MATHEMATICS OF OPERATIONS RESEARCH
Vol. 9, No. 4, November 1984
Printed in U.S.A.

DEPARTURES FROM A QUEUE WITH MANY BUSY SERVERS*

WARD WHITT

AT&T Bell Laboratories

To analyze networks of queues, it is important to be able to analyze departure processes from single queues. For the $M/M/s$ and $M/G/\infty$ models, the stationary departure process is simple (Poisson), but in general the stationary departure process is quite complicated. As a basis for approximations, this paper shows that the stationary departure process is approximately Poisson when there are many busy slow servers in a large class of stationary $G/GI/s$ congestion models having s servers, infinite waiting room, the first-come first-served discipline, and mutually independent and identically distributed service times that are independent of a stationary arrival process. Limit theorems are proved for the departure process in a $G/GI/s$ system in which the number of servers and the offered load (arrival rate divided by the service rate) both increase. The asymptotic behavior of the departure process depends on the way the arrival rate changes. If the arrival rate is held fixed, so that the offered load increases by slowing down the service rate, then the departure process converges to a Poisson process. For this result, the service-time distribution is assumed to be phase-type. Other limiting behavior occurs if the arrival rate approaches zero or infinity. Convergence is established in each case by applying previous heavy-traffic limit theorems.

1. Introduction and summary. To analyze networks of queues, it is important to be able to analyze departure processes from single queues. Useful for this purpose are the results by Burke (1958) and Mirasol (1963) showing that the departure process is Poisson in stationary $M/M/s$ and $M/G/\infty$ congestion models; see Bremaud (1981, Chapter V). Other research has shown, however, that the departure process tends to be quite complicated when these assumptions are relaxed; see Disney and König (1984, Chapter VII) and references there. Thus it is natural to look for approximations. In particular, it is of interest to know when the departure process is approximately Poisson. By continuity arguments, as in Franken et al. (1981, Chapter 3) and references there, it is not difficult to show that the departure process is approximately Poisson in a stationary $G/G/s$ model (built from stationary point processes; see Franken et al.) that is appropriately close to a stationary $M/M/s$ model. Similarly, it is not difficult to show that the departure process in a stationary $M/G/s$ system approaches a Poisson process as $s \to \infty$.

We might also expect the stationary departure process to be approximately Poisson when the equilibrium number of busy servers tends to be large in a $G/GI/s$ model having s servers, infinite waiting room, the first-come first-served discipline and i.i.d. service times that are independent of a stationary arrival process. With many busy servers, the departure process might behave like the superposition of many renewal processes having service times as renewal intervals. Hence, some variant of the classical superposition limit theorem might apply; see Çinlar (1972). Note, however, that some critical assumptions are not satisfied: The component renewal processes are typically neither independent nor stationary. The residual service time will differ

*Received November 1, 1982; revised May 11, 1983.
AMS 1980 subject classification. Primary: 60K25. Secondary: 90B22.
OR/MS Index 1978 subject classification. Primary: 697 Queues/output process. Secondary: 692 Queues/limit theorems.
Key words. Departure process, limit theorems, queueing networks.

0364-765X/84/0904/0534$01.25

depending on how long the customers have been in service. There are also some obvious counterexamples that indicate limitations of this heuristic reasoning. First, if the service times are deterministic, then the departure process is just a shifted version of the arrival process when there are infinitely many servers. Second, if the service-time distribution has positive mass at zero, then the departure process will have multiple points when there are finitely many servers and customers waiting in queue. Third, if the interarrival-time and service-time distributions have common lattice support, e.g., the integers, so will the departure process.

The purpose of this paper, nevertheless, is to justify the heuristic reasoning. After restricting the class of service-time distributions to avoid the counterexamples mentioned above, we prove that the stationary departure process does converge to a Poisson process in stationary $G/GI/s$ models as $s \to \infty$ and $\alpha \to \infty$, where $\alpha = \lambda/\mu$, λ is the arrival rate and μ is the service rate. The quantity α is often referred to as the offered load; it is the expected number of busy servers; see (4.2.3) of Franken et al. If $s < \infty$, then $\rho = \alpha/s$ is the traffic intensity.

We consider a sequence of $G/GI/s$ systems indexed by n in which $s_n \to \infty$ and $\alpha_n = \lambda_n/\mu_n = n$. The limiting behavior of the associated sequence of departure processes also depends on the way the arrival rates λ_n and service rates μ_n change with n. Since the departure rate is just the arrival rate, the only way we can have the associated sequence of departure processes without normalization converge to a nondegenerate limit is to have $\lambda_n \to \lambda$ as $n \to \infty$, $0 < \lambda < \infty$. Accordingly, we establish convergence to a Poisson process when $\lambda_n = 1$ for all n (Theorem 2).

Our proof involves only a few ideas. First, we adopt the martingale or Strasbourg view and look at our departure process as a point process with a stochastic intensity with respect to an appropriate history or filtration; see Bremaud (1981). However, we make no real use of the history. Second, as in Brown (1978, 1983), we obtain convergence to the Poisson process by showing convergence of the compensators (integral of the intensity). Third, we assume the service-time distribution is phase type as in Neuts (1981) or Whitt (1982) and represent the intensity as a function of the number of phases of each type in service. Finally, we obtain convergence of the compensators by changing the time scale and applying previous heavy-traffic limit theorems in Borovkov (1967), Halfin and Whitt (1981) and Whitt (1982). Our result and proof are closely related to Brown and Pollett (1982) and Pollett (1982), which contain Poisson approximations for the flows in Markovian networks of queues.

