i is intended to be the arrival rate and τ_i the mean service time of class i. Let $\rho_i = \lambda_i \tau_i$ be the associated traffic intensity for class i. Let $\lambda = \lambda_1 + \ldots + \lambda_k$ be the total arrival rate, $\tau = \lambda^{-1} \Sigma_{i=1}^k \lambda_i \tau_i$ the overall mean service time and $\rho = \rho_1 + \ldots + \rho_k = \lambda\tau$ the total traffic intensity. We assume that the server is eventually continuously busy, which means that $\rho \geq 1$. Obviously the overall departure rate must be τ^{-1} if the server is always busy. Hence, we should have $\delta_i = \lambda_i/\lambda\tau$. Our main result in this section is a FCLT for the departure processes given a joint FCLT for the arrival processes and service times. The resulting approximation under the standard independence and moment conditions appears in (19) below.

THEOREM 1

If $[\hat{A}_{1n}, \ldots, \hat{A}_{kn}, \hat{V}_{1n}, \ldots, \hat{V}_{kn}] \Rightarrow [\hat{A}_1, \ldots, \hat{A}_k, \hat{V}_1, \ldots, \hat{V}_k]$ in D^{2k} where $P(\hat{A}_i \in C) = P(\hat{V}_i \in C) = 1, 1 \le i \le k$, and $\rho > 1$, then

$$[\hat{A}_{1n}, \ldots, \hat{A}_{kn}, \hat{V}_{1n}, \ldots, \hat{V}_{kn}, \hat{D}_{1n}, \ldots, \hat{D}_{kn}]$$
$$\Rightarrow [\hat{A}_1, \ldots, \hat{A}_k, \hat{V}_1, \ldots, \hat{V}_k, \hat{D}_1, \ldots, \hat{D}_k] \quad \text{in } D^{3k},$$

where $\delta_i = \lambda_i/\lambda\tau$ and

$$\hat{D}_i(t) = \left(1 - \frac{\rho_1}{\rho}\right)\rho^{-1/2}\hat{A}_i(t) - \left(\frac{\lambda_i}{\rho}\right)^{3/2}\hat{V}_i(t) - \frac{\lambda_i}{\rho^{3/2}}\sum_{\substack{j=1\\ j\neq i}}^{k}[\lambda_j^{1/2}\hat{V}_j(t) + \tau_j\hat{A}_j(t)].$$

Proof

Let $T(t)$ be the process representing the total work to arrive in the interval $[0, t]$ and let $C(t)$ be an associated inverse process, defined by

$$C(t) = \sup\{s \ge 0 : T(s) \le t\}.$$

Then $D_i(t) = A_i(C(t))$, $t \ge 0$, by virtue of the continuously-busy assumption. Since $\rho > 1$, the continuously-busy assumption is eventually satisfied, so that the limiting behavior is unaffected by any initial discrepancy (idleness). (This can be rigorously justified by theorem 4.1 of [4]; we only consider the continuously busy case.) Define the following random functions in D:

$$\hat{T}_{in}(t) = n^{-1/2}\left[\sum_{j=1}^{A_i(nt)} \nu_{ij} - \lambda_i\tau_i nt\right], \qquad \hat{T}_n = \hat{T}_{1n} + \ldots + \hat{T}_{kn},$$

$$\hat{C}_n(t) = n^{-1/2}\left[C(nt) - \left(\sum_{i=1}^{k}\lambda_i\tau_i\right)^{-1} nt\right], \quad t \ge 0. \tag{15}$$

As in [9],

$$[\hat{A}_{1n},\ldots,\hat{A}_{kn}, \hat{V}_{1n},\ldots,\hat{V}_{kn}, \hat{T}_{1n},\ldots,\hat{T}_{kn}, \hat{T}_n, \hat{C}_n]$$
$$\Rightarrow [\hat{A}_1,\ldots,\hat{A}_k, \hat{V}_1,\ldots,\hat{V}_k, \hat{T}_1\ldots,\hat{T}_k, \hat{T}, \hat{C}]$$

in D^{3k+1}, where

$$\hat{T}_i(t) = \hat{V}_i(\lambda_i t) + \tau_i \hat{A}_i(t), \quad \hat{T} = \hat{T}_1 + \cdots + \hat{T}_k,$$

and

$$\hat{C}(t) = -\left(\sum_{i=1}^{k} \lambda_i \tau_i\right)^{-1} \hat{T}\left(t/\sum_{i=1}^{k} \lambda_i \tau_i\right), \quad t \geq 0,$$

by theorems 5.1, 4.1 and 7.3 of [30]; see the remark after theorem 5.1 and the corollary to lemma 7.6. Applying theorem 5.1 of [30] again, we see that $[\hat{D}_{1n},\ldots,\hat{D}_{kn}] \Rightarrow [\hat{D}_1,\ldots,\hat{D}_k]$ jointly with all the processes above, where

$$\begin{aligned}
\hat{D}_i(t) &= A_i(t/\lambda\tau) = (\lambda_i/\lambda\tau) \sum_{j=1}^{k} [\hat{V}_j(\lambda_j t/\lambda\tau) + \tau_j \hat{A}_j(t/\lambda\tau)] \\
&\stackrel{\mathrm{d}}{=} (\lambda\tau)^{-1/2}\left(\hat{A}_i(t) - (\lambda_i/\lambda\tau) \sum_{j=1}^{k} [\lambda_j^{1/2} \hat{V}_j(t) + \tau_j \hat{A}_j(t)]\right) \\
&= (\lambda\tau)^{-1/2}\left(1 - \frac{\rho_i}{\rho}\right)\hat{A}_i(t) - \left(\frac{\lambda_i}{\lambda\tau}\right)^{3/2} \hat{V}_i(t) \\
&\quad - \frac{\lambda_i}{(\lambda\tau)^{3/2}} \sum_{\substack{j=1 \\ j\neq i}}^{k} [\lambda_j^{1/2} \hat{V}_j(t) + \tau_j \hat{A}_j(t)],
\end{aligned}$$

with $\stackrel{\mathrm{d}}{=}$ (equality in distribution, as processes) holding by the normalization in (13); e.g. the limits must satisfy $\hat{A}_i(ct) \stackrel{\mathrm{d}}{=} c^{1/2}\hat{A}_i(t)$. □

Remarks

(2.1) A FCLT for the total departure process $D(t) = D_1(t) + \ldots + D_k(t)$ with $\rho > 1$ was previously established in theorem 4.2 of [16].

(2.2) Under standard additional independence assumptions, the limit process $[\hat{A}_1,\ldots,\hat{A}_k, \hat{V}_1,\ldots,\hat{V}_k]$ is composed of independent Brownian motions (BMs). Then the limit in theorem 1 is multivariate BM. If the limit processes

$\hat{A}_1, \ldots, \hat{A}_k$, $\hat{V}_1, \ldots, \hat{V}_k$ in theorem 1 are independent BMs with zero means and variances $\alpha_1^2, \ldots, \alpha_k^2$, $\beta_1^2, \ldots, \beta_k^2$, respectively, then $[\hat{D}_1, \ldots, \hat{D}_k]$ is a BM in C^k with zero means, variances σ_i^2 and covariances σ_{ij}^2, where

$$\sigma_i^2 = \left(1 - \frac{\rho_i}{\rho}\right)^2 \frac{\alpha_i^2}{\rho} + \left(\frac{\lambda_i}{\rho}\right)^3 \beta_i^2 + \frac{\lambda_i^2}{\rho_3} \sum_{\substack{j=1 \\ j \neq i}}^{k} [\lambda_j \beta_j^2 + \tau_j^2 \alpha_j^2] \tag{16}$$

and

$$\sigma_{ij}^2 = -\left(1 - \frac{\rho_i}{\rho}\right) \frac{\lambda_j \tau_i \alpha_i^2}{\rho^2} - \left(1 - \frac{\rho_j}{\rho}\right) \frac{\lambda_i \tau_j \alpha_j^2}{\rho^2} + \frac{\lambda_i \lambda_j}{\rho^3} (\lambda_i \beta_i^2 + \lambda_j \beta_j^2). \tag{17}$$

