Adv. Appl. Prob. **23**, 188–209 (1991)
Printed in N. Ireland
© *Applied Probability Trust* 1991

A NEW VIEW OF THE HEAVY-TRAFFIC LIMIT THEOREM FOR INFINITE-SERVER QUEUES

PETER W. GLYNN,* *Stanford University*
WARD WHITT, ***AT & T Bell Laboratories*

Abstract

This paper presents a new approach for obtaining heavy-traffic limits for infinite-server queues and open networks of infinite-server queues. The key observation is that infinite-server queues having deterministic service times can easily be analyzed in terms of the arrival counting process. A variant of the same idea applies when the service times take values in a finite set, so this is the key assumption. In addition to new proofs of established results, the paper contains several new results, including limits for the work-in-system process, limits for steady-state distributions, limits for open networks with general customer routes, and rates of convergence. The relatively tractable Gaussian limits are promising approximations for many-server queues and open networks of such queues, possibly with finite waiting rooms.

MANY-SERVER QUEUES; HEAVY TRAFFIC; DIFFUSION APPROXIMATIONS; QUEUEING NETWORKS; GAUSSIAN DISTRIBUTIONS; STRONG APPROXIMATIONS; $G/G/\infty$ QUEUES

1. Introduction

In this paper, we describe a new approach for obtaining heavy-traffic limits for open infinite-server queueing systems. Heavy traffic is achieved by sending the arrival rate to infinity while holding the distribution of the service times fixed. As a consequence, the number of busy servers goes to infinity, thereby justifying the use of the term 'heavy traffic'. The limit processes obtained in this setting are typically non-Markov Gaussian processes having mean and covariance functions that depend on the detailed form of the service-time distributions. This is in contrast to the heavy-traffic limit theory associated with systems having a finite (fixed) number of servers. The limit processes obtained there are typically reflecting Brownian motion Markov diffusion processes having infinitesimal means, variances, and reflection terms that depend on the service-time distributions only through their mean and variance parameters; see for example Reiman (1984).

Even though the limit process is not Markov and it depends on the distribution of

Received 18 January 1990.

* Postal address: Department of Operations Research, Stanford University, Stanford, CA 94305-4022, USA.

Research supported by the U.S. Army Research Office under Contract DAAL-03-88-K-0063 and by a Natural Sciences and Engineering Research Council of Canada grant.

** Postal address: AT & T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974-2070, USA.

the service times in a relatively complicated way, the one-dimensional marginal distributions (and the limiting stationary distribution) are remarkably simple, being Gaussian with tractable expressions for the means and variances. The relatively simple Gaussian limits occur here in part because the boundary at the origin disappears in the limit. Indeed, it is evident that all servers being idle will be a very rare event in a many-server queue with a high arrival rate.

In this paper we consider only infinite-server systems, but it is important to note that the boundary-free Gaussian limits may serve as useful approximations for systems with a large finite number of servers, possibly with a finite waiting room. Indeed, the Gaussian limits also are limits for s-server models with r waiting spaces, $0 \leqq r \leqq \infty$, if s goes to infinity sufficiently fast as the arrival rate increases (so that having all s servers busy is asymptotically negligible).

While Gaussian approximations and heavy-traffic limits for infinite-server queueing systems have already been established by Iglehart (1965), Borovkov (1967), (1984), Newell (1973) and Whitt (1982), we believe that the approach in this paper provides some additional insight. The key observation that we exploit throughout this paper is that infinite-server queues having deterministic service times can easily be analyzed in terms of the counting process that records the cumulative number of arrivals. We then note that a variant of the same idea applies to systems in which service times take values in a finite set. Hence, we consider only service-time distributions that take values in a finite set. From a practical standpoint, we do not view this assumption as being particularly restrictive. In fact, it is similar in spirit to the assumption of Whitt (1982) that the service-time distributions are phase-type.

Our approach has several advantages.