Our limit theorem is useful for generating approximations. For example, for the approximation method described in Whitt (1983a) it suggests that the variability parameter c^2 (the squared coefficient of variation of the renewal interval in an approximating renewal process) in an approximation of the departure process in a $G/GI/s$ model should approach 1 as $s \to \infty$, $\alpha \to \infty$ and $\mu \to 0$.

Our limit theorem is also useful for generating approximations for networks of queues, as in Whitt (1983a). First, since the arrival process can be quite general, the result applies to general nodes in the network having nonrenewal input. Second, under these limiting conditions, the departure process is not only asymptotically Poisson but also asymptotically independent of the arrival process in any bounded time interval. (This is easy to prove; we do not give the details.) Hence, such a node with many busy slow servers tends to decouple the network. It can be replaced (approximately) by an independent Poisson source. (More on this will appear in another paper.)

We also prove limit theorems when $\lambda_n \to 0$ or $\lambda_n \to \infty$ with $s_n \to \infty$ and $\alpha_n \to \infty$. For these the departure process is normalized. In §4 we discuss the case of fast arrivals in which we fix the service rate instead of the arrival rate, and let $\mu_n = 1$ and $\lambda_n = n$ for all n. The main limit theorem in this case (Theorem 3) is due to Borovkov (1967). However, there is a gap in the theory because we are primarily interested in the

stationary version of the departure process, whereas Borovkov assumes the systems are initially empty. We can obtain a limit theorem for the stationary version in the case of renewal arrival processes (Theorem 5), but there is another gap because it remains to show that the limit process is consistent with Borovkov's results.

It is significant that the normalization constant in the conjectured limit theorem for the stationary departure process, obtained from Borovkov (1967) by taking the iterated limit in the wrong order, is not the same as if the departure process were Poisson. The variability of the departure process is affected by the variability of the service-time distribution and the variability of the arrival process in a rather complicated way; see (4.11). Just as Wolff (1977) observed for the number of busy servers, greater variability in the service times need not cause greater variability in the departure process; the qualitative behavior depends on the arrival process. Obviously, this limit theorem also can serve as a valuable guide when developing approximations for departure processes.

There is another important way to interpret the different kinds of limiting behavior that occur for the departure process depending on the way λ_n changes as $n \to \infty$. The choice of λ_n is equivalent to a choice of the time scale. Instead of varying λ_n, we can let $\lambda_n = 1$ and vary the time. Indeed, suppose $\lambda_n = 1$ and that n is large. Let $D_n(t)$ be the number of departures in the interval $[0, t)$ in the nth system. Our results say that $D_n(t)$ behaves like a Poisson process when $t = O(1)$, like a point process with a rather complex Gaussian stochastic intensity when $t = O(n)$, and like the arrival process when $t \gg n$, i.e., when $t^{-1} = o(n^{-1})$. In other words, the departure process for large n is still complicated. Its structure depends on the way you look at it. For example, the asymptotic method in Whitt (1982a) corresponds to $t \gg n$ and approximates the departure process by the arrival process. On the other hand, the stationary-interval method in Whitt (1982a) corresponds to $t = O(1)$ and our results indicate that it approximates the departure process by a Poisson process. For intermediate values of t, the variability of the departure process, as measured by the normalization constant in the central limit theorem when $t = O(n)$, is between these two extremes. (See Remark 4.2.)

We illustrate these ideas in §6, where we conclude by discussing the asymptotic behavior of two service facilities in series, the first being an infinite-server system with many busy servers and the second being a single-server system. Hence, the arrival process to the second system is the departure process studied in the previous sections of this paper. We show how the different kinds of limiting behavior for the departure process are reflected in the second facility.

2. A representation for the departure process. Consider a $G/GI/s$ queueing model and let the arrival process $A(t)$ be a stationary point process with a predictable stochastic intensity $\lambda(t)$ with respect to its history $\mathcal{F}(A(s), s \leq t)$; see Chapter II of Bremaud (1981). Suppose that $E\lambda(t) < \infty$ and let the arrival rate be $\lambda = E\lambda(1)$.

Let each service time consist of a random (finite and positive) number of phases, with the length of each phase being exponentially distributed with mean β^{-1}. A customer in service upon completing phase k leaves the system with probability p_k and moves on to phase $k+1$ with probability $1 - p_k$. Let m be the maximum number of phases, i.e., assume that $p_m = 1$. This phase-type distribution was used in Whitt (1982). We could of course also use other phase-type distributions, e.g., Neuts (1981, Chapter 2). Our class of phase-type distributions, like most others, is dense in the family of all probability distributions on the nonnegative real line, using the topology of weak convergence as in Billingsley (1968); see p. 179 of Whitt (1982).