(2.3) We obtain the resulting asymptotic method (AM) approximation for the departure process from (16) if we work with squared coefficients of variation. Let the AM parameters be

$$c_{Ai}^2 = \lambda_i^{-1} \alpha_i^2, \quad c_{Di}^2 = \delta_i^{-1} \sigma_i^2 \quad \text{and} \quad c_{Si}^2 = \tau_i^{-2} \beta_i^2. \tag{18}$$

We treat c_{Ai}^2 and c_{Di}^2 differently from c_{Si}^2 in (18) because $\hat{A}_{in}(t)$ and $\hat{D}_{in}(t)$ are random functions associated with counting processes, while $\hat{V}_{in}(t)$ is not; see section 2 of [31]. From (16) and (18), we obtain

$$c_{Di}^2 = (1 - q_i)^2 c_{Ai}^2 + q_i^2 c_{Si}^2 + \sum_{\substack{j=1 \\ j \neq i}}^{k} q_j^2 (p_i/p_j)(c_{Aj}^2 + c_{Sj}^2), \tag{19}$$

where $q_i = \rho_i/\rho$ and c_{Ai}^2 and c_{Si}^2 are the AM variability parameters determined by the FCLT for the arrival and service processes based on (13), (16) and (18). Note that most of c_{Di}^2 in (19) is dimensionless, as it must be. Note that most of the weights in (19) are functions of $q_i = \rho_i/\rho$ instead of $p_i = \lambda_i/\lambda$; i.e. the relative traffic intensities appear in (19) as well as the relative arrival rates in (1)–(11). In the special case of two classes, if $\tau_2 = 0$, then $q_1 = 1$ and $c_{D_1}^2 = c_{S1}^2$, as it should; if $\tau_1 = 0$, then $q_1 = 0$, $q_2 = 1$ and $c_{D_1}^2 = c_{A1}^2 + (p_1/p_2)(c_{A2}^2 + c_{S2}^2)$. If p_2 and q_1 are both very small (a rather pathological case), then $c_{D_1}^2$ is very large. However, this is realistic; then class 2 must be contributing rare exceptionally long service times, as in the case of service interruptions, e.g. machine down times.

(2.4) A trivial case arises when $k = 1$. Then the departure process assuming the server is continuously busy is obviously just the counting process associated with the service times. From $\hat{V}_{in} \Rightarrow \hat{V}_1$ plus the corollary on p. 83 of [30], we get $\hat{D}_{1n} \Rightarrow \hat{D}_1$ where $\delta_1 = 1/\tau_1$ and $\hat{D}_1(t) = -\tau_1^{-1} \hat{V}_1(t/\tau_1) \stackrel{\text{d}}{=} -\tau_1^{-3/2} \hat{V}_1(t)$. Note that this is consistent with theorem 1; then $(1 - \rho_1/\rho) = 0, \lambda_1/\lambda = 1$, $\lambda_j = 0$ and $\hat{A}_j(t) = 0$ for $j \neq 1$.

(2.5) Theorem 1 extends relatively easily to queues with m parallel servers. The limit holds with t replaced by t/m in $\hat{D}_{in}(t)$ and $\hat{D}_i(t)$, so that the AM approximation c_{Di}^2 in (19) is unchanged. The proof of the extended version of theorem 1 is complicated by the fact that $D_i(t)$ does not coincide exactly with $A_i(C(t/m))$ when $m > 1$, but the difference is asymptotically negligible. Theorem 4.1 of [4] can be applied because the normalized difference in the FCLT is dominated by

$$\max_{1 \le i \le k} \sup_{\substack{1 \le j \le A_i(nt) \\ 0 \le t \le T}} n^{-1/2} \nu_{ij},$$

which converges to 0 in probability because $\hat{T}_n \Rightarrow \hat{T}$ where $P(\hat{T} \in C) = 1$; apply the maximum jump functional with theorem 5.1 of [4].

(2.6) As noted in [31, 33], the AM approximation is asymptotically correct in heavy traffic, where heavy traffic applies at subsequent queue where the point process is the arrival process. In fact, we can simply combine theorem 1 here with theorem 1 of [17]. In the setting of remark 2.2, this means that if the departure process $D_i(t)$ serves as the sole arrival process at another queue of the same type (where the service times are i.i.d. and independent of $D_i(t)$) with traffic intensity ρ', then the standard heavy-traffic limit holds for this second queue as $\rho' \to 1$ and the limit depends on the process $D_i(t)$ only through c_{Di}^2 in (19) and its rate via the contribution to ρ'. □

2.2. A SPECIFIC MULTI-CLASS MODEL WITH BATCH ARRIVALS

We now describe one fairly general special case of the model in section 2.1 that was considered in [9, 10]. For class i let the service times be i.i.d. with mean τ_i and squared coefficient of variation c_{si}^2; let arrivals be generated in i.i.d. batches with batch size having mean m_i and squared coefficient of variation c_{bi}^2; let the arrivals within a batch be separated by i.i.d. spacings with mean ξ_i and squared coefficient of variation c_{xi}^2; let the interval between the last arrival of one batch and the first arrival of the next batch be the sum of one spacing and an idle time; let the successive idle times be i.i.d. with mean η_i and squared coefficient of variation c_{yi}^2; and let the service times, batch sizes, spacings and idle times for all the classes be mutually independent. Let $\gamma_i = m_i\xi_i/(m_i\xi_i + \eta_i)$. The parameter γ_i measures the long-run proportion of time that the arrival process is in a busy state (not an idle time). The arrival rate for class i is $\lambda_i = m_i/(m_i\xi_i + \eta_i)$ and the traffic intensity is $\rho_i = \lambda_i\tau_i$.

COROLLARY

If $\rho > 1$ for this particular multi-class model, then the conditions of theorem 1 are satisfied with the limit process $[\hat{A}_1, \ldots, \hat{A}_k, \hat{V}_1, \ldots, \hat{V}_k]$ being composed of independent BMs, so that $[\hat{D}_1, \ldots, \hat{D}_k]$ is a BM in C^k and

$$\begin{aligned} c_{Ai}^2 &\equiv \lambda_i^{-1} \alpha_i^2 = m_i(1-\gamma_i)^2(c_{bi}^2 + c_{yi}^2) + \gamma_i^2 c_{xi}^2, \\ c_{Si}^2 &\equiv \tau_i^{-2} \beta_i^2 = c_{si}^2, \\ c_{Di}^2 &\equiv \delta_i^{-1} \sigma_i^2 = \rho \lambda_i^{-1} \sigma_i^2 \\ &= a(1-q_i)^2 [m_i(1-\gamma_i)^2(c_{bi}^2 + c_{yi}^2) + \gamma_i^2 c_{xi}^2] + q_i^2 c_{si}^2 \\ &\quad + \sum_{\substack{j=1 \\ j \neq i}}^{k} q_j^2 (p_i/p_j)[(m_j(1-\gamma_j)^2(c_{bj}^2 + c_{yj}^2) + \gamma_j^2 c_{xj}^2 + c_{sj}^2)], \end{aligned} \tag{20}$$

where again $q_i = \rho_i/\rho$ and $p_i = \lambda_i/\lambda$.

