A Light-Traffic Approximation For Single-Class Departure Pr
Whitt, Ward
Management Science; Nov 1988; 34, 11; ProQuest
pg. 1333

MANAGEMENT SCIENCE
Vol. 34, No. 11, November 1988
Printed in U.S.A.

A LIGHT-TRAFFIC APPROXIMATION FOR SINGLE-CLASS DEPARTURE PROCESSES FROM MULTI-CLASS QUEUES*

WARD WHITT
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

This paper discusses an approximation for single-class departure processes from multi-class queues: If the arrival rate of one class upon one visit to the queue is a small proportion of the total arrival rate there, then the departure process for that class from that visit should be nearly the same as the arrival process for that class for that visit. This can be regarded as a light-traffic approximation, but only the one class must be in light traffic; the overall traffic intensity of the queue need not be low. As a consequence, in a queueing network if the routing for one class is deterministic, and if the light-traffic condition applies at every queue this class visits, then the arrival and departure processes for this class at each visit to each queue should be nearly the same as its external arrival process. This approximation is explained in terms of different time scales, and is justified here by a limit theorem in a special case. There are important implications for parametric decomposition approximation techniques: the variability parameter partially characterizing the departure process at any visit to any queue of such a low-intensity class should be nearly the same as the variability parameter partially characterizing the arrival process for that class at that visit to that queue. The approximation principle in this form was recently proposed by G. Bitran and D. Tirupati while developing improved parametric-decomposition approximations for low-variability multi-class queueing networks with deterministic routing, which have important applications in manufacturing. The approximation principle also has important implications for data networks, showing how burstiness in originated traffic can pass through heavily shared network facilities where it has relatively little effect and then reappear at the destination.
(MULTI-CLASS QUEUES; QUEUEING NETWORKS; APPROXIMATIONS; LIGHT TRAFFIC; DEPARTURE PROCESSES)

1. Introduction and Summary

Many communication, computer and manufacturing systems can be modeled as multi-class queueing networks. However, realistic models often are very large and complex, having many queues and many classes. Consequently, it often is very difficult to analyze these models in detail. Of course, greater computer power and more powerful analytical techniques will eventually enable us to analyze the performance of more complex models in microscopic detail. However, just as in statistical mechanics, we should also look for new statistical regularity emerging from a macroscopic view of the increased complexity; we may well find useful simple approximations.

1.1 *Statistical Regularity Emerging from Increased Complexity*

Examples of statistical regularity emerging from increased complexity in queueing are the ways large closed queueing network models (with many queues or many customers) behave like more elementary open models as the size increases; see Whitt (1984a). The statistical regularity often involves a decoupling of the network (the separate queues become independent), sometimes subject to a relatively simple constraint such as a system of equations, as occurs with reduced-load or Erlang-fixed-point approximations, see Kelly (1986), Whitt (1985a) and references there. Macroscopic statistical regularity is also associated with the many heavy-traffic diffusion approximations, e.g., Iglehart and Whitt (1970), Reiman (1984) and Harrison (1988).

* Accepted by Linda Green; received June 22, 1987. This paper has been with the author 3 weeks for 1 revision.

0025-1909/88/3411/1333$01.25
Copyright © 1988, The Institute of Management Sciences

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Perhaps the most familiar example of statistical regularity emerging from increased complexity in queueing occurs with superposition arrival processes, when an arrival process is composed of many independent streams each of very small intensity. For example, the overall arrival process of calls at a telephone exchange is composed of the relatively infrequent requests from many separate subscribers. It is well known that under very general conditions the overall superposition arrival process can be well approximated by a simple Poisson process, even if the component streams are not nearly Poisson or renewal; in fact, the quality of the approximation tends to improve as the complexity (number of component streams) increases; see Palm (1943), Khintchine (1960), Franken (1963), and Çinlar (1972). (See Albin 1982, 1984, Newell 1984, Whitt 1985b, Sriram and Whitt 1986, and Fendick, Saksena and Whitt 1988a, b for recent investigations of the applicability of this approximation in queueing when the number of component streams is not exceptionally large.)

Another justification for assuming Poisson arrival processes to queues in complex queueing networks comes from random routing. If each departure from some queue is routed to another queue with probability p, independently of everything else up to that time, then the resulting flow from the first queue to the second is approximately Poisson if p is sufficiently small, under minimal conditions on the departure process from the first queue. Moreover, the same result holds when the independence is relaxed in various ways. These Poisson approximations are justified by limit theorems for thinning and partitioning of point processes; see Serfozo (1977, 1985), Böker and Serfozo (1983) and references there. (From these theorems, we see that the departure process only needs to satisfy a weak law of large numbers.)

Consistent with the superposition and thinning limit theorems discussed above, conventional wisdom concludes that it is often reasonable to model complex open queueing networks by Markovian Jackson queueing network models and their product-form relatives. In particular, it is generally considered reasonable to assume that the steady-state numbers of customers at the queues are independent, each having the same distribution as if the arrival process to that queue were a Poisson process. With multiple classes, the class identity of each arrival can be chosen by independent trials or, equivalently, the arrival processes of the classes can be regarded as independent Poisson processes. Note that the superposition and thinning limit theorems help justify independence, in both the time-dependent behavior, and the product-form steady-state distribution, as well as the Poisson distribution. With enough random thinning or superposition, the arrival processes will be nearly Poisson for any service discipline and any service-time distributions at the queues. Hence, there is a very strong theoretical basis for analyzing large queueing networks as a collection of independent queues with Poisson arrival processes.

A primary purpose of this paper is to point out an important limitation to this conventional wisdom for queueing networks when the routing is *deterministic* instead of random. By deterministic routing, we mean that the customer classes follow specified routes through the network; i.e., the customers from each class visit a specified sequence of queues. In a route for one class, queues can appear in this sequence more than once or not at all. The customers from one class all enter the network at the first queue on their route and leave the network after completing service at the last queue on their route. In many applications the routing is indeed primarily deterministic. In many data networks, the data travel over virtual circuits. In manufacturing, individual products usually have a well defined sequence of operations.