For the $G/GI/s$ model described above, let $N^k(t)$ be the number of customers in phase k of service at time t. We shall work with the vector-valued process $\mathbf{N}(t) = [N^1(t), \ldots, N^m(t)]$. When there are infinitely many servers, the process $\mathbf{N}(t)$ is

conditionally Markov given the arrival process; when there are finitely many servers, the process $(\mathbf{N}(t), Y(t))$ obtained by appending the number waiting in queue, $Y(t)$, is also conditionally Markov. In order to obtain a homogeneous Poisson limit, we assume that we have a stationary version of the process $\mathbf{N}(t)$, so that $D(t)$ is a stationary point process. It is easy to verify that such a stationary version of $\mathbf{N}(t)$ exists and is unique when the arrival process is a renewal process and $s = \infty$ or $\rho < 1$ because then, with phase-type service times, the empty system is a regeneration point with finite mean regeneration time; see Whitt (1972). Much more general conditions are given by Franken et al. (1981).

Let $D(t)$ be the number of departures in the interval $[0, t]$. The process $D(t)$ is a point process with the predictable stochastic intensity

$$\Lambda(t) = \sum_{k=1}^{m} \beta p_k N^k(t-), \qquad t \geqslant 0, \tag{2.1}$$

with respect to the history $\{\mathscr{F}_t\} = \{\mathscr{F}(\mathbf{N}(s),\ s \leqslant t)\}$, see Chapter II of Bremaud (1981). (The left-continuous version makes $\Lambda(t)$ predictable.) Hence, $D(t)$ can be represented as the random-time transformation

$$D(t) = \Pi(C(t)), \qquad t \geqslant 0, \tag{2.2}$$

where $\Pi(t)$ is a Poisson process with unit intensity and $C(t)$ is the compensator associated with $\Lambda(t)$, defined by

$$C(t) = \int_0^t \Lambda(u)\, du, \qquad t \geqslant 0. \tag{2.3}$$

In fact, $\Pi(t)$ is defined as

$$\Pi(t) = D\big(C^{-1}(t)\big), \qquad t \geqslant 0, \qquad \text{where} \tag{2.4}$$

$$C^{-1}(t) = \inf\{u : C(u) > t\}, \qquad t \geqslant 0; \tag{2.5}$$

see p. 41 of Bremaud (1981) and references there. Note, however, that $D(t)$ is not a doubly-stochastic or conditional Poisson process as in Serfozo (1972) and Grandell (1976) because the processes $\Pi(t)$ and $C(t)$ in (2.2) and (2.3) are generally dependent.

We use this martingale setting only to obtain the representation (2.2), which enables us to prove the desired limit theorems by exploiting the continuity of the composition map on an appropriate function space.

3. A sequence of systems. We begin this section by giving simple sufficient conditions for the convergence of a sequence of point processes to a Poisson process. Then we apply these results to the sequence of departure processes in a sequence of queueing systems.

3.1. *Convergence of point processes.* For each n, let $D_n(t)$ be a general point process with a stochastic intensity $\Lambda_n(t)$ and associated compensator $C_n(t)$, as defined in (2.3). Our notation is motivated by the application to queues, but the results here are not limited to queues.

We now give conditions on $C_n(t)$ for $D_n(t)$ to converge in distribution. Let $\Rightarrow$ denote convergence in distribution of random elements of the real line R or $D[0, \infty)$; see Billingsley (1968), Whitt (1980) and references there. Let e and ω be the special random elements of $D[0, \infty)$ defined by

$$e(t) = t \quad \text{and} \quad \omega(t) = 1, \qquad t \geqslant 0. \tag{3.1}$$

Given the representation (2.2), the following result is an elementary consequence of

the continuity of composition; see §17 of Billingsley (1968). Related results are contained in Brown (1978), Serfozo (1977) and Whitt (1980). We include the short proof to be complete.

THEOREM 1. *If $C_n(t) \Rightarrow ct$ in R as $n \to \infty$ for each t, then $D_n \Rightarrow \Pi_c$ in $D[0,\infty)$ as $n \to \infty$, where Π_c is a Poisson process with intensity c.*

PROOF. If we can show that $(\Pi_n, C_n) \Rightarrow (\Pi, ce)$, then we obtain $\Pi_n(C_n) \Rightarrow \Pi(ce)$ by the continuity of the composition function; §5, 17 of Billingsley (1968). Since ce is nonrandom, in order to have $(\Pi_n, C_n) \Rightarrow (\Pi, ce)$ it suffices to show that $\Pi_n \Rightarrow \Pi$ and $C_n \Rightarrow ce$ separately; Theorem 4.4 of Billingsley (1968). Note that Π_n and C_n are, in general, dependent. Since Π_n is distributed as Π for each n, $\Pi_n \Rightarrow \Pi$ trivially. Since $C_n(t)$ is nondecreasing and ce is strictly increasing and continuous, to have $C_n \Rightarrow ce$ it suffices to have convergence of the finite-dimensional distributions; see Straf (1972). Since c is nonrandom, it suffices to have $C_n(t) \Rightarrow ct$ in R for each t.