The key supporting FCLT for $A_i(t)$ and the formula for c_{Ai}^2 are established in [9]. (The heavy traffic limit for the workload (virtual waiting time) and waiting time processes as $\rho \to 1$ is also established in [9].) The corollary to theorem 1 expresses c_{Di}^2 in terms of the $4k$ variability parameters $(c_{sj}^2, c_{bj}^2, c_{xj}^2, c_{yj}^2)$, and the $4k$ means $(\tau_j, m_j, \xi_j, \eta_j)$, $1 \le j \le k$. Note that the means only affect c_{Di}^2 via the ratios p_i/p_j, $\gamma_i = m_i \xi_i/(m_i \xi_i + \eta_i)$, $q_i = (\rho_i/\rho)$ and the mean batch size m_i.

3. Stationary-interval approximation with a continuously busy server

We now determine the squared coefficient of variation of a stationary interval between departures in the departure process for one class assuming that the server is continuously busy. For this result, the model assumptions are much stronger than in section 2, but the exact results for these special cases can provide the basis for quite general approximations, as we indicated in section 1. We call the resulting squared coefficient of variation c_{d1}^2 the stationary-interval (SI) variability parameter for the limiting case of a continuously busy server.

As in [5], we only consider the two-class case. When there are more than two classes, we assume that all classes but the one of interest are aggregated into one. We assume that customers in the class of interest arrive in a batch-renewal process: Successive batches are i.i.d. with the batch sizes having mean m_1 and squared coefficient of variation c_{b1}^2; successive interarrival times of batches are also i.i.d., having mean $\hat{\lambda}_1^{-1}$ and squared coefficient of variation $\hat{c}_{a1}^2$. (The overall arrival rate is $\lambda_1 = \hat{\lambda}_1 m_1$ and the overall arrival variability parameter is c_{a1}^2. In general, c_{a1}^2 is not uniquely defined, but it is if the batch sizes have a geometric distribution, because that makes the arrival process a renewal process; see section 3.1.) The service times of class 1 are i.i.d. with mean τ_1 and squared coefficient of variation c_{s1}^2. The other class is assumed to be either a batch-Poisson (B-P) process or a batch-deterministic (B-D) process. Successive batches are i.i.d. with the batch sizes having mean m_2 and squared coefficient of variation c_{b2}^2; successive interarrival times of batches are also i.i.d. being exponentially distributed with mean $\hat{\lambda}_2^{-1}$ and squared coefficient 1 in the B-P case, and being constant with mean $\hat{\lambda}_2^{-1}$ and squared coefficient of variation 0 in the B-D case. (The overall class-2 arrival rate and variability parameter are $\lambda_2 = \hat{\lambda}_2 m_2$ and c_{a2}^2. When the batch-size distribution is geometric, c_{a2}^2 is well defined, but otherwise not.) The class-2 service times are i.i.d. with mean τ_2 and squared coefficient of variation c_{s2}^2. All the batch sizes, interarrival times and service times are assumed to be mutually independent.

In section 3.1 we indicate how to approximate a general arrival process partially specified by its arrival rate and variability parameter by these special batch processes. Then in sections 3.2 and 3.3 we calculate the SI variability parameter for the class-1 departure process, assuming that the class-2 arrival process is one of these special batch processes.

3.1. APPROXIMATING GENERAL PROCESSES BY THESE SPECIAL BATCH PROCESSES

The arrival process for the second class is quite special, being B-P or B-D, but we can treat more general processes for the second class by first approximating them by one of our special processes. Such approximations can be done in many ways; we suggest obtaining a specific approximation by working with geometric batch-size distributions. Let the batch size B be distributed as

$$P(B=k) = (1-p)\,p^{k-1}, \quad k = 1,2,\ldots. \tag{21}$$

The batch-size distribution thus has mean $m_2 = 1/(1-p)$ and squared coefficient of variation $p = (m_2 - 1)/m_2$. The geometric distribution is particularly useful because the associated B-P and B-D processes are then two-parameter renewal processes. The two parameters are the mean batch size m_2 (or, equivalently, p in (21)) and the mean of the interarrival time of batches $\hat{\lambda}_2^{-1}$. In each case, the overall arrival rate for the process is $\lambda_2 = \hat{\lambda}_2 m_2$. For the B-P process, the squared coefficient of variation of an interarrival time is $c_{a2}^2 = 2m_2 - 1$, which can assume any value greater than or equal to 1. For the B-D process, the squared coefficient of variation of an interarrival time is $c_{a2}^2 = m_2 - 1$, which can assume any value greater than or equal to 0.

Hence, given a general class-2 arrival process partially characterized by rate λ_2 and variability c_{a2}^2, we can approximate it by a renewal process with these same parameters. For any c_{a2}^2, we can use a B-D renewal process by setting $\lambda_2 = \hat{\lambda}_2 m_2$ and $(m_2 - 1) = c_{a2}^2$, i.e.

$$m_2 = c_{a2}^2 + 1 \quad \text{and} \quad \hat{\lambda}_2 = \lambda_2/m_2 = \lambda_2/(c_{a2}^2 + 1). \tag{22}$$

For any $c_{a2}^2 \geq 1$, we can use a B-P renewal process by setting $\lambda_2 = \hat{\lambda}_2 m_2$ and

$(2m_2 - 1) = c_{a2}^2$, i.e.

$$m_2 = (c_{a2}^2 + 1)/2 \quad \text{and} \quad \hat{\lambda}_2 = \lambda_2/m_2 = 2\lambda_2/(c_{a2}^2 + 1). \tag{23}$$

We recommend using a B-D process for $c_{a2}^2 < 1$ and a batch-Poisson process for $c_{a2}^2 \geq 1$, but a full process is not needed in sections 3.2 and 3.3.

3.2. WHEN THE SECOND CLASS IS BATCH-POISSON

In this section we assume that the class-2 arrival process is B-P. We determine the squared coefficient of variation of a stationary interval between successive class-1 departures, assuming that the server is continuously busy.

Given that the server is continuously busy, a stationary interval between departures of class-1 customers, say D_1, is one class-1 service time plus the sum of the class-2 service times of all class-2 customers to arrive during a class-1 inter-arrival time. As in [5], the Poisson property associated with the class-2 B-P process makes this class-1 interdeparture time well defined and relatively easy to analyze. Since the class-1 process is batch renewal, the class-1 interarrival time is of length 0 with probability $(m_1 - 1)/m_1$ and of positive length (having mean $\hat{\lambda}_1^{-1}$ and squared coefficient of variation $\hat{c}_{a1}^2$) with probability $1/m_1$. Since $c_{d1}^2 + 1 = E(D_1^2)/(ED_1)^2$, we obtain the desired variability parameter c_{d1}^2 from the first two moments of D_1.

THEOREM 2

If the server is continuously busy and the class-2 arrival process is B-P, then the first two moments of D_1 are

$$E(D_1) = \tau_1 + \frac{\hat{\lambda}_2 m_2 \tau_2}{\hat{\lambda}_1 m_1} = \frac{\rho}{\lambda_1} \tag{24}$$

and

$$\begin{aligned} E(D_1^2) &= (c_{s1}^2 + 1)\tau_1^2 + \frac{2\tau_1 \hat{\lambda}_2 m_2 \tau_2}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2 m_2 c_{s2}^2 \tau_2^2}{\hat{\lambda}_1 m_1} \\ &\quad + \frac{\hat{\lambda}_2 m_2^2 (c_{b2}^2 + 1)\tau_2^2}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2^2 m_2^2 (\hat{c}_{a1}^2 + 1)\tau_2^2}{\hat{\lambda}_1^2 m_1} \\ &= \tau_1^2 \left(c_{s1}^2 + \frac{(\rho + \rho_2)}{\rho_1} + \frac{\rho_2 \tau_2}{\rho_1 \tau_1}[c_{s2}^2 + m_2(c_{b2}^2 + 1)] + \frac{\rho_2^2}{\rho_1^2} m_1 (\hat{c}_{a1}^2 + 1)\right), \end{aligned} \tag{25}$$

so that

$$c_{d1}^2 = q_1^2 c_{s1}^2 + (1-q_1)^2 (p_1/p_2)[c_{s2}^2 + m_2(c_{b2}^2+1)] + (1-q_1)^2[m_1(\hat{c}_{a1}^2+1)-1], \tag{26}$$

where $q_1 = \rho_1/\rho$ and $p_1 = \lambda_1/\lambda$ as before.