1.2 *A Light-Traffic Approximation for Networks with Deterministic Routing*

In this paper, we examine queueing networks with deterministic routing. As with the thinning associated with random routing, we assume that the proportion p of customers served at each queue that belong to the designated class is very small, and we focus on

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the departure process associated with this individual class. With deterministic routing, it turns out that this thinned flow is *not* in general nearly Poisson, but as before there is useful statistical regularity emerging from the increased complexity. The new statistical regularity is based on a light-traffic limit theorem for the departure processes of individual classes. However, unlike many recent light-traffic results, light traffic here does not mean that the overall traffic intensity is becoming negligible, but that the contribution of one class to the whole is becoming negligible.

One version of the approximation principle can be stated in rough general terms as follows:

Light-Traffic Approximation Principle for Single-Class Departure Processes from Multi-Class Queues. If the arrival rate of one class upon one visit to some queue is a small proportion of the total arrival rate there, then the departure process for that class from that visit to that queue tends to be nearly the same as the arrival process for that class for that visit to that queue.

This approximation principle for single-class departure processes from individual multi-class queues in turn yields an associated approximation principle for multi-class queueing networks with deterministic routing:

Light-Traffic Approximation Principle for Multi-Class Queueing Networks with Deterministic Routing. Consider a class that follows a fixed deterministic route through the network. If the contribution to the arrival rate by this class at each visit to each queue is a small proportion of the total arrival rate at that queue, then the departure process of that class from each visit to each queue, and thus from the entire network, is nearly the same as the external arrival process of this class to the network.

The network approximation principle also applies to sojourns in subnetworks, because the network under discussion could be a subnetwork of a larger network.

The two approximation principles have been stated without carefully defining the conclusions, i.e., without explaining what is meant by the departure process being nearly the same as the arrival process. In fact, we mean in the strongest possible sense: sample path by sample path, i.e., with probability one (w.p.1). As a consequence, of course, this implies that the distribution of the entire departure process (as a stochastic process) is nearly the same as the distribution of the entire arrival process. As a further consequence, it means that the stationary distribution of the interdeparture times is nearly the same as the stationary distribution of the interarrival times. However, the w.p.1 conclusion is much stronger; e.g., each exceptionally long interarrival time occurring by chance in one sample path of the arrival process will tend to have a counterpart in the corresponding departure process sample path.

The two approximation principles have also been stated without specifying conditions on the numbers of servers, the waiting space, the service disciplines, the service-time distributions or even much about the arrival processes. Of course, some conditions are needed, but in fact the conditions are very general, so that the statements above without these extra conditions capture the spirit of the truth. There is one important condition worth mentioning, however. Both the arrival rate and the offered load (arrival rate times the mean service time) for the class of interest must be small. This is achieved by simply letting the arrival rate get small, but the service time cannot be allowed to grow at the same time; the offered load of this class must also be small.

In fact, we do not rigorously establish these light-traffic approximations in their full generality. We obtain a fairly nice result with simple sufficient conditions (Proposition 3) only for the case of a single queue without feedback, i.e., with exogenous arrival processes.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.3 *An Explanation: Different Time Scales*

Upon reflection, the approximation principles are easily understood, and soon seem quite trivial. Nevertheless, we contend that they are worth expressing carefully because they are vital theoretical reference points for understanding the behavior of complex queueing networks.

A proper understanding of this light-traffic approximation follows from focusing on the relevant time scales. The light-traffic condition amounts to assuming that the time scale for arrivals and departures from the designated class is much longer than the time scale for the aggregate of other classes at the queue. Arrivals and departures occur much more slowly for customers of the designated class than arrivals and departures occur for the aggregate of all classes at the queue. For example, interarrival times for the designated class might be in seconds, while interarrival times and service times at the queue are in milliseconds. The service times for the designated class are also much shorter than the interarrival times for the class. Therefore, the delay and service time experienced by a customer from the designated class are in the short time scale, so that they are negligible compared to the interarrival times. Thus, in the time scale of the arrival epochs of the designated class, the departure epochs are nearly the same as the arrival epochs. This is what we meant by trivial: The low-intensity assumption is tantamount to assuming that the service times and delays are negligible compared to the interarrival times. When a slow arrival process comes to a fast service facility, it is indeed obvious that the departure process is nearly the same as the arrival process. From the perspective of the individual class, it is almost as if the service facility were not there.

Once this low-intensity approximation principle is understood, it is natural to ask if it could possibly be practically meaningful. In fact, the principle becomes more and more meaningful as the size and complexity of physical systems grow. For example, in data communication the speed of switching and transmission is growing rapidly. Already it is common to have great sharing of resources and great differences in time scales. Characters often are transmitted from individual terminals in the time scale of seconds, while packets arrive and are processed in a packet network facility in the time scale of milliseconds. (We give an example in §3.) Even greater differences in time scales will be common in the data communication in the near future.

Similarly, manufacturing systems are increasing in size and complexity. Today it is common to have manufacturing lines producing hundreds of products, each with their own sequence of operations. The low runners (less frequently produced products) might be started once or twice a month, while the interarrival times and processing times on individual machines are in minutes. Thus, in the time scale of the low runners (weeks or months), the product completion times are nearly the same as the product start times. Moreover, even when the time scales are not exceptionally far apart, this extreme case helps provide insight into the behavior of queueing networks. For further discussion about different time scales in queueing, see pp. 536, 543 of Whitt (1984b) and pp. 835, 842 of Sriram and Whitt (1986).

It is of course obvious that the deterministic routing considered here violates the conditions of the random thinning and partitioning theorems in Serfozo (1985) and its references, so that we should perhaps not be surprised that the system behavior is quite different. On the other hand, on pp. 281–282 of Serfozo (1985), communications networks and manufacturing lines are cited as examples of potential applications of the theorems, without any caveat that the actual behavior of these systems might be radically different than the theory predicts because of the deterministic routing. A primary purpose of this paper is to make this caveat.