COROLLARY 1. *If $\Lambda_n \Rightarrow c\omega$ in $D[0,\infty)$ as $n \to \infty$, where ω is defined in* (3.1), *then $D_n \Rightarrow \Pi_c$ in $D[0,\infty)$.*

PROOF. Apply the continuous mapping theorem with (2.3). ∎

REMARK (3.1). By different methods, a bound on the distance between a point process and the Poisson process can be established; see Brown (1983) and Brown and Pollett (1982). Brown's bound on the total-variation distance d_t over any interval $[0,t]$ is

$$d_t(D_n, \Pi_c) \leq \int_0^t E|\Lambda_n(u) - c|\,du. \tag{3.2}$$

The applications here are also related to Brown and Pollett (1982) and Pollett (1982), but they consider only Markovian queueing networks.

3.2. *Applications to queues.* We now let $D_n(t)$ represent the nth departure process in a sequence of $G/GI/s$ queueing systems. We are interested in the case in which $\lambda_n = 1$ and $\mu_n = n^{-1}$, but we want to apply previously proved heavy-traffic limit theorems for the case in which $\lambda_n = n$ and $\mu_n = 1$; see Borovkov (1967), Halfin and Whitt (1981) and Whitt (1982). Let $D_n(t)$, $C_n(t)$, etc., be the processes with $\lambda_n = 1$ and $\mu_n = n^{-1}$ and let $\hat{D}_n(t)$, $\hat{C}_n(t)$, etc., be the processes with $\lambda_n = n$ and $\mu_n = 1$. Obviously the processes can be related if we obtain one case from the other simply by rescaling time. Suppose this is done. Then

$$D_n(t) = \hat{D}_n(t/n) \quad \text{and} \tag{3.3}$$

$$C_n(t) = \hat{C}_n(t/n) \tag{3.4}$$

for $t > 0$. Hence, we also have the following variant of Theorem 1.

COROLLARY 2. *If $\hat{C}_n(t/n) \Rightarrow ct$ in R as $n \to \infty$ for each t, where $\hat{C}_n$ is based on $\lambda_n = n$ and $\mu_n = 1$, then $D_n \Rightarrow \Pi_c$ in $D[0,\infty)$, where D_n is the rescaled departure process in* (3.3).

We now give sufficient conditions for D_n to converge in terms of the process $\hat{N}_n^k(t)$ representing the number of customers in phase k of service based on $\lambda_n = n$ and $\mu_n = 1$.

THEOREM 2. *If*

$$\hat{N}_n^k/n \Rightarrow \xi_k \omega \qquad \text{in } D[0,\infty) \quad \text{as} \quad n \to \infty \tag{3.5}$$

for each k, $1 \leq k \leq m$, where ω is defined in (3.1) *and ξ_k is a nonrandom number, then $D_n \Rightarrow \Pi_c$ in $D[0,\infty)$, where Π_c is a Poisson process with intensity*

$$c = \beta \sum_{k=1}^{m} p_k \xi_k . \tag{3.6}$$

PROOF. By Theorem 4.4 of Billingsley, (3.5) implies the convergence

$$\left(\frac{\hat{N}_n^1}{n}, \ldots, \frac{\hat{N}_n^m}{n} \right) \Rightarrow (\xi_1 \omega, \ldots, \xi_m \omega)$$

in $D[0,\infty)^m$. By (2.1) and the continuous mapping theorem, $\hat{\Lambda}_n/n \Rightarrow c\omega$ as $n \to \infty$ in $D[0,\infty)$, where c is as in (3.6). Moreover, by a change of variables,

$$\hat{C}_n(t/n) = \int_0^{t/n} \hat{\Lambda}_n(u)\, du = \int_0^t \frac{\hat{\Lambda}_n(u/n)}{n}\, du, \qquad t \geq 0,$$

so that $\hat{C}_n(t/n) \Rightarrow ct$ in R by the continuous mapping theorem. Finish by applying Corollary 2. ∎

It still remains to determine conditions under which condition (3.5) in Theorem 2 is satisfied. We do not investigate this question in detail here. We observe that in several cases (3.5) is a consequence of existing heavy-traffic limit theorems. In particular, in $GI/GI/\infty$ systems and $GI/GI/s$ systems in which $s_n \to \infty$ quickly, where the arrival process is a renewal process, (3.5) is an immediate consequence of Theorem 3 of Whitt (1982). We have used the same phase-type service-time distributions to make the connection clear. For more general arrival processes, Borovkov (1967) can be applied. For $GI/M/s$ systems in which $s_n \to \infty$ more slowly, so that the probability of delay converges to a nontrivial limit, i.e., so that $(1-\rho_n)\sqrt{s_n} \to \theta$, $0 < \theta < \infty$, (3.5) is an immediate consequence of Theorem 3 of Halfin and Whitt (1981).

For these heavy-traffic limit theorems in which $\lambda_n = n$ and $\mu_n = 1$, the service-time distribution is held fixed while the arrival rate and number of servers go to infinity. For example, the arrival process in the nth system, $\hat{A}_n(t)$, can be defined in terms of a fixed arrival process $A(t)$ by scaling, i.e.,

$$\hat{A}_n(t) = A(nt), \qquad t \geq 0. \tag{3.7}$$

This definition of $\hat{A}_n(t)$ in (3.7) is not the most general, but it is instructive for interpreting the scaling. The scaling in (3.7) causes the rate of both the arrival process $\hat{A}_n(t)$ and the departure process $\hat{D}_n(t)$ to go to infinity as $n \to \infty$. However, as n changes, $\hat{D}_n(t)$ is also approaching a Poisson process, while $\hat{A}_n(t)$ is not changing. If we rescale $\hat{A}_n(t)$ and $\hat{D}_n(t)$ as in (3.3), then we just get back the original arrival process $A_n(t) = A(t)$, but $D_n(t)$ approaches a Poisson process.