Proof

Let U_1 be a class-1 batch interarrival time, ν_1 a class-1 service time, $\{\nu_{2n} : n \geq 1\}$ a sequence of i.i.d. service times for class 2, $\{B_{2n} : n \geq 1\}$ a sequence of i.i.d. batch sizes for class-2, and $\{N(t) : t \geq 0\}$ a Poisson process with rate $\hat{\lambda}_2$. With probability $(m_1-1)/m_1$, $D_1 = \nu_1$; with probability $1/m_1$, $D_1 = \nu_1 + \Sigma_{i=1}^{N_2} \nu_{2i}$ where $N_2 = \Sigma_{i=1}^{N(U_1)} B_{2i}$. Hence,

$$\begin{aligned}
E(N_2) &= \lambda_2(EU_1)(EB_{21}) = \hat{\lambda}_2 m_2/\hat{\lambda}_1, \\
\mathrm{Var}(N_2) &= E[N(U_1)]\,\mathrm{Var}(B_{21}) + \mathrm{Var}[N(U_1)(EB_{21})^2 \\
&= \frac{\hat{\lambda}_2 c_{b2}^2 m_2^2}{\hat{\lambda}_1} + \left[\frac{\hat{\lambda}_2}{\hat{\lambda}_1} + \frac{\hat{\lambda}_2^2 \hat{c}_{a1}^2}{\hat{\lambda}_1^2}\right] m_2^2 = \left(\frac{\hat{\lambda}_2(c_{b2}^2+1)}{\hat{\lambda}_1} + \frac{\hat{\lambda}_2 \hat{c}_{a1}^2}{\hat{\lambda}_1^2}\right) m_2^2, \\
E\left(\sum_{i=1}^{N_2} \nu_{2i}\right) &= \hat{\lambda}_2 m_2 \tau_2/\hat{\lambda}_1, \\
\mathrm{Var}\left(\sum_{i=1}^{N_2} \nu_{2i}\right) &= (EN_2)\,\mathrm{Var}(\nu_{21}) + \mathrm{Var}(N_2)(E\nu_{21})^2 \\
&= \frac{\hat{\lambda}_2 m_2 c_{s2}^2 \tau_2^2}{\hat{\lambda}_1} + \left[\frac{\hat{\lambda}_2(c_{b2}^2+1)}{\hat{\lambda}_1} + \frac{\hat{\lambda}_2^2 \hat{c}_{a1}^2}{\hat{\lambda}_1^2}\right] \tau_2^2 m_2^2, \\
E(D) &= \tau_1 + \hat{\lambda}_2 m_2 \tau_2/\hat{\lambda}_1 m_1, \\
E(D^2) &= \tau_1^2(c_{s1}^2+1) + \frac{2\hat{\lambda}_2 m_2 \tau_2 \tau_1}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2 m_2 c_{s2}^2 \tau_2^2}{\hat{\lambda}_1 m_1} \\
&\quad + \left[\frac{\hat{\lambda}_2(c_{b2}^2+1)}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2^2(c_{a1}^2+1)}{\hat{\lambda}_1^2 m_1}\right] \tau_2^2 m_2^2.
\end{aligned}$$

Remarks

(3.1) If the class-1 arrival process is characterized by the general parameters

λ_1 and c_{a1}^2, then we can obtain c_{d1}^2 from (26) by letting $m_1 = 1$ and replacing $\hat{c}_{a1}^2$ by c_{a1}^2. Then

$$c_{d1}^2 = q_1^2 c_{s1}^2 + (1-q_1)^2 (p_1/p_2)[c_{s2}^2 + m_2(c_{b2}^2 + 1)] + (1-q_1)^2 c_{a1}^2. \tag{27}$$

Furthermore, if the batch-size distribution for class-2 is geometric as in (21), then $c_{b2}^2 = (m_2 - 1)/m_2$ and $c_{a2}^2 = 2m_2 - 1 = m_2(c_{b2}^2 + 1)$, so that (27) becomes

$$c_{d1}^2 = q_1^2 c_{s1}^2 + (1-q_1)^2 (p_1/p_2)[c_{s2}^2 + c_{a2}^2] + (1-q_1)^2 c_{a1}^2. \tag{28}$$

Note that (28) is consistent with the AM approximation in (19) in the two-class case. This occurs because all the approximations now make the class-1 departure process assuming that the server is continuously busy a renewal process. (This is not difficult to prove using the lack of memory property associated with the Poisson process and the geometric distribution.)

If the class-1 and class-2 service times also have a common distribution, then (28) becomes

$$c_{d1}^2 = q_1 c_s^2 + q_1(1-q_1) c_{a2}^2 + (1-q_1)^2 c_{a1}^2, \tag{29}$$

which agrees with (5) because then $p_1 = q_1$ and $c_d^2 = c_s^2$.

(3.2) We obtain the desired formula for c_{n1}^2 in (2) directly from (27) by setting $\tau_1 = \tau_2 = 1$ and $c_{s1}^2 = c_{s2}^2 = 0$ (because all service times are identically 1). The general formula is

$$c_{n1}^2 = q_1(1-q_1)[m_2(c_{b2}^2 + 1)] + (1-q_1)^2 c_{a1}^2. \tag{30}$$

With geometric batch sizes for class 2, we apply (28) to obtain

$$c_{n1}^2 = q_1(1-q_1) c_{a2}^2 + (1-q_1)^2 c_{a1}^2, \tag{31}$$

which reduces to (4) because $p_1 = q_1$. Obviously (28) and (19) provide a simple modification to treat different service-time distributions. It seems intuitively reasonable that we should weight c_{a2}^2 more compared to c_{a1}^2 as τ_2/τ_1 increases. (Recall that $(1-q_1)^2 (p_1/p_2) = q_1(1-q_1)(\tau_2/\tau_1)$.) □

3.3. WHEN THE SECOND PROCESS IS BATCH-DETERMINISTIC

The general approximation formulas in (28) and (31) are easy to apply for all values of c_{a2}^2, but the B-P model of section 3.2 only applies to the case $c_{a2}^2 \geq 1$. To treat lower class-2 variability we consider the B-D process in this section. However, it turns out that the analysis here supports simply using (28) and (31) for all $c_{a2}^2 \geq 0$.

Given that the class 1 and 2 arrival processes are independent and stationary versions, the arrival point of an arbitrary class-1 batch is uniformly distributed over the deterministic interval between the arrival points of class-2 batches. Using this property, we can calculate the first two moments $E(D_1)$ and $E(D_1^2)$ exactly, given the distribution of U_1, the class-1 batch interarrival time. Thus, given $\hat{\lambda}_1$ and $\hat{c}_{a1}^2$, we can fit a distribution to them, as in section 3 of [31], and then calculate $E(D_1)$ and $E(D_1^2)$.