It is important to realize that indeed deterministic routing can make a great difference. Of course, if the external single-class arrival process is nearly Poisson, then the result is

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

essentially the same as with random routing. (Experience with such systems in the past may be the cause of current misperceptions.) However, if the external single-class arrival process is not nearly Poisson, neither will be the single-class departure process. The practical importance of the light-traffic approximation principle, then, stems from the fact that in many applications the single-class external arrival processes are indeed not nearly Poisson. As Bitran and Tirupati (1988) observe, in manufacturing the single-product external arrival processes are often substantially less variable than Poisson. In packet communications networks the single-source packet arrival processes are often substantially more variable than Poisson; see Sriram and Whitt (1986), Fendick et al. (1988a, b) and §3 here.

1.4 *Implications for Parametric-Decomposition Approximations*

The light-traffic approximation principle above has important implications for parametric-decomposition approximation methods, such as are used in the Queueing Network Analyzer (Whitt 1983, 1987b and Segal and Whitt 1988). In that context, it specifically suggests that (under the assumptions above) the variability parameter partially characterizing the single-class departure process should be nearly the same as the variability parameter partially characterizing the external arrival process of that class. The approximation principle in this more restricted form was suggested by Bitran and Tirupati (1988), in the process of developing an improved parametric-decomposition approximation for multi-class open queueing networks with deterministic routing. Bitran and Tirupati show that parametric-decomposition combined with aggregation as in Whitt (1983) can perform quite poorly with deterministic routing, so that improvement is needed. Moreover, if a multi-class open queueing network has a very large number of classes, each with relatively small arrival rate at every queue, then the light-traffic decomposition principle indicates that for the variability parameters we can completely ignore the dependence among the queues and analyze the individual queues separately. For such networks, we can determine both the net arrival rates and variability parameters at the queues without solving any system of equations, as in §§4.1 and 4.2 of Whitt (1983). In these cases the resulting approximation is obtained with great ease and there is virtually no constraint on the size of the model that can be analyzed. As with the other approximations emerging from increasing complexity, these should perform better as the model grows and becomes more complex.

Further work on parametric-decomposition approximations is contained in Albin (1984, 1986), Albin and Kai (1986), Fendick et al. (1988a, b), Holtzman (1982), Whitt (1987a, b) and references there.

1.5 *Connections to the Superposition Approximation*

At first glance, it may seem as if the new approximations for the congestion at the queues should be covered by the Poisson approximation associated with the previously discussed superposition limit theorem, but this is not the case. Note that only the class of interest must have relatively small arrival rates; other classes might have nonnegligible arrival rates; e.g., there might be only one other class. Then the overall superposition process need not be remotely close to Poisson.

Even when there are many classes at the queue and all classes have small arrival rates, the new approximation principle has something important to say. Of course, in this case the overall arrival process at the queue can be approximated by a Poisson process. Moreover, since the arrival rate of the class of interest is very small, that class can be regarded as outside observers that do not interact significantly with the system. Thus this class tends to see time averages just as if it were a Poisson arrival process, see Wolff (1982). (See §III.C of Sriram and Whitt (1986) for related discussion.) Therefore, from the point of view of that queue, we can approximate the arrival processes of all classes by Poisson

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

processes. Indeed, if such small arrival rates prevailed for all classes at all queues, then the arrival processes at every queue can be approximated by Poisson process (which is an important observation). Note that this property does not require exponential service-time distributions.

However, the present approximation principle still applies to the departure processes of individual classes. If at some other queue the conditions for the full Poisson approximation are not applicable, i.e., if the number of classes is no longer large or if not all classes have very small rates there, then the full Poisson approximation is not appropriate there. For example, suppose that the different classes are routed to separate queues with much slower service rates after completing service at the multi-class queue. To be more concrete, think of a packet communication network with traffic traveling over virtual circuits from one terminal to another. The packets from one source pass through some switches and then go to the destination terminals. The full Poisson approximation might be appropriate at the heavily shared switches but not at the separate destinations. Dependence in the originating packet arrival processes would thus pass through the switch, without being apparent there, and reappear at the destination. Obviously such dependence effects could not be captured by naive decomposition approximation methods.

The results here imply that if the departure process of one class satisfying the assumptions is routed to a separate queue from all other classes that the congestion at this second queue will be nearly the same as if the arrival process at this last queue were the original external arrival process of that class at its first queue. This follows from the results here plus continuity theorems for queues; see Chapter 3 of Franken et al. (1981).

The light-traffic approximation principles above lead to approximate characterizations of arrival processes of individual classes. However, when there are multiple classes at a queue, then the congestion experienced by the class of interest still depends on the arrival processes of the other classes. The congestion experienced by an arbitrary customer might still be approximated by the methods in Albin (1984), Fendick et al. (1988a, b), and Whitt (1983, 1987b). Moreover, if all the classes do not have identical arrival processes, then the congestion experienced by the different classes is typically different. To address this complication, one can use approximations such as those proposed by Holtzman (1982) and Albin (1986). We do not address this issue here.

We now proceed to a theoretical justification of the approximation principles above. In §3 we describe a simulation experiment.

2. Supporting Light-Traffic Limits

From the discussion in §1.3, it should be clear that the light-traffic approximation is valid in great generality, provided that the difference in time scales is indeed sufficiently great. We provide theoretical justification in this section.

Let T_k be the arrival epoch of customer from the designated class at some queue; let D_k be the departure epoch of this customer; and let S_k be its sojourn time (the time spent by customer k in the system). These variables are obviously related by

$$D_k = T_k + S_k, \qquad k \geq 1, \tag{1}$$

regardless of the nature of the queue and the other classes.