REMARK (3.2). It is easy to obtain a discrete analog of §§2–3. The phases should be geometrically distributed instead of exponentially distributed and the arrival process should have all its jumps on the integers. The stationary departure process then approaches a Poisson process on the integers. The number of departures at $1, 2, \ldots, k$ are mutually independent random variables with a Poisson distribution.

4. Fast arrivals. If $\lambda_n \to 0$ or $\lambda_n \to \infty$ as $n \to \infty$, then we must normalize the departure process to get a nondegenerate limit. The case of $\lambda_n = n$ and $\mu_n = 1$ for all n is of particular interest. A limit theorem for this case was obtained by Borovkov (1967) under the assumption that the system is initially empty. To state Borovkov's result, let the service time have the general (not necessarily phase-type) cdf $F(t)$, and let the

system start off empty for each n. Let $\mathbf{D}'_n$ be the normalized process defined by

$$\mathbf{D}'_n \equiv \mathbf{D}'_n(t) = \frac{D_n(t) - n\int_0^t F(t-u)\,du}{\sqrt{n}}, \qquad t \geq 0. \tag{4.1}$$

Let $\hat{\mathbf{A}}_n$ be the normalized processes defined as in §3 in terms of a fixed arrival process A by

$$\hat{\mathbf{A}}_n \equiv \hat{\mathbf{A}}_n(t) = \frac{A(nt) - nt}{\sqrt{n}}, \qquad t \geq 0. \tag{4.2}$$

THEOREM 3 (BOROVKOV). *Consider a sequence of $G/GI/s$ systems, initially empty, with $(s_n - n)/\sqrt{n} \to \infty$. If $\hat{\mathbf{A}}_n \Rightarrow c_a\mathbf{B}$ for $\hat{\mathbf{A}}_n$ in* (4.2), *where B is standard Brownian motion, then*

$$\mathbf{D}'_n \Rightarrow c_a\mathbf{Y}'_2 - \mathbf{Y}'_1 \qquad \text{in } D[0,\infty), \tag{4.3}$$

where $\mathbf{D}'_n$ is defined in (4.1),

$$\mathbf{Y}'_2(t) = \int_0^t F(t-u)\,d\mathbf{B}'(u), \tag{4.4}$$

$\mathbf{B}'$ is a standard Brownian motion, and $\mathbf{Y}_1$ is a nonstationary entered Gaussian process independent of $\mathbf{B}'$ and $\mathbf{Y}'_2$ having covariance function

$$\mathrm{Cov}(\mathbf{Y}'_1(x), \mathbf{Y}'_1(x+t)) = \int_0^x F(u)\big[1 - F(t+u)\big]\,du. \tag{4.5}$$

However, we are primarily interested in the stationary version of the departure process. Let $\hat{\mathbf{D}}_n$ be the normalized random element of $D[0,\infty)$ defined by

$$\hat{\mathbf{D}}_n = \hat{\mathbf{D}}_n(t) = \frac{D_n(t) - nt}{\sqrt{n}}, \qquad t \geq 0, \tag{4.6}$$

where $D_n(t)$ is the stationary version of the departure process, assuming it exists. The following is the stationary analog of Theorem 3, but we do not have a proof.

Conjecture. If, in addition to the assumptions of Theorem 3, a stationary version of the departure process exists for each n, then

$$\hat{\mathbf{D}}_n \Rightarrow c_a\mathbf{Y}_2 - \mathbf{Y}_1 \qquad \text{in } D[0,\infty), \tag{4.7}$$

where $\hat{\mathbf{D}}_n$ is defined in (4.6),

$$\mathbf{Y}_2(t) = \int_0^\infty \big[F(u) - F(u-t)\big]\,d\mathbf{B}'(u), \qquad t \geq 0, \tag{4.8}$$

with $\mathbf{B}'$ a standard Brownian motion and $\mathbf{Y}_1(t) = \mathbf{Z}_1(t) - \mathbf{Z}_1(0)$, where $\mathbf{Z}_1$ is a stationary centered Gaussian process independent of $\mathbf{B}'$ and $\mathbf{Y}_2$ having covariance function

$$r_1(t) = \int_0^\infty F(u)\big[1 - F(t+u)\big]\,du. \tag{4.9}$$

Heuristic reasoning. We obtain the conjecture from Theorem 3 by letting $x \to \infty$ in $\mathbf{D}'_n(t+x) - \mathbf{D}'_n(x)$ and $\mathbf{Z}(t+x) - \mathbf{Z}(x)$ for $\mathbf{Z}(t) = c_a\mathbf{Y}'_2(t) - \mathbf{Y}'_1(t)$. First note that the translation term in $\mathbf{D}'_n(t+x) - \mathbf{D}'_n(x)$ is