1. We are able to give an elementary proof (using standard weak convergence arguments such as the continuous mapping theorem) to obtain heavy-traffic limits for the queue-length (number of busy servers) and departure processes for a single station. Although these limits have been obtained previously, our proof is entirely probabilistic and avoids the explicit analytical calculation present in the earlier work. (For example, both Iglehart (1965) and Whitt (1982) calculate infinitesimal means and variances of related Markov processes; Borovkov (1967), (1984) works with weak convergence of finite-dimensional distributions.)
2. Our limit processes are explicitly described as functionals of increments of a vector-valued Brownian motion process. This seems attractive from a computational viewpoint, particularly in comparison with the multivariate Ornstein–Uhlenbeck processes that underlie the analysis of Whitt (1982); see Section 5 for further discussion.
3. We obtain new heavy traffic limits for the work-in-system process of an infinite-server queue; see Theorem 3.
4. Our approach enables us to model stochastic dependencies among and between interarrival and service times that can not be analyzed using previous

techniques. (For example, previous papers require that the service time sequence at a given station be i.i.d.) A wide variety of dependencies can be analyzed. Furthermore, our analysis reveals the statistical information that must be gathered to cope successfully with correlated service times; see Section 3.

5. We are able to apply 'strong approximation' techniques to obtain rates of convergence for our limit theorems when the interarrival and service times are independent and i.i.d.; see Theorem 4.
6. We obtain explicit expressions for the steady-state distributions of our queueing processes and show that they, in turn, converge to the steady-states of our Gaussian limits. This complements work of Whitt (1982) and verifies the conjecture on p. 540 of Whitt (1984); see Theorem 6.
7. We are able to obtain other new results for the $GI/D/\infty$ queue (single-station queue with renewal arrivals and deterministic service times); see Theorems 5 and 7.
8. We are able to apply our method to the analysis of networks of infinite-server queues; see Theorem 8 for the heavy-traffic behavior of the vector queue-length process in heavy traffic. As mentioned in (2) above, we believe that our representation in terms of increments of a vector-valued Brownian motion process permits more efficient computation of the limiting distribution than the Ornstein–Uhlenbeck approach in Whitt (1982).

This paper is organized as follows. In Section 2, we describe our basic limit theorems for the queue-length, departure, and work-in-system processes in the setting of a single station. In Section 3 we focus on the basic assumption used to derive these limit theorems. The 'fully independent' $GI/GI/\infty$ model previously analyzed in the literature (in which arrivals come from a renewal process independent of i.i.d. service times) is a special case of our framework, as is the $G/GI/\infty$ model considered by Borovkov (1967) in which independence among the interarrival times is relaxed. In Section 4, we return to the analysis of the single-station system. We obtain rate-of-convergence results, as well as limit theorems for the steady-state distributions of the various queueing processes. In Section 5, we show how to extend the theory of Sections 2–4 to networks. Finally, Section 6 contains all proofs.

2. A simple proof for a single-station system

Suppose that $N \equiv (N(t)\colon\ t \geqq 0)$ is the cumulative number of arrivals to a single infinite-server station during the interval $[0, t]$. Assume that the station is idle at time $t = 0$. (If this were not the case, we would be obligated to describe the residual service times for each of the active servers, in order to obtain a well-defined description of the state of the system at time 0. We avoid this by making our current assumption.) Our fundamental observation is that the process $Q \equiv (Q(t)\colon\ t \geqq 0)$

describing the number of customers at the station at time t is easily described in terms of N when the service times are deterministic. In particular, if the service times equal x almost surely, then

$$Q(t) = N(t) - N(t-x), \qquad t \geqq 0,$$

where we adopt the convention that $N(t) = 0$ for $t < 0$.

To send the station into heavy traffic, we let the arrival rate go to infinity. In particular, we consider a sequence of queueing systems in which the arrival rate to the nth system is of order n. This can be modelled by letting the arrival process $N_n \equiv (N_n(t)\colon t \geqq 0)$ to the nth system be defined by scaling time by n, i.e., through the relation $N_n(t) = N(nt)$. If $Q_n \equiv (Q_n(t)\colon t \geqq 0)$ is the number of customers in the nth system at time t, then

$$Q_n(t) = N(nt) - N(n(t-x)), \qquad t \geqq 0.$$

Thus, Q_n is a simple function of an 'accelerated' version of the process N. Since accelerated counting processes typically satisfy strong laws of large numbers (SLLNs) and functional central limit theorems (FCLTs), it follows that Q_n ought to inherit the same type of behavior.