Now we formalize what we mean by our light-traffic limit. We consider a sequence of systems indexed by n with variables D_{nk}, T_{nk} and S_{nk} defined as above. Obviously $D_{nk} = T_{nk} + S_{nk}$ for each n as well as for each k. It is natural to assume that $T_{nk} \to \infty$ as $n \to \infty$ for each k in order to let the arrival rate of the designated class get small, while having $S_{nk} \to S_k$ as $n \to \infty$ for each k to reflect the rest of the system being unchanged, but we want to focus on the arrival and departure process of the designated class in their time scale. Hence, instead we assume that

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$\lim_{n\to\infty} T_{nk} = T_k \qquad \text{w.p.1 for all } k \tag{2}$$

and assume that otherwise the time scale for the queue is speeded up as $n \to \infty$. For the designated class, then, what matters is how the sojourn times change with n. The general case is covered by the following elementary proposition.

PROPOSITION 1. *If $T_{nk} \to T_k$ and $S_{nk} \to 0$ as $n \to \infty$ w.p.1 (in probability) for each k, then $[D_{n1}, \ldots, D_{nk}] \to [T_1, \ldots, T_k]$ in R^k as $n \to \infty$ w.p.1 (in probability) for each k.*

Of course, it remains to verify the conditions of Proposition 1. To be more concrete, we focus on a special case. We consider an s-server queue ($1 \le s < \infty$) with unlimited waiting room and some work-conserving queue discipline. (No servers are free when a customer is waiting in queue; the service time and interarrival times are unaffected by the discipline; examples are first-come first-served (FCFS), last-come first-served (LCFS), random order of service (ROS) and nonpreemptive priority.) For simplicity, we also assume that each customer stays in service until departure once service has begun.

Two classes of customers arrive at this multi-server queue. Let class 1 be the class of interest with arrival epochs T_k, departure epochs D_k and sojourn times S_k. We make no assumptions about the joint distributions $[T_1, \ldots, T_k]$; e.g., we do not assume independence or identical distributions for the interarrival times. Let the associated sequence of class 1 service times be $\{v_{1k}: k \ge 1\}$. Let $C_1(0)$ be the initial (finite) number of customers in the system at time 0 and let $W_1(0)$ be the initial (finite) amount of class-1 work in remaining service time of these customers. Let class 2 represent the remaining customers, which will typically be the aggregate of several other classes. We construct the sequence of models indexed by n by leaving the class-1 arrival epochs unchanged, i.e., by assuming that $T_{nk} = T_k$ for all n and k. Assuming that class-1 arrival epochs are independent of n is tantamount to having the class-1 arrival process be an exogenous input to the queue, so that the following analysis only applies to an external arrival process. (Of course, the concrete result in this case supports the approximation more generally, but more supporting theorems are still needed.)

At the same time we set $T_{nk} = T_k$, we also speed up time by n. Hence, the new class-1 variables are $v_{1nk} = n^{-1}v_{1k}$, $C_{1n}(0) = C_1(0)$ and $W_{1n}(0) = n^{-1}W_1(0)$. This construction for the class-1 variables v_{1nk}, $C_{1n}(0)$ and $W_{1n}(0)$ is tantamount to assuming that they are also independent of class 2. Let $N_{2n}(t)$ represent the number of servers busy serving class-2 customers at time t in the nth model. We now show that the condition of Proposition 1 is satisfied under a condition on $N_{2n}(t)$, which is usually satisfied because of the time scaling.

PROPOSITION 2. *For this special case involving an exogenous class-1 arrival process, if* (i) *$s = \infty$ or* (ii) *if $s < \infty$ and $\int_0^t N_{2n}(u)du \to \rho_2 st$ as $n \to \infty$ w.p.1 for all t where $\rho_2 < 1$, then $S_{nk} \to 0$ as $n \to \infty$ w.p.1, as required for Proposition 1.*

PROOF. First pick k and condition on $(T_1, \ldots, T_k) = (t_1, \ldots, t_k)$. In the nth system, the sum of the initial work plus the first k service times for class 1 is $n^{-1}[W_1(0) + v_{11} + \cdots + v_{1k}]$, which converges to 0 as $n \to \infty$. Hence, for any $\epsilon > 0$, the arrival at epoch t_i will depart before $t_i + \epsilon$, $1 \le i \le k$, for all sufficiently large n provided there is sufficient spare service capacity in the time interval $(t_i, t_i + \epsilon)$. By the condition in (ii),

$$\int_{t_i}^{t_i+\epsilon} N_{2n}(u)du \to \rho_2 s\epsilon \qquad \text{as} \qquad n \to \infty \text{ w.p.1},$$

when $s < \infty$, so that there is indeed at least $\epsilon s(1 - \rho_2)$ spare capacity in $(t_i, t_i + \epsilon)$. For all n sufficiently large, $n^{-1}[W_1(0) + v_{11} + \cdots + v_{1k}] < \epsilon s(1 - \rho_2)$. ■

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

We now add further detail about class 2 in order to provide easily verifiable sufficient conditions for the condition of Proposition 2. Obviously it suffices to restrict attention to the case $s < \infty$. The idea is to exploit $L = \lambda W$ arguments, as in Franken et al. (1981), Glynn and Whitt (1986) and Stidham (1974), with the system being the servers (excluding the waiting room) in order to establish convergence of the integral of the number of busy servers.

We begin by specifying class 2 in more detail before constructing the sequence of models indexed by n. Let $\{A_2(t): t \geq 0\}$ be the original counting process governing class-2 arrivals and let $\{v_{2k}: k \geq 1\}$ be the sequence of successive class-2 service times. Let $C_2(0)$ be the initial (finite) number of class-2 customers in the system at time 0, and let $W_2(0)$ be the initial (finite) amount of class-2 work in service time of these customers. Our basic assumption for class 2 is that certain averages obey strong laws of large numbers (SLLNs) with stable limits; i.e., we assume that

$$t^{-1}A_2(t) \to \lambda_2 \quad \text{w.p.1} \quad \text{as} \quad t \to \infty \quad \text{and} \tag{3}$$

$$m^{-1}\sum_{i=1}^{m} v_{2i} \to \tau_2 \quad \text{w.p.1} \quad \text{as} \quad m \to \infty \quad \text{where} \tag{4}$$

$$\rho_2 \equiv \lambda_2\tau_2/s < 1. \tag{5}$$

We assume that each server works at rate 1, so that the system is stable for class 2 alone by (5).