$$n\int_x^{x+t} F(u)\,du = n\left[t - \int_x^{x+t}\big[1 - F(u)\big]\,du\right] \to nt \tag{4.10}$$

as $x \to \infty$. Next note that

$$\begin{aligned}\mathbf{Y}_2'(t+x) - \mathbf{Y}_2'(x) &= \int_0^{x+t} \big[F(x+t-u) - F(x-u)\big]\, d\mathbf{B}'(u) \\ &\overset{d}{=} \int_0^{x+t} \big[F(u) - F(u-t)\big]\, d\mathbf{B}'(u) \\ &\Rightarrow \int_0^{\infty} \big[F(u) - F(u-t)\big]\, d\mathbf{B}'(u) \qquad \text{as} \quad x \to \infty,\end{aligned}$$

where $\overset{d}{=}$ means equal in distribution. We also easily obtain (4.9) as the limit of (4.5) as $x \to \infty$.

We now describe the conjectured limit process in more detail.

THEOREM 4. *For the conjectured limit process in* (4.7),

$$\operatorname{Var}\big[c_a \mathbf{Y}_2(t) - \mathbf{Y}_1(t)\big] = t + (c_a^2 - 1)\int_0^{\infty} \big[F(u) - F(u-t)\big]^2\, du. \tag{4.11}$$

PROOF. We use the fact that

$$\int_0^{\infty} \big[F(u) - F(u-t)\big]\, du = t \tag{4.12}$$

for all t and any cdf F on the positive half line. Since $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent,

$$\begin{aligned}\operatorname{Var}\big[c_a \mathbf{Y}_2(t) - \mathbf{Y}_1(t)\big] &= c_a^2 \operatorname{Var} \mathbf{Y}_2(t) + \operatorname{Var} \mathbf{Y}_1(t) \\ &= c_a^2 \int_0^{\infty} \big[F(u) - F(u-t)\big]^2\, du + 2\big[\operatorname{Var} Z_1(0) - r_1(t)\big] \\ &= c_a^2 \int_0^{\infty} \big[F(u) - F(u-t)\big]^2\, dt \\ &\quad + 2\int_0^{\infty} F(u)\big[F(t+u) - F(u)\big]\, du \\ &= (c_a^2 - 1)\int_0^{\infty} \big[F(i) - F(u-t)\big]^2 + \int_0^{\infty} \big[F(u)^2 - F(u-t)^2\big]\, du \\ &= t + (c_a^2 - 1)\int_0^{\infty} \big[F(u) - F(u-t)\big]^2\, du,\end{aligned}$$

where (4.12) with the cdf $F(t)^2$ is used in the last step. ∎

REMARKS (4.1). From Mirasol (1963), we know that the departure process is Poisson in an $M/G/\infty$ system. This is consistent with (4.11) because $c_a^2 = 1$ for a Poisson arrival process, so that $\operatorname{Var}[c_a \mathbf{Y}_2(t) - \mathbf{Y}_1(t)] = t$ as it should. When the service time is deterministic, the stationary departure process is just the arrival process, so that (4.11) should be tc_a^2. Since

$$\int_0^{\infty} \big[F(u) - F(u-t)\big]^2\, du = t \tag{4.13}$$

for deterministic service times with mean 1,

$$\operatorname{Var}\big[c_a Y_2(t) - Y_1(t)\big] = tc_a^2$$

as it should.

(4.2) The variance expression (4.11) enables us to do sensitivity analysis. We can see how changes in the arrival process or service times affect the variability of the

departure process. The integral in (4.11) is a measure of the variability of the service-time distribution, with larger values indicating less variability. As Wolff (1977) noted for the number in system, the way greater variability in the service times affects the variability of the departure process depends on the sign of $c_a^2 - 1$. Since (4.12) holds, the integral in (4.11) is maximized by a deterministic service-time distribution. Hence, the most (least) variable departure processes, as measured by (4.11), are obtained with deterministic service times when $c_a^2 > 1$ ($c_a^2 < 1$).

For related results, see Whitt (1984). The set of possible values of (4.11) for all service-time distributions of mean 1 is the interval $t[c_a^2 \wedge 1, c_a^2 \vee 1]$.

(4.3) Note that

$$\lim_{t\to\infty} t^{-1}\int_0^\infty \left[F(u) - F(u-t)\right]^2 du = 1 \tag{4.14}$$

as $t \to \infty$, so that

$$\lim_{t\to\infty} t^{-1}\mathrm{Var}\left[c_a \mathbf{Y}_2(t) - \mathbf{Y}_1(t)\right] = c_a^2 . \tag{4.15}$$

In the special case of a renewal arrival process and a phase-type service-time distribution, we can obtain a limit theorem for $\hat{\mathbf{D}}_n$ in (4.6). As in (2.12) of Whitt (1982), let $\hat{\mathbf{X}}_n$ be the normalized process induced by $[\hat{N}_n^1(t), \ldots, \hat{N}_n^m(t)]$, where $\hat{N}_n^k(t)$ represents the number of customers in phase k of service at time t in system n, as in (3.5),

$$\hat{\mathbf{X}}_n^k \equiv \hat{\mathbf{X}}_n^k(t) = \frac{\hat{N}_n^k(t) - \xi_k n}{\sqrt{n}}, \qquad t \geqslant 0. \tag{4.16}$$

The following theorem closely parallels Theorem 3 of Whitt (1982) and Theorem 3.2 of Whitt (1984a), so we omit the details. Here too we can apply simple criteria for weak convergence of Markov chains in Stroock and Varadhan (1979).