Instead, here we propose a simple approximation: We approximate the number of class-2 batches to arrive during U_1 by $\hat{\lambda}_2 U_1$. In particular, we use $\hat{\lambda}_2 EU_1 = \hat{\lambda}_2/\hat{\lambda}_1$ as its mean and $\hat{\lambda}_2^2(\hat{c}_{a1}^2+1)/\hat{\lambda}_1^2$ as its second moment. Of course, the approximate mean is exact, but the approximate second moment is not. With the notation in theorem 2 and its proof, under this approximating assumption we obtain

$$E(N_2) = \hat{\lambda}_2 m_2/\hat{\lambda}_1,$$

$$\operatorname{Var}(N_2) = \frac{\hat{\lambda}_2 c_{b2}^2 m_2^2}{\hat{\lambda}_1} + \left[\frac{\hat{\lambda}_2^2(\hat{c}_{a1}^2+1)}{\hat{\lambda}_1^2} - \frac{\hat{\lambda}_2^2}{\hat{\lambda}_1^2}\right] m_2^2$$

$$= \frac{\hat{\lambda}_2 c_{b2}^2 m_2^2}{\hat{\lambda}_1} + \frac{\hat{\lambda}_2^2 \hat{c}_{a1}^2 m_2^2}{\hat{\lambda}_1^2},$$

$$E\sum_{i=1}^{N_2} \nu_{2i} = \frac{\hat{\lambda}_2 m_2 \tau_2}{\hat{\lambda}_1},$$

$$\operatorname{Var}\left(\sum_{i=1}^{N_2} \nu_{2i}\right) = E(N_2)\operatorname{Var}(\nu_{21}) + \operatorname{Var}(N_2)(E\nu_{21})^2$$

$$= \frac{\hat{\lambda}_2 m_2 c_{s2}^2 \tau_2^2}{\hat{\lambda}_1} + \left[\frac{\hat{\lambda}_2 c_{b2}^2}{\hat{\lambda}_1} + \frac{\hat{\lambda}_2^2 \hat{c}_{a1}^2}{\hat{\lambda}_1^2}\right]\tau_2^2 m_2^2,$$

so that

$$E(D_1) = \frac{\hat{\lambda}_1 m_1 \tau_1 + \hat{\lambda}_2 m_2 \tau_2}{\hat{\lambda}_1 m_1} = \frac{\rho}{\lambda_1}, \tag{32}$$

$$E(D_1^2) = \tau_1^2(c_{s1}^2+1) + \frac{2\hat{\lambda}_2 m_2 \tau_2 \tau_1}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2 m_2 c_{s2}^2 \tau_2^2}{\hat{\lambda}_1 m_1} + \left[\frac{\hat{\lambda}_2 c_{b2}^2}{\hat{\lambda}_1 m_1} + \frac{\hat{\lambda}_2^2(\hat{c}_{a1}^2+1)}{\hat{\lambda}_1^2 m_1}\right]\tau_2^2 m_2^2$$

$$= \tau_1^2\left(c_{s1}^2 + \frac{(\rho+\rho_2)}{\rho_1} + \frac{\rho_2\tau_2}{\rho_1\tau_1}[c_{s2}^2 + m_2 c_{b2}^2] + \frac{\rho_2^2}{\rho_1^2} m_1(\hat{c}_{a1}^2+1)\right)$$

and

$$c_{d1}^2 = q_1^2 c_{s1}^2 + (1-q_1)^2 (p_1/p_2)[c_{s2}^2 + m_2 c_{b2}^2] + (1-q_1)^2 [m_1(\hat{c}_{a1}^2 + 1) - 1], \quad (33)$$

where $q_1 = \rho_1/\rho$.

If, as in remark 3.1, the class-1 arrival process is characterized by general parameters λ_1 and c_{a1}^2, then we can let $m_1 = 1$ and replace $\hat{c}_{a1}^2$ by c_{a1}^2 in (33). Furthermore, if the class-2 batch-size distribution is geometric as in (21), then $c_{b2}^2 = (m_2 - 1)/m_2$ and $c_{a2}^2 = m_2 - 1 = m_2 c_{b2}^2$, so that (33) agrees with (28). Thus, (19), (28) and (33) all support the same approximation.

4. Hybrid approximations

We now illustrate how the results for the continuously-busy limiting case in sections 2 and 3 can be used to construct heuristic hybrid approximations to cover the usual cases in which the queue is not continuously busy. Paralleling [1], the idea is to consider convex combinations that are consistent with established results in various limiting cases.

4.1. A DIRECT TWO-CLASS HYBRID APPROXIMATION

Let $c_{d1}^2(\rho)$ represent the approximate class-1 departure variability parameter as a function of the traffic intensity ρ; let c_{d1}^2 be the continuously-busy approximation in (28). (It helps that the two-class AM and SI approximations for c_{d1}^2 in (19), (28) and (33) agree.) A natural hybrid approximation based on the two-class case is

$$\begin{aligned} c_{d1}^2(\rho) &= \rho^2 c_{d1}^2 + (1-\rho^2) c_{a1}^2 \\ &= \rho^2 [q_1^2 c_{s1}^2 + (1-q_1)^2 (p_1/p_2)[c_{s2}^2 + c_{a2}^2] + (1-q_1)^2 c_{a1}^2] + (1-\rho^2) c_{a1}^2 \\ &= \rho_1^2 c_{s1}^2 + \rho_2^2 (p_1/p_2)[c_{s2}^2 + c_{a2}^2] + (1 - 2\rho_1\rho + \rho_1^2) c_{a1}^2, \end{aligned} \quad (34)$$

where c_{s2}^2 and c_{a2}^2 are aggregate variability parameters for all other classes when $k > 2$. Formula (34) was chosen because it satisfies certain limiting consistency conditions. For all ρ, as $q_1 \to 0$, $\rho_1 \to 0$ and $c_{d1}^2(\rho) \to c_{a1}^2$, which is consistent with [36]. For all ρ, as $q_1 \to 1$, $\rho_1 \to \rho$ and $c_{d1}^2(\rho) \to c_d^2(\rho)$, the one-class SI approximation in (6). For all q_1, as $\rho \to 1$, $\rho_1 \to q_1$ and $c_{d1}^2(\rho) \to c_{d1}^2$ in (19), (28) and (33), as it should because the server approaches being continuously busy. Finally, for all q_1, as $\rho \to 0$, $c_{d1}^2(\rho) \to c_{a1}^2$, which can be shown to be the appropriate pure light-traffic approximation.

4.2. A MULTI-CLASS HYBRID APPROXIMATION

When there are more than two classes, (34) involves aggregating all classes except the first in order to determine τ_2, c_{s2}^2 and c_{a2}^2. Instead of doing this aggregation, we can use (19). Then (34) becomes (13).

4.3. COMMON SERVICE-TIME DISTRIBUTIONS

In the special case of common service-time distributions, $\tau_1 = \tau_2 = \tau$ and $c_{s1}^2 = c_{s2}^2 = c_s^2$ so that (34) becomes

$$c_{d1}^2(\rho) = \rho_1(\rho - \rho_1)c_{a2}^2 + (1 - 2\rho_1\rho + \rho_1^2)c_{a1}^2 \tag{35}$$

and (13) becomes

$$c_{d1}^2(\rho) = \rho\rho_1 c_s^2 + \rho_1 \sum_{j=2}^{k} \rho_j \tilde{c}_{aj}^2 + (1 - 2\rho_1\rho + \rho_1^2)c_{a1}^2. \tag{36}$$

Moreover, (36) coincides with (35) when we use the asymptotic method (10) to approximate the aggregate variability parameter c_{a2}^2 in (35) in terms of the individual variability parameters $\tilde{c}_{aj}^2$, $2 \leq j \leq k$. Even with non-identical service-time distributions, (35) and (36) remain candidate approximations, using (7) and (8) of [32] to determine c_s^2. Of course, it remains to determine c_{a2}^2 when class-2 is an aggregate of other classes. The AM approximation is (10); the other natural simple alternative is the QNA hybrid $c_{a2}^2 = wc_{AM}^2 + 1 - w$ where c_{AM}^2 is (10) and the weight w comes from (29) and (30) of [32]. Formulas (13) and (35) coincide if c_{a2}^2 is the AM approximation in (10).