Next we assume that classes 1 and 2 are independent, in the sense that $(\{(T_k, v_{1k}): k \geq 1\}, C_1(0), W_1(0))$ is independent of $(\{A_2(t): t \geq 0\}, \{v_{2k}: k \geq 1\}, C_2(0), W_2(0))$, and that these variables are independent of the past state of the system. This independence for class 2 can be circumvented by assuming appropriate uniform convergence as $n \to \infty$, but then the conditions become harder to verify.

Let $Q_2(t)$ be the number of class-2 customers that would be in the queue (waiting before beginning service) at time t under the assumption that class-1 customers were not present. In addition, we assume that

$$t^{-1}Q_2(t) \to 0 \quad \text{w.p.1} \quad \text{as} \quad t \to \infty. \tag{6}$$

Recall that we have assumed that a customer stays in service until departure once service has begun. Let $B_2(t)$ be the number of class-2 customers that would begin service in $[0, t]$, if no class 1 customers were present. Obviously $B_2(t) = A_2(t) + C_2(0) - Q_2(t)$. By (3) and (6),

$$t^{-1}B_2(t) \to \lambda_2 \quad \text{w.p.1} \quad \text{as} \quad t \to \infty. \tag{7}$$

Let $D_2(t)$ be the number of class-2 departures in $[0, t]$ under the assumption that class-1 customers were not present. Since $B_2(t) - s \leq D_2(t) \leq B_2(t)$ for all t,

$$t^{-1}D_2(t) \to \lambda_2 \quad \text{w.p.1} \quad \text{as} \quad t \to \infty. \tag{8}$$

Given (3)–(5), the limits (6), (7) and (8) are obviously equivalent. We have not yet determined when these limits can be deduced directly from (3)–(5). For the case of a single-server queue, this implication is established in Theorem 4.2 of Gelenbe and

For the models indexed by n, we treat class 2 by scaling time by n; i.e., the new quantities are: $A_{2n}(t) = A_2(nt)$, $v_{2ni} = n^{-1}v_{2i}$, $Q_{2n}(t) = Q_2(nt)$, $B_{2n}(t) = B_2(nt)$ and $D_{2n}(t) = D_2(nt)$. (Note that n appears in the final representations only in the time scaling. For $A_{2n}(t)$ and v_{2ni}, we use the independence assumption above, for $Q_{2n}(t)$, $B_{2n}(t)$ and $D_{2n}(t)$, we use the fact that these processes do not count the class 1 customers; we have not defined the corresponding quantities when class 1 is included.)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

PROPOSITION 3. *If $T_k \to \infty$ as $k \to \infty$ w.p.1 and (3)–(6) hold with the class independence condition specified above, then $\int_0^t N_{2n}(u)du \to \rho_2 st$ as $n \to \infty$ w.p.1 for each t, so that the conditions and conclusions of Propositions 1 and 2 are satisfied.*

PROOF. For any t given, choose k so that $t_k \le t < t_{k+1}$, which is possible by the first condition above. By standard inequalities (related to the proof of $L = \lambda W$),

$$\sum_{i=1}^{D_{2n}(t-n^{-1}Z)} v_{2ni} \le \int_0^t N_{2n}(u)du \le \sum_{i=1}^{A_{2n}(t)} v_{2ni} + n^{-1}Z \tag{9}$$

where $Z = W_1(0) + W_2(0) + \sum_{i=1}^{k} v_{1i}$. Note that $D_{2n}(t)$ and $A_{2n}(t)$ in the bounds are class-2 processes that are independent of class 1; they count class-2 events without class 1 present. Equivalently, by the normalization, (9) can be expressed as

$$n^{-1}\sum_{i=1}^{D_2(nt-Z)} v_{2i} \le \int_0^t N_{2n}(u)du \le n^{-1}\sum_{i=1}^{A_2(nt)} v_{2i} + n^{-1}Z.$$

By (3), (4) and (8), the two bounds converge to $\rho_2 st$ w.p.1. ■

Results about the class-2 work seen by a class-1 arrival also follow easily from the proof of Proposition 3. Let $V_{2n}(t)$ ($V_2(t)$) be the amount of class-2 work in service time in the nth system (the original system without class 1) at time t with the scaling previously specified for n. Let $\Rightarrow$ denote convergence in distribution (weak convergence), as in Billingsley (1968).

PROPOSITION 4. (a) *If, in addition to the assumptions of Proposition 3, $V_2(t) \Rightarrow V_2$ as $t \to \infty$, then $nV_{2n}(T_k) \Rightarrow V_2$ as $n \to \infty$ for each k.*

(b) *If, in addition, $V_2(t+s)$ and $V_2(t)$ are asymptotically independent as $s \to \infty$ for each t, then*

$$[nV_{2n}(T_1), \ldots, nV_{2n}(T_k)] \Rightarrow [X_1, \ldots, X_k] \quad \text{in } R_k \quad \text{as} \quad n \to \infty \quad \text{for all } k$$

where $X_1, \ldots, X_k$ are i.i.d. (independent and identically distributed) random variables with distribution V_2.

Proposition 4 draws the relatively obvious conclusion that in the limit class-1 customers act like independent outside observers that do not interact with the system. Like Poisson arrivals, they see time averages (Wolff 1982).