THEOREM 5. *For GI/GI/s systems with phase-type service,* $(\hat{\mathbf{X}}_n, \hat{\mathbf{D}}_n) \Rightarrow (\mathbf{X}, \mathbf{D})$ *in* $D([0,\infty), R^{m+1})$, *where* $\mathbf{X}_n$ *and* $\mathbf{D}_n$ *are defined in* (4.16) *and* (4.6) *in terms of the stationary processes, and* $(\mathbf{X}, \mathbf{D})$ *is a stationary multivariate diffusion process with infinitesimal mean vector* $m(x) = Mx$ *and infinitesimal covariance function* $\Sigma(x) = \Sigma$, *where* M *and* Σ *are* $(m+1) \times (m+1)$ *matrices.*

To obtain a proof of the conjecture under the conditions of Theorem 5, it remains to show that the stationary Gaussian limit process $\mathbf{D}$ can be represented as $c_a\mathbf{Y}_2 - \mathbf{Y}_1$ in (4.7). This has not yet been done.

5. Other limiting behavior. In this section we briefly describe what happens in other cases when $\lambda_n \to 0$ or $\lambda_n \to \infty$ with $\alpha_n = \lambda_n/\mu_n = n$ and $s_n \to \infty$. We can exploit the basic relation

$$D(t) = A(t) - Q(t) + Q(0), \tag{5.1}$$

where $Q(t)$ is the number of customers in the system at time t, to obtain

$$\frac{D_n(t) - \lambda_n t}{\beta_n} = \frac{A_n(t) - \lambda_n t}{\beta_n} - \frac{Q_n(t) - Q_n(0)}{\beta_n} . \tag{5.2}$$

We assume that

$$\frac{Q_n(t) - Q_n(0)}{\sqrt{n}} \Rightarrow X(t) \qquad \text{as} \quad n \to \infty, \tag{5.3}$$

which we can obtain from the heavy-traffic limit theorems in special cases.

Case 1. First suppose that $\lambda_n \to \infty$. We assume that $A_n(t)$ satisfies a functional limit

theorem of the form

$$\frac{A_n(t) - \lambda_n t}{\sqrt{\lambda_n}} \Rightarrow Y(t) \qquad \text{as} \quad n \to \infty. \tag{5.4}$$

If $\mu_n = \lambda_n/n \to \infty$, then $\beta_n = \sqrt{\lambda_n}$ and

$$\frac{D_n(t) - \lambda_n t}{\sqrt{\lambda_n}} \Rightarrow Y(t) \qquad \text{as} \quad n \to \infty, \tag{5.5}$$

i.e., $D_n(t)$ has the same asymptotic behavior as $A_n(t)$. As a consequence, $D_n(t)$ has the same central limit behavior as a Poisson process if $A_n(t)$ is a renewal process and the renewal intervals have squared coefficient of variation 1.

Case 2. On the other hand, if $\lambda_n \to \infty$ and $\mu_n = \lambda_n/n \to 0$, then $\beta_n = \sqrt{n}$ and

$$\frac{D_n(t) - \lambda_n t}{\sqrt{n}} \Rightarrow X(t) \qquad \text{as} \quad n \to \infty. \tag{5.6}$$

Suppose that $s = \infty$, so that all customers waiting are in service. For the convergence, $Q_n(t) - Q_n(0)$ is asymptotically equivalent to the number of the $Q_n(0)$ customers in service that depart, which has mean and variance of order λ_n. Hence, in this case (5.3) holds with $X(t) = 0$.

Case 3. Finally, suppose $\lambda_n \to 0$. Then $\mu_n = \lambda_n/n \to 0$, $\beta_n = \sqrt{n}$ and the behavior is the same as in Case 2.

6. Two facilities in series. We believe that a good way to examine departure processes and generate approximations for them is to see what kind of congestion they cause as arrival processes to other service facilities, e.g., see Whitt (1983).

Suppose that the departure process we have been analyzing is the arrival process to a second facility with a single server, unlimited waiting room, the first-come first-served discipline and i.i.d. general service times that are independent of the departure process from the first facility. Let n index the average number of busy servers in the first facility ($\alpha_n = n$) and let $\lambda_n = 1$. Let $G_n(t)$ and μ_n be the service-time distribution and the service rate, respectively at the second facility, also indexed by n. A consequence of §3 is that if $G_n(t) = G(t)$ with $\mu_n = \mu > 1$, then as $n \to \infty$ the second facility behaves like a stable $M/G/1$ queue. To see the effect of all the different kinds of limiting behavior possible for the departure process from the first facility, we put the second facility in heavy-traffic by letting μ_n approach 1 from above as $n \to \infty$. By Theorem 1(a) of Iglehart and Whitt (1970), the heavy-traffic behavior at the second facility depends on the central limit theorem behavior for the arrival process to that facility. We only must relate the way $(\mu_n - 1)$ approaches 0 with n. There are three cases:

Case 1. $\sqrt{n}(\mu_n - 1) \to 0$.