4.4. EXTENSION OF THE BITRAN–TIRUPATI APPROXIMATION

Another candidate approximation is obtained from (2), i.e. (6) of [5], using (6) and (31). Let

$$\begin{aligned} c_{d1}^2(\rho) &= p_1 c_d^2(\rho) + c_{n1}^2 \\ &= p_1[\rho^2 c_s^2 + (1-\rho^2)c_a^2] + c_{n1}^2 \\ &= p_1\rho^2 c_s^2 + p_1(1-\rho^2)c_a^2 + q_1(1-q_1)c_{a2}^2 + (1-q_1)^2 c_{a1}^2. \end{aligned} \tag{37}$$

(Note that p_1 and q_1 both appear in (37).) Formula (37) behaves the same as (34) if $p_1 \to 0$ and $q_1 \to 0$ or if $p_1 \to 1$ and $q_1 \to 1$, but behaves differently as $\rho \to 0$ and $\rho \to 1$. However, qualitatively (37) and (34) are quite similar. Note that (37) could be modified using (19), just as (34) was converted to (13).

With common service-time distributions, (37) reduces to (7). If, in addition, we express c_a^2 in terms of c_{a1}^2 and c_{a2}^2 using the AM approximation in (8), i.e. $c_a^2 = q_1 c_{a1}^2 + (1 - q_1)c_{a2}^2$, then (37) and (7) become (9) as noted in section 1.

4.5. THE CASE OF ZERO SERVICE TIMES

Another consistency condition to consider involves what happens as the mean service time for one class goes to zero. If $\tau_2 \to 0$ with λ_2 fixed, then for class-1 it is as if class-2 were not present. Consistent with this exact theoretical reference point, (34) approaches (6) for class-1 alone. However, (37) fails to satisfy this condition; it is smaller by the factor p_1, which could be arbitrarily small. Similarly, we can consider what happens when $\tau_1 \to 0$ with λ_1 fixed, but the exact behavior is more complicated; (34) approaches $c_{a1}^2 + \rho_2^2(p_1/p_2)(c_{a2}^2 + c_{s2}^2)$, while (7) is unchanged, except c_s^2 changes as $\tau_1 \to 0$, i.e. $c_s^2 \to p_2(c_{s2}^2 + 1) - 1$. Hence, (34) captures, at least qualitatively, the real explosion in variability that occurs as $\tau_1 \to 0$ and $p_1 \to 1$, while (7) and (37) do not. Thus, (34) is our proposed two-class procedure and (13) is our proposed full multi-class approximation.

4.6. SUMMARY

A summary of the candidate approximations for c_{d1}^2 discussed here appears in table 1. There are three procedures that work with all k classes and five procedures that work with only 2 classes (the class of interest plus the rest aggregated). The two

Table 1
A summary of candidate approximations for c_{d1}^2, the class-1 departure-process variability parameter.

k-class methods	
(11)	extension of INTl in [5] using (4) instead of (3), ignores class-dependent service times
(13)	based on (19) and (34), addresses class-dependent service times
(36) with c_s^2 via [32]	based on (19) and (34), but ignoring class-dependent service times
2-class methods	
(11) with $c_{ai}^2 = 1$	INTl from [5], ignores class-dependent service times
(7) with c_s^2, c_a^2, c_{a2}^2 calculated via [32]	extension of INTl in [5] using (4) instead of (3) and QNA superposition approximation instead of (8) and (10)
(34) with c_{a2}^2 and c_{s2}^2 via [32]	based on (28), addresses class-dependent service times
(35) with c_s^2 and c_{a2}^2 via [32]	based on (28) and (34), but ignores class-dependent service times
(37) with c_{a2}^2 and c_{s2}^2 via [32]	based on (2), (6) and (31), only partially addresses class-dependent service times via c_{n1}^2 in (31)

Erlang-based two-class procedures INT2 and INT3 from [5] are not included in this list. All procedures in table 1 have c_{d1}^2 a linear function of the arrival variability parameters c_{aj}^2, so that for an open network of queues the net arrival-process parameters can be obtained by solving a system of linear equations. Moreover, it is easy to see that this system of equations always has a unique solution.

5. The Bitran–Tirupati experiments with common service-time distributions

In this section we compare our approximations for $c_{d1}^2(\rho)$ with the approximations and simulation results of Bitran and Tirupati [5]. Throughout this section, we follow [5] and assume that all classes have a common service-time distribution. Our leading candidates are (35) and (36) which are special cases of (34) and (13). (These are our first choices because they satisfy all the consistency conditions.) Our third candidate is (7) which coincides with (37), and is based on (2), (4), (6) and (31). Our fourth candidate is (11).

Variants of candidate approximations (35) and (7) are obtained depending on how we approximate the variability parameters c_a^2 and c_{a2}^2 for the respective superposition arrival processes. The AM approximations for c_a^2 and c_{a2}^2 are given in (8) and (10). (It was already noted that the AM approximation converts (7) into (11).) The QNA hybrid is the convex combination $wc_{AM}^2 + 1 - w$ where c_{AM}^2 is the AM approximation and the weight w comes from (29) and (30) of [32]. An SI approximation and other hybrids are discussed in [1] and [31]. We only consider the QNA hybrid approximation for superposition processes here (and the AM via (11)).

The first experiments from [5] that we consider involve two or more i.i.d. arrival processes. Consequently, c_{ai}^2 is the same for all classes. In this case (but not more generally), (11) and (36) coincide, both reducing to

$$c_{d1}^2(\rho) = \rho_1 \rho c_s^2 + (1 - \rho_1 \rho) c_{a1}^2. \tag{38}$$

There thus remain three new candidate approximations: (7), (11) and (35). Approximations (7) and (35) involve applying the superposition approximations in [32] to c_a^2 and c_{a2}^2, and c_{a2}^2, respectively. The QNA hybrid approximations for c_{a2}^2 and c_a^2 do not agree. Since the component streams are i.i.d., the effective number of streams ν in (30) of [32] is just the actual number of streams. For c_{a2}^2, this is obviously one less than for c_a^2. Following [5], we consider the cases $\nu = 2, 3, 5$ and 10. For treating c_{a2}^2, we thus need to consider 1, 2, 4 and 9. The resulting weights w for (29) of [32] and approximate variability parameters for the six cases involving $c_{a1}^2 = 0.500, 0.333, 0.250$ and $\rho = 0.6, 0.9$ are given in table 2. As can be seen from table 2, the QNA hybrid recognizes the tendency for superposition processes to converge to Poisson processes as the number of streams increases: c_a^2 and c_{a2}^2 increase toward 1 as ν increases. The limiting case in which c_{a2}^2 is replaced by 1 was used as Bitran and Tirupati to obtain (3).

Table 2
Approximate variability parameters for superposition arrival processes via QNA: $wc_{AM}^2 + 1 - w$ with the weight w coming from (29) and (30) of [32].

	$\rho = 0.9$				$\rho = 0.6$			
Number of streams	Weight	Single-stream c_{a1}^2			Weight	Single-stream c_{a1}^2		
ν	w	0.500	0.333	0.250	w	0.500	0.333	0.250
1	1.000	0.500	0.333	0.250	1.000	0.500	0.333	0.250
2	0.962	0.519	0.359	0.279	0.610	0.695	0.593	0.543
3	0.926	0.537	0.382	0.306	0.439	0.781	0.707	0.671
4	0.893	0.554	0.405	0.330	0.342	0.829	0.772	0.744
5	0.862	0.569	0.425	0.354	0.281	0.860	0.813	0.789
9	0.758	0.621	0.494	0.431	0.163	0.919	0.891	0.878
10	0.735	0.632	0.510	0.449	0.148	0.926	0.901	0.889

Table 3
Approximate single-class departure-process variability parameters c_{d1}^2 for one multi-class queue with independent and identically distributed component streams and service times: the case of $\rho = 0.9$ and $c_s^2 = 0.333$ (cf. table 1 of [5]).