3. A Simulation Experiment

To test the light-traffic approximation for single-class departure processes from multi-class queues, we consider a packet-queue model from Fendick et al. (1988a, b). For each class i, service times are i.i.d. with mean τ_i and squared coefficient of variation c_{si}^2 (the packets are not of fixed length); arrivals are generated in i.i.d. batches of size having mean m_i and squared coefficient of variation c_{bi}^2; the arrivals within a batch are separated by i.i.d. spacings having mean ξ_i and squared coefficient of variation c_{xi}^2; the interval between the last arrival of one batch and the first arrival of the next batch is the sum of one spacing and an idle period; the successive idle times are i.i.d. with mean η_i and squared coefficient of variation c_{yi}^2; and the service times, batch sizes, spacings and idle periods for all classes are mutually independent. The overall arrival process is a superposition of these independent streams. Customers are served in the order of their arrival by a single server with unlimited waiting space.

In particular, for this test we assume that all service times and spacings are deterministic, so that $c_{si}^2 = c_{xi}^2 = 0$; all idle periods are exponential, so that $c_{yi}^2 = 1$; and all batch sizes are geometric, so that $c_{bi}^2 = (m_i - 1)/m_i$. As a consequence, the arrival process of individual customers of each class is a renewal process with arrival rate

$$\lambda_i = m_i/(m_i\xi_i + \eta_i); \tag{10}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

i.e., the interarrival times are i.i.d. with mean λ_i^{-1}. An interarrival time of class i has squared coefficient of variation (variance divided by the square of the mean)

$$c_{ai}^2 = m_i(1 - \gamma_i)^2(c_{bi}^2 + 1) = (2m_i - 1)(1 - \gamma_i)^2, \tag{11}$$

where $\gamma_i = m_i\xi_i/(m_i\xi_i + \eta_i)$, the proportion of time the class-i source is active (not in an idle period). The overall model can be referred to as a $\sum (GI_i/D_i)/1$ queue. In particular, because of this structure, it is easy to verify the conditions of Proposition 3 for a sequence of models indexed by n. The following is an easy consequence of Proposition 3.

PROPOSITION 5. *Consider the multi-class packet queue model above. In the nth system, let all service times be divided by n; let the class-1 arrival process be unchanged; let the spacings and idle periods of all other classes be divided by n. Then the conditions of Propositions 3 and 4 are satisfied; e.g., the class-1 departure process converges w.p.1 as* $n \to \infty$ *to the class-1 arrival process.*

However, it remains to see what happens for any given n. We consider a fairly realistic case containing 50 classes. There are 25 classes representing originating traffic and 25 classes representing acknowledgments. These streams are not tightly coupled because transmission is full duplex; i.e., there are separate lines going in each direction. In any case, for the simulation, we assume that the 50 arrival processes are independent. Of the 25 originating classes, 20 are "interactive" with parameters $m_i = 2$, $\tau_i = 400$, $\xi_i = 1000$ and $\eta_i = 86{,}000$; and 5 are "batch" with parameters $m_i = 30$, $\tau_i = 800$, $\xi_i = 2000$ and $\eta_i = 570{,}000$. (Time is measured in milliseconds (ms.).) The acknowledgments are assumed to be of the same mix: There are 20 "interactive acks" with parameters $m_i = 2$, $\tau_i = 40$, $\xi_i = 1000$ and $\eta_i = 86{,}000$; and 5 "batch acks" with parameters $m_i = 30$, $\tau_i = 40$, $\xi_i = 2000$ and $\eta_i = 570{,}000$. The acknowledgment streams are just like the originating traffic streams, except the service times are shorter, because the acknowledgment packets are shorter.

For this model, the average service time is 288.75 ms., the total arrival rate is 0.001385 pkts./ms., and the total traffic intensity is $\rho = 0.4$. The case of $\rho = 0.8$ is also considered by multiplying all service times by 2.

Simulation estimates of the squared coefficient of variation of a stationary interval between departures for each class are displayed in Table 1 together with the squared coefficient of variation of an interarrival time in the renewal arrival process of each class, computed via (11). As indicated above, there are four kinds of classes, but each of the 50 classes is treated separately.

The results strongly support the simple light-traffic approximation principle for single-class departure processes from multi-class queues. Even though the number of classes is not exceptionally large, the departure process variability parameter of each class is quite close to the corresponding (exact) arrival process variability parameter, so that simple "back-of-the-envelope" calculations based on this principle seem justified.

It is intuitively obvious that the simple approximation ought to perform better as the total traffic intensity decreases, with everything else held fixed, and the results support this principle: The approximation performs better at $\rho = 0.4$ than at $\rho = 0.8$.

The deviations observed at $\rho = 0.8$ naturally suggest looking for more accurate refined approximations, but that is not our purpose here. Such refined approximations are developed in Bitran and Tirupati (1988) and Whitt (1987a). In contrast to the light-traffic analysis here, this approximation is based on a heavy-traffic limit. In particular, the suggested approximation for class i, (13) in Whitt (1987a), is

$$c_{di}^2 = (1 - 2\rho_i\rho + \rho_i^2)c_{ai}^2 + \rho_i^2 c_{si}^2 + p_i \sum_{\substack{j=1 \\ j\neq i}}^{50} \rho_j^2(1/p_j)(c_{aj}^2 + c_{sj}^2), \tag{12}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 1

A Comparison of Simulation Estimates of c_{di}^2, the Squared Coefficient of Variation of a Stationary Interval between Departures of Class i with the Corresponding Exact Squared Coefficient of Variation of an Interarrival Time from (11).