To pursue this approach for non-deterministic service times, we use the following problem formulation. We assume that all customers that enter the station have service times belonging to the finite set $\{x_1, \cdots, x_m\}$. We say that a customer requiring service time x_i is a customer of type i. For $1 \leqq i \leqq m$, let $N^i(t)$ denote the cumulative number of arrivals of customers of type i to the station by time t. Assuming that the station is idle at time $t = 0$, the number of customers of type i in the system at time t is given by

$$Q^i(t) = N^i(t) - N^i(t - x_i), \qquad t \geqq 0.$$

Suppose that we increase the arrival rate by a factor of n, so that the arrival process for type-i customers in the nth system is given by $N_n^i(t) = N^i(nt)$. Then, the total number of type-i customers in the nth system at time t is given by

$$Q_n^i(t) = N^i(nt) - N^i(n(t - x_i)), \qquad t \geqq 0. \tag{2.1}$$

Let $\vec{N}(t) = (N^1(t), \cdots, N^m(t))$. To obtain suitable limit theorems for $\vec{Q}_n \equiv ((Q_n^1(t), \cdots, Q_n^m(t))\colon t \geqq 0)$ and $Q_n \equiv (Q_n(t)\colon t \geqq 0)$, where $Q_n(t) = Q_n^1(t) + \cdots + Q_n^m(t)$, we shall require that $\vec{N} \equiv (\vec{N}(t)\colon t \geqq 0)$ satisfy an SLLN and an FCLT. Here is the SLLN:

(2.2) there exists a (deterministic) vector $\vec{\lambda} \equiv (\lambda_1, \cdots, \lambda_m) \in \mathbb{R}^m$ such that

$$t^{-1}\vec{N}(t) \to \vec{\lambda} \qquad \text{a.s. as } t \to \infty.$$

The following result is a simple consequence of (2.2). Let $[x]^+ = \max\{x, 0\}$.

Proposition 1. If (2.2) is satisfied, then $n^{-1}\vec{Q}_n(t) \to \vec{m}_Q(t) = (m_Q^1(t), \cdots, m_Q^m(t))$ and $n^{-1}Q_n(t) \to m_Q(t)$ a.s. as $n \to \infty$, where $m_Q^i(t) = \lambda_i(t - [t - x_i]^+)$ and $m_Q(t) = m_Q^1(t) + \cdots + m_Q^m(t)$.

Proposition 1 states that $\vec{Q}_n(t) \approx n\vec{m}_Q(t)$ when n is large (where $\approx$ means 'approximately equal'). From a practical standpoint, this suggests using the approximation $Q(t) \approx \sum_{i=1}^{m} \lambda_i(t-[t-x_i]^+)$, when the arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_m$ are large. To refine this approximation, we require the FCLT hypothesis.

Let $\Rightarrow$ denote convergence in distribution (weak convergence) and let $D_m \equiv D_m[0, \infty)$ be the space of right-continuous $\mathbb{R}^m$-valued functions with left limits having domain $[0, \infty)$, endowed with the standard Skorohod (1956) J_1 topology; see Billingsley (1968), Whitt (1980) and Ethier and Kurtz (1986). Let θ_s: $D_m \rightarrow D_m$ be the shift operator defined by $\theta_s(x)(t) = x(t+s)$, $t+s \geqq 0$ and $\theta_s(x)(t) = 0$, $t+s<0$, $t \geqq 0$. We exploit the fact that θ_s is a continuous operator for $s<0$. (Note that it is *not* continuous for $s>0$. To see this, consider $s=1$, $x(t) = 1_{[1,\infty)}(t)$, and $x_n(t) = 1_{[1+n^{-1},\infty)}(t)$, where $1_A(t)$ is the indicator function of the set A.) For $x \in D_m$, let $\mathrm{Disc}(x)$ be the set of discontinuity points of x in $[0, \infty)$. Here is our key FCLT assumption.

(2.3) Assume that there exist a process $\tilde{Z} \equiv (\tilde{Z}(t)$: $t \geqq 0)$ in D_m for which $P(\mathrm{Disc}(\tilde{Z}_i) \cap \mathrm{Disc}(\theta_s(\tilde{Z}_j))) = 0$ for all i, j and $s \leqq 0$ (for which $i \neq j$ or $i=j$ and $s \neq 0$), a vector $\vec{\lambda}$ and a constant $\gamma > 0$ such that $\tilde{Z}_n \Rightarrow \tilde{Z}$ in D_m as $n \rightarrow \infty$, which we denote by

$$\tilde{Z}_n(t) \equiv n^{\gamma}(n^{-1}\vec{N}(nt) - \vec{\lambda}t) \Rightarrow \tilde{Z}(t) \qquad \text{in } D_m \text{ as } n \rightarrow \infty.$$

Remark 2.1. Under (2.2) and (2.3), the two $\vec{\lambda}$