Number of component streams (products)	One arrival stream c_{a1}^2	Aggregate parameters from QNA [32]: Arrival hybrids $\hat{c}_{a2}^2$	Arrival hybrids c_a^2	Departures c_d^2	Variability parameter c_{d1}^2: New (11), (36) and (38)	QNA hybrids (7)	QNA hybrids (35)	From Bitran–Tirupati [5]: INT1	INT3	Simulation
$\nu = 2$	0.500	0.500	0.519	0.369	0.433	0.434	0.433	0.559	0.460	0.475
$(p_1 = 0.5)$	0.333	0.333	0.359	0.338	0.333	0.336	0.333	0.502	0.369	0.373
	0.250	0.250	0.279	0.323	0.284	0.287	0.284	0.474	0.323	0.328
$\nu = 3$	0.500	0.519	0.537	0.372	0.455	0.461	0.458	0.568	0.469	0.486
$(p_1 = 0.333)$	0.333	0.359	0.382	0.342	0.333	0.342	0.338	0.484	0.351	0.37
	0.250	0.279	0.306	0.328	0.273	0.282	0.278	0.442	0.290	0.303
$\nu = 5$	0.500	0.554	0.569	0.378	0.473	0.484	0.480	0.553	0.496	0.522
$(p_1 = 0.2)$	0.333	0.405	0.425	0.351	0.333	0.348	0.343	0.440	0.340	0.361
	0.250	0.330	0.354	0.337	0.264	0.280	0.274	0.394	0.270	0.287
$\nu = 10$	0.500	0.621	0.632	0.390	0.487	0.500	0.495	0.532	0.488	0.515
$(p_1 = 0.1)$	0.333	0.494	0.510	0.367	0.333	0.351	0.345	0.393	0.334	0.355
	0.250	0.431	0.449	0.355	0.257	0.277	0.270	0.324	0.259	0.279

Table 4
Approximate single-class departure-process variability parameters c_{d1}^2 for one multi-class queue with independent and identically distributed component streams and service times: the case of $\rho = 0.6$ and $c_s^2 = 0.333$ (cf. table 3 of [5]).

Number of component streams (products)	One arrival stream c_{a1}^2	Aggregate departure from [32] c_d^2	Variability parameter c_{d1}^2					
			New					
			(11), (36) and (38)	QNA hybrids		From Bitran–Tirupati [5]		
				(7)	(35)	INT1	INT3	Simulation
$\nu = 2$	0.500	0.452	0.470	0.476	0.470	0.595	0.498	0.500
$(p_1 = 0.5)$	0.333	0.350	0.333	0.342	0.333	0.499	0.369	0.371
	0.250	0.299	0.266	0.275	0.266	0.453	0.304	0.303
$\nu = 3$	0.500	0.464	0.480	0.492	0.496	0.591	0.493	0.507
$(p_1 = 0.333)$	0.333	0.364	0.333	0.349	0.353	0.481	0.351	0.377
	0.250	0.316	0.260	0.278	0.283	0.427	0.277	0.287
$\nu = 5$	0.500	0.484	0.488	0.506	0.507	0.568	0.493	0.510
$(p_1 = 0.2)$	0.333	0.392	0.333	0.357	0.358	0.44	0.340	0.360
	0.250	0.347	0.256	0.282	0.287	0.376	0.262	0.276
$\nu = 10$	0.500	0.524	0.494	0.513	0.508	0.539	0.495	0.507
$(p_1 = 0.1)$	0.333	0.446	0.333	0.359	0.352	0.393	0.334	0.354
	0.250	0.407	0.253	0.282	0.283	0.32	0.255	0.274

The three approximations for $c_{d1}^2(\rho)$, (38) and the QNA hybrids plus (7) and (35), are compared to simulation and the Bitran–Tirupati approximations (INT1 and INT3 from [5]) in tables 3 and 4. All these approximations perform reasonably well (much better than a direct application of [32], as shown in [5]). The most elementary approximations are (38) and INT1; (38) is better for small numbers ν of component arrival processes, but INT1 improves as ν increases, reflecting the convergence to Poisson. The two QNA hybrids perform essentially the same, both being somewhat better than (38) and INT1. The performance of the QNA hybrids is roughly comparable to INT3; however, the QNA hybrids may be preferred because they are more elementary and generalize to other cases.

It is of course of interest to see how these departure-process approximations perform when the departure process serves as an arrival process to a subsequent queue. Even a perfect match of $c_{d1}^2(\rho)$ with simulation does not guarantee good congestion approximations because the parameter $c_{d1}^2(\rho)$ only partially characterizes the departure process. Moreover, the departure process is typically not renewal. However, experience indicates that good congestion approximations usually require $c_{d1}^2(\rho)$ to be close to the actual value [33], so that the comparisons in tables 3 and 4 are meaningful. To illustrate how the approximations apply to the congestion measures, we consider one case from [5], let the number of arrival processes be 5, $c_{a1}^2 = c_s^2 = 0.333$, and $\rho = 0.6$ (the eighth row of table 4). Let the departure process of each class be routed to a separate single-server queue with i.i.d. Erlang service times ($c_s^2 = 0.333$) and traffic intensity 0.8. The observed simulation average number of customers in one of these queues was 1.79. Using the approximation (45) and (47) of [32] the approximate values by (7), (35), (38), INT3 and INT1 are, respectively 1.78, 1.78, 1.75, 1.77 and 1.96. In contrast, simple $M/M/1$ and $M/G/1$ approximations are 4.00 and 2.93, respectively.

We also consider a second experiment from [5]. There are two arrival processes with the arrival-rate proportion $p_1 = \lambda_1/\lambda = j/10$, $1 \le j \le 9$. Let all the arrival and service variability parameters be 0.333 and let $\rho = 0.6$. Since there are only two streams, c_{a2}^2 does not require aggregation and (35) coincides with (38). Moreover, since $c_s^2 = c_{a1}^2 = c_{a2}^2 = 0.333$, by these methods $c_{d1}^2(\rho) = 0.333$ for all ρ and p_1. However, for the QNA hybrid based on (7), c_a^2 must be calculated. Since the streams have unequal intensity (except in the case $p_1 = 0.5$), the equivalent number of streams ν from (30) of [32] is less than two. The calculations for the QNA approximation of c_a^2 appear in table 5 together with the various approximations for $c_{d1}^2(\rho)$. The approximations perform reasonably well, but are not exceptionally accurate, having relative errors of about 5–20%. As noted in [5], these approximations evidently perform better as the number of component streams increases (unlike (1) when there is deterministic routing).

Table 5

Comparison of approximations for the single-class departure process variability parameter: the case of two component arrival processes with proportions p_1 and $(1-p_1)$, $c_{a1}^2 = c_{a2}^2 = c_s^2 = 0.333$, and $\rho = 0.6$. (cf. table 7 of [5]).