Customer Class	Squared Coefficient of Variation of Single-Class Arrival Process c_{ai}^2	Simulation Estimate of c_{di}^2	
		$\rho = 0.4$	$\rho = 0.8$
Originating Interactive (20)	2.865	2.856 (±0.019)	3.081 (±0.021)
Originating Batch (5)	48.3	48.0 (±1.69)	44.5 (±0.36)
Interactive Acknowledgments (20)	2.865	2.8653 (±0.015)	3.137 (±0.013)
Batch Acknowledgments (5)	48.3	47.4 (±0.50)	48.1 (±0.86)

Note. 95% confidence intervals are given below the simulation estimates in parentheses.

where ρ_j is the traffic intensity of class j alone and $p_j = \lambda_j/\lambda$ is the proportion of the total arrival rate due to class j. Since $c_{si}^2 = 0$ for all i, (12) is equivalent to

$$c_{di}^2 = (1 - 2\rho_i\rho + \rho_i^2)c_{ai}^2 + p_i \sum_{\substack{j=1 \\ j\neq 1}}^{50} \rho_j^2(1/p_j)c_{aj}^2$$

$$= 1 - \rho^2[q_i(2 - q_i)]c_{ai}^2 + \rho^2 \sum_{\substack{j=1 \\ j\neq 1}}^{50} q_j^2(p_i/p_j)c_{aj}^2, \tag{13}$$

where $q_j = \rho_j/\rho$ is the proportion of the total traffic intensity due to class j. From (12), we see that the approximation is of the form $c_{di}^2 = \alpha_i c_{ai}^2 + \beta_i$. The coefficients α_i and β_i for this experiment are given in Table 2. The approximations based on (12) and (13) are compared with the simulation results for this example in Table 3.

The approximations obviously are consistent with the light-traffic approximation principle and correctly predict, at least qualitatively, the effect of increasing ρ. However, only in some cases does the refinement in (13) provide a truly significant improvement over the simple approximation $c_{di}^2 \approx c_{ai}^2$. Relatively, the refinement in (13) seems to do better

TABLE 2

The Coefficients α_i and β_i in the Linear Approximation $c_{di}^2 = \alpha_i c_{ai}^2 + \beta_i$ in (12) *and* (13).

Customer Class	c_{aj}^2	$\rho = 0.4$		$\rho = 0.8$	
		α_i	β_i	α_i	β_i
Originating Interactive (20)	2.87	0.9928	0.172	0.9712	0.688
Originating Batch (5)	48.3	0.9710	0.298	0.8838	1.192
Interactive Acknowledgments (20)	2.87	0.9993	0.172	0.9971	0.688
Batch Acknowledgments (5)	48.3	0.9985	0.352	0.9939	1.408

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

TABLE 3

A Comparison of the Heuristic Heavy-traffic Approximation for c_{di}^2 in (12) and (13) with the Simulation Estimates from Table 1.

	Model Parameters				$\rho = 0.4$		$\rho = 0.8$	
Customer Class	c_{aj}^2	λ_j	p_j	$q_j = \rho_j/\rho$	Approx. (13)	Sim.	Approx. (13)	Sim.
Originating Interactive (20)	2.87	2.27×10^{-5}	0.0164	0.0227	3.02	2.86	3.48	3.08
Originating Batch (5)	48.3	4.76×10^{-5}	0.0344	0.0953	47.2	48.0	43.9	44.5
Interactive Acknowledgments (20)	2.87	2.27×10^{-5}	0.0164	0.00227	3.04	2.87	3.55	3.14
Batch Acknowledgments (5)	48.3	4.76×10^{-5}	0.0344	0.00476	48.6	48.7	49.4	48.1

predicting the variability of the more highly variable batch processes than the interactive sources. The noisy batch processes may distort the interactive processes.

Table 1 only displays the squared coefficients of variation of single interdeparture times. However, simulation results also show that the entire single-class departure process

TABLE 4

A Comparison of Histograms for the Interarrival-Time Distribution (Exact) and the Interdeparture-Time Distribution (Simulated with Sample Size 50,000) for Originating Traffic from an Interactive Source (Mean Interarrival Time = 44 Seconds).

Lower Limit of Histogram Interval		Interarrival-Time Distribution (Exact)	Interdeparture-Time Distribution (Simulated)	
In Seconds	In Units of Mean Interarrival Times		$\rho = 0.4$	$\rho = 0.8$
0.0	0.0	0.5433	0.5468	0.5525
8.8	0.2	0.0444	0.0443	0.0505
17.6	0.4	0.0401	0.0392	0.0422
26.4	0.6	0.0362	0.0358	0.0364
35.2	0.8	0.0327	0.0327	0.0309
44.0	1.0	0.0295	0.0297	0.0273
52.8	1.2	0.0266	0.0263	0.0248
61.6	1.4	0.0240	0.0248	0.0220
70.4	1.6	0.0217	0.0206	0.0193
79.2	1.8	0.0196	0.0199	0.0179
88.0	2.0	0.0177	0.0173	0.0161
96.8	2.2	0.0160	0.0156	0.0153
105.6	2.4	0.0144	0.0141	0.0124
114.4	2.6	0.0130	0.0123	0.0128
123.2	2.8	0.0117	0.0118	0.0112
132.0	3.0	0.0106	0.0106	0.0097
140.8	3.2	0.0096	0.0090	0.0083
149.6	3.4	0.0086	0.0091	0.0081
158.4	3.6	0.0078	0.0075	0.0074
167.2	3.8	0.0070	0.0072	0.0073
176.0	4.0	0.0654	0.0653	0.0675

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

tends to be close to the corresponding single-class arrival process, as predicted by Proposition 5. For example, Table 4 compares histograms of the exact interarrival-time distribution for an interactive source with the estimated interdeparture-time distribution for this class obtained by simulation with sample size 50,000. The histogram has 40 intervals with the width of each interval being 8.8 seconds, which is 0.2 times a mean interarrival time for this class (apply (10)). These histogram intervals were chosen to be in the time scale of this single-class arrival process, where the mean interarrival time is 44 seconds. In a shorter time scale, e.g., in the order of the overall mean service time (0.289 secs.), differences between the single-class arrival process and the corresponding single-class departure process are readily apparent. Note that the spacings between packets within a batch of this class are 1 sec., so that they too are in the shorter time scale. In the time scale of the single-class arrival process (in tens of seconds), the distinction between the given process and a pure-batch process (without any spacing) is small.