The traffic intensity at the second facility is going to the critical value one quickly relative to n, so that we are in Case 1 of §5. The second facility has the same heavy-traffic behavior as if the first facility were not there, i.e., as if the arrival process to the second facility were the arrival process to the first facility.

Case 2. $\sqrt{n}(\mu_n - 1) \to \gamma$ for $0 < \gamma < \infty$.

We are in the setting of §4. The departure process from the first facility has the complicated central limit behavior described there. The heavy-traffic behavior at the second facility is even more complicated. The limiting process can be characterized, but not in a very useful way; see Iglehart and Whitt (1970).

Case 3. $\sqrt{n}(\mu_n - 1) \to \infty$.

The traffic intensity at the second facility is going to the critical value one slowly relatively to n, so that we are in the setting of §3. The heavy-traffic behavior at the second facility is the same as if its arrival process were Poisson.

In closing we remark that Borovkov (1967) has also considered several multi-server facilities in series, where each has many busy servers. His results correspond to the difficult Case 2 above.

Acknowledgement. I am grateful to G. J. Foschini, M. I. Reiman, M. Segal, R. F. Serfozo and the referees for helpful suggestions.

References

[1] Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley and Sons, New York.
[2] Borovkov, A. A. (1967). On Limit Laws for Service Processes in Multi-Channel Systems. *Siberian Math. J.* **8** 746–763.
[3] Bremaud, P. (1981). *Point Processes and Queues*. Springer-Verlag, New York.
[4] Brown, T. C. (1978). Martingales and Point Process Convergence. *Ann. Probab.* **6** 615–628.
[5] ———. (1983). Some Poisson Approximations Using Compensators. *Ann. Probab.* **11** 726–744.
[6] ——— and Pollett, P. K. (1982). Some Distributional Approximations in Markovian Queueing Network. *Adv. Appl. Probab.* **14** 654–671.
[7] Burke, P. J. (1958). The Output Process of a Stationary $M/M/s$ Queueing System. *Ann. Math. Statist.* **39** 1144–1152.
[8] Çinlar, E. (1972). Superposition of Point Processes. In *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis, ed., John Wiley and Sons, New York, 549–606.
[9] Disney, R. L. and König, D. (1984). Queueing Networks: A Survey of Their Random Processes. *SIAM Rev.*, to appear.
[10] Franken, P., König, D., Arndt, U. and Schmidt, V. (1981). *Queues and Point Processes*, Akademie-Verlag, Berlin.
[11] Grandell, J. (1976). *Doubly Stochastic Poisson Processes*. Lecture Notes in Mathematics 529, Springer-Verlag, New York.
[12] Halfin, S. and Whitt, W. (1981). Heavy-Traffic Limits for Queues with Many Exponential Servers. *Oper. Res.* **29** 567–588.
[13] Iglehart, D. L. and Whitt, W. (1970). Multiple Channel Queues in Heavy Traffic II: Sequences, Networks and Batches. *Adv. Appl. Probab.* **2** 355–369.
[14] Mirasol, N. M. (1963). The Output of an $M/G/\infty$ Queueing System is Poisson. *Oper. Res.* **11** 282–284.
[15] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. The Johns Hopkins University Press, Baltimore.
[16] Pollett, P. K. (1982). Distributional Approximations for Networks of Quasireversible Queues. Statistical Laboratory, University of Cambridge.
[17] Serfozo, R. F. (1972). Conditional Poisson Processes. *J. Appl. Probab.* **9** 288–302.
[18] ———. (1977). Compositions, Inverses and Thinnings of Random Measures. *Z. Wahrsch. verw. Gebiete* **37** 253–265.
[19] Straf, M. L. (1972). Weak Convergence of Stochastic Processes with Several Parameters. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.* **2** 187–221.
[20] Stroock, D. W. and Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*. Springer-Verlag, New York.
[21] Whitt, W. (1972). Embedded Renewal Processes in the $GI/G/s$ Queue. *J. Appl. Probab.* **9** 650–658.
[22] ———. (1980). Some Useful Functions for Functional Limit Theorems. *Math. Oper. Res.* **5** 67–85.
[23] ———. (1982). On the Heavy-Traffic Limit Theorem for $GI/G/\infty$ Queues. *Adv. Appl. Probab.* **14** 171–190.
[24] ———. (1982a). Approximating a Point Process by a Renewal Process, I: Two Basic Methods. *Oper. Res.* **30** 125–147.
[25] ———. (1983). Queue Tests for Renewal Processes. *Oper. Res. Letters* **2** 7–12.
[26] ———. (1983a). The Queueing Network Analyzer. *Bell System Tech. J.* **62** 2779–2815.
[27] ———. (1984). Minimizing Delays in the $GI/G/1$ Queue. *Oper. Res.*, to appear.
[28] ———. (1984a). Queues with Superposition Arrival Processes in Heavy Traffic. under review.
[29] Wolff, R. W. (1977). The Effect of Service Time Regularity on System Performance. In *Computer Performance*, K. M. Chandy and M. Reiser, eds. North-Holland, Amsterdam, 297–304.

AT&T BELL LABORATORIES, HOLMDEL, NEW JERSEY 07733