	From [32]				New		From Bitran–Tirupati [5]		
	Equivalent number of streams	Hybrid weight	Aggregate variability parameters						
p_1	ν	w	c_a^2	c_d^2	(11), (35) (36) and (38)	(7)	Simulation (Erlang)	INT3	INT1
0.1	1.22	0.876	0.416	0.386	0.333	0.339	0.351	0.335	0.393
0.2	1.47	0.769	0.487	0.432	0.333	0.353	0.352	0.340	0.439
0.3	1.72	0.685	0.543	0.468	0.333	0.374	0.362	0.346	0.473
0.4	1.92	0.629	0.581	0.492	0.333	0.397	0.358	0.356	0.493
0.5	2.00	0.610	0.593	0.500	0.333	0.417	0.372	0.370	0.500
0.6	1.92	0.629	0.581	0.492	0.333	0.428	0.381	0.383	0.493
0.7	1.72	0.685	0.543	0.468	0.333	0.428	0.401	0.394	0.473
0.8	1.47	0.769	0.487	0.432	0.333	0.412	0.398	0.397	0.440
0.9	1.22	0.876	0.416	0.386	0.333	0.381	0.403	0.383	0.393

6. Conclusions

In sections 2 and 3 we presented theoretical results characterizing the departure processes of individual customer classes from multi-class queues under the assumption that the server is continuously busy. As noted in remark 2.5, the AM result also applies to multi-server queues. Obviously, these results can be used to describe the queue in the special limiting case, in which the server is almost always continuously busy, but they also can be used to develop hybrid approximations for more general cases. In section 4 we proposed relatively simple hybrid approximations based on our theoretical results, especially (13) and (34), and in section 5 we showed that these hybrid approximations perform reasonably well when compared to simulations. Overall we have established a basis for improvements in the parametric-decomposition method for approximating open queueing networks. It remains to refine the approximations and do more extensive experiments. This is intended for a future paper. Experiments are especially needed in the case of class-dependent service times. One such experiment is described in section 3 of [36].

The approximations developed here and in [5] offer significant improvements over the random splitting formula (1) when the routing is deterministic. Conversely, when the routing is primarily random, (1) and [32] are preferred. Of course, in many realistic networks both random routing and deterministic routing are present, so that it is appropriate to account for both kinds of routing in the network analysis. A hybrid routing approximation has been implemented in [24]. Further work in this direction seems worthwhile.

References

[1] S.L. Albin, Approximating a point process by a renewal process, II: Superposition arrival processes to queues, Oper. Res. 32 (1984) 1133–1162.
[2] S.L. Albin, Delays for customers from different arrival streams to a queue, Manag. Sci. 32 (1986) 329–340.
[3] S.L. Albin and S. Kai, Approximation for the departure process of a queue in a network, Nav. Res. Log. Qtrly. 33 (1986) 129–143.
[4] P. Billingsley, *Convergence of Probability Measures* (Wiley, New York, 1968).
[5] G.R. Bitran and D. Tirupati, Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference, Manag. Sci. 34 (1988) 75–100.
[6] J.A. Buzacott and J.G. Shanthikumar, On approximate queueing models of dynamic job shops, Manag. Sci. 31 (1985) 347–366.
[7] K.M. Chandy and C.H. Sauer, Approximate methods for analyzing queueing network models of computer systems, Comp. Surveys 10 (1978) 281–317.
[8] J.G. Dai, V. Nguyen and M.I. Reiman, Sequential bottleneck decomposition: An approximation method for open queueing networks, Oper. Res., to appear.
[9] K.W. Fendick, V.R. Saksena and W. Whitt, Dependence in packet queues, IEEE Trans. Commun. COM-37 (1989) 1173–1183.
[10] K.W. Fendick, V.R. Saksena and W. Whitt, Investigating dependence in packet queues with the index of dispersion for work, IEEE Trans. Commun. COM-39 (1991) 1231–1244.
[11] W. Fischer and D. Stanford, Approximations for the per-class waiting time and interdeparture time in the $\Sigma_i GI_i/GI_i/1$ queue, Siemens, München (1989).

[12] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems* (Academic Press, New York, 1980).

[13] J.M. Harrison and V. Nguyen, The QNET method for two-moment analysis of open queueing networks, Queueing Syst. 6 (1990) 1–32.

[14] J.M. Harrison and V. Nguyen, Brownian models of multiclass queueing networks: Current status and open problems, Queueing Syst. 13 (1993) 5–40.

[15] J.M. Holtzman, Mean delays of individual streams into a queue: The $\Sigma G_i/M/1$ queue, *Applied Probability-Computer Science: The Interface*, eds. R.L. Disney and T.J. Ott (Birkhäuser, Boston, 1982) pp. 417–430.

[16] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I, Adv. Appl. Prob. 2 (1970) 150–177.

[17] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, II: Sequences, networks and batches, Adv. Appl. Prob. 2 (1970) 355–369.

[18] L. Kleinrock, *Communication Nets* (McGraw–Hill, New York, 1964).

[19] W. Kraemer and M. Langenbach-Belz, Approximate formulae for the delay in the queueing system $GI/G/1$, *Proc. 8th Int. Teletraffic Congress*, Melbourne (1976).

[20] P.J. Kuehn, Approximate analysis of general queueing networks by decomposition, IEEE Trans. Commun. COM-27 (1979) 113–126.

[21] B. Pourbabai and D. Sonderman, Approximation of departure process from a $G/M/1/0$ queueing system, Oper. Res. Lett. 4 (1986) 201–205.

[22] M.I. Reiman, Asymptotically exact decomposition approximations for open queueing networks, Oper. Res. Lett. 9 (1990) 363–370.

[23] M. Reiser and H. Kobayashi, Accuracy of the diffusion approximation for some queueing systems, IBM J. Res. Dev. 18 (1974) 110–124.

[24] M. Segal and W. Whitt, A queueing network analyser for manufacturing, *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12*, ed. M. Bonatti (Elsevier Science, Amsterdam, 1989) pp. 1146–1152.

[25] K.C. Sevcik, A.I. Levy, S.K. Tripathi and J.L. Zahorjan, Improving approximations of aggregated queueing network subsystems, *Computer Performance*, eds. K.M. Chandy and M. Reiser (North-Holland, Amsterdam, 1977) pp. 1–22.

[26] J.G. Shanthikumar and J.A. Buzacott, Open queueing network models of dynamic job shops, Int. J. Prod. Res. 19 (1981) 255–266.

[27] D.A. Stanford and W. Fischer, The interdeparture-time distribution for each class in the $\Sigma M_i/G_i/1$ queue, Queueing Syst. 4 (1989) 177–190.

[28] D.A. Stanford and W. Fischer, Characterizing interdeparture times for bursty input streams in the queue with pooled renewal arrivals, Stoch. Mod. 7 (1991) 311–320.

[29] S. Suresh and W. Whitt, The heavy-traffic bottleneck phenomenon in open queueing networks, Oper. Res. Lett. 9 (1990) 355–362.

[30] W. Whitt, Some useful functions for functional limit theorems, Math. Oper. Res. 5 (1980) 67–85.

[31] W. Whitt, Approximating a point process by a renewal process, I: Two basic methods, Oper. Res. 30 (1982) 125–147.

[32] W. Whitt, The queueing network analyzer, Bell Syst. Tech. J. 62 (1983) 2779–2815.

[33] W. Whitt, Approximations for departure processes and queues in series, Nav. Res. Log. Qtrly. 31 (1984) 499–521.

[34] W. Whitt, Approximations for the $GI/G/m$ queue, AT&T Bell Laboratories, Murray Hill, NJ (1985).

[35] W. Whitt, Approximations for single-class departure processes from multi-class queues, AT&T Bell Laboratories, Murray Hill, NJ (1987).

[36] W. Whitt, A light-traffic approximation for single-class departure processes from multi-class queues, Manage. Sci. 34 (1988) 1333–1346.

[37] W. Whitt, Large fluctuations in a deterministic multiclass network of queues, Manag. Sci. 39 (1993), to appear.