Finally, individual sample paths of successive single-class departure epochs and arrival epochs also match up well in the time scale of the single-class arrival process, thus providing empirical verification for the light-traffic approximations principle in its strongest form. (Typical sample paths of 400 successive customers from the middle of the simulation runs appear in an unpublished appendix available from the author.) The overall accuracy of the approximation thus seems to be reasonably well described by Table 1.[1]

[1] I am grateful to Kerry Fendick for conducting the simulation in §3 and for many helpful discussions about multi-class queues.

References

ALBIN, S. L., "On Poisson Approximations for Superposition Arrival Processes in Queues," *Management Sci.*, 28 (1982), 126–137.

———, "Approximating a Point Process by a Renewal Process. II. Superposition Arrival Processes to Queues," *Oper. Res.*, 32 (1984), 1133–1162.

———, "Delays for Customers from Different Arrival Streams to a Queue," *Management Sci.*, 32 (1986), 329–340.

——— AND S. KAI, "Approximation for the Departure Process of a Queue in a Network," *Naval Res. Logist. Quart.*, 33 (1986), 129–143.

BILLINGSLEY, P., *Convergence of Probability Measures*, Wiley, New York, 1968.

BITRAN, G. R. AND D. TIRUPATI, "Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference," *Management Sci.*, 34 (1988), 75–100.

BÖKER, F. AND R. SERFOZO, "Ordered Thinnings of Point Processes and Random Measures," *Stochastic Process. Appl.*, 15 (1983), 113–132.

ÇINLAR, E., "Superposition of Point Processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis (Ed.), Wiley, New York, 1972, 549–606.

FENDICK, K. W., V. R. SAKSENA AND W. WHITT, "Dependence in Packet Queues," *IEEE Trans. Commun.* (1988a), to appear.

———, ———, AND ———, "Dependence in Packet Queues. II," 1988b, in preparation.

FRANKEN, P., "A Refinement of the Limit Theorem for the Superposition of Independent Renewal Processes," *Theory Probab. Appl.*, 8 (1963), 320–328.

———, D. KÖNIG, U. ARNDT AND V. SCHMIDT, *Queues and Point Processes*, Akademie-Verlag, Berlin, 1981.

GELENBE, E. AND D. FINKEL, "Stationary Deterministic Flows: II. The Single Server Queue," *Theor. Comp. Sci.*, 52 (1987), 269–280.

GLYNN, P. W. AND W. WHITT, "A Central-Limit-Theorem Version of $L = \lambda W$," *Queueing Systems*, 2 (1986), 191–215.

HARRISON, J. M., "Brownian Models of Queueing Networks with Heterogeneous Customer Populations," *Stochastic Differential Systems, Stochastic Control Theory and Applications*, W. Fleming and P. L. Lions (Eds.), Springer-Verlag, New York, 1988, 147–186.

HOLTZMAN, J. M., "Mean Delays of Individual Streams into a Queue: The $\Sigma\ GI_i/M/1$ Queue," *Applied Probability—Computer Science: The Interface*, Vol. I, R. L. Disney and T. J. Ott (Eds.), Birkhäuser, Boston, 1982, 417–430.

IGLEHART, D. L. AND W. WHITT, "Multiple Channel Queues in Heavy Traffic. II. Sequences, Networks and Batches," *Adv. in Appl. Probab.*, 2 (1970), 355–369.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

KELLY, F. P., "Blocking Probabilities in Large Circuit-Switched Networks," *Adv. in Appl. Probab.*, 18 (1986), 473–505.

KHINTCHINE, A. Y., *Mathematical Methods in the Theory of Queueing*, Griffin, London, 1960.

NEWELL, G. F., "Approximations for Superposition Arrival Processes in Queues," *Management Sci.*, 30 (1984), 623–632.

PALM, C., "Variation in Intensity in Telephone Conversation," *Ericsson Technics* (1943–4), 1–189.

REIMAN, M. I., "Open Queueing Networks in Heavy Traffic," *Math. Oper. Res.*, 9 (1984), 441–458.

SEGAL, M. AND W. WHITT, "A Queueing Network Analyzer for Manufacturing," *Proc. 12th Internat. Teletraffic Congress*, Torino, Italy, June 1988.

SERFOZO, R. F., "Compositions, Inverses and Thinnings of Random Measures," *Z. Wahrsch. Verw. Gebiete*, 37 (1977), 253–265.

———, "Partitions of Point Processes: Multivariate Poisson Approximations," *Stochastic Process. Appl.*, 20 (1985), 281–294.

SRIRAM, K. AND W. WHITT, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE J. Sel. Areas Commun.*, SAC-4 (1986), 833–846.

STIDHAM, S., JR., "A Last Word on $L = \lambda W$," *Oper. Res.*, 22 (1974), 417–421.

WHITT, W., "Approximating a Point Process by a Renewal Process. I. Two Basic Methods," *Oper. Res.*, 39 (1982), 125–147.

———, "The Queueing Network Analyzer," *Bell System Tech. J.*, (1983), 2779–2815.

———, "Open and Closed Models for Networks of Queues," *AT&T Bell Lab. Tech. J.*, 63 (1984a), 1911–1979.

———, "Departures from a Queue with Many Busy Servers," *Math. Oper. Res.*, 9 (1984b), 534–544.

———, "Blocking When Service Is Required from Several Facilities Simultaneously," *AT&T Tech. J.*, 64 (1985a), 1807–1856.

———, "Queues with Superposition Arrival Processes in Heavy Traffic," *Stochastic Process. Appl.*, 21 (1985b), 81–91.

———, "Approximations for Single-Class Departure Processes from Multi-Class Queues," 1987a, in preparation.

———, "A Queueing Network Analyzer for Manufacturing," 1987b, in preparation.

WOLFF, R. W., "Poisson Arrivals See Time Averages," *Oper. Res.*, 30 (1982), 223–231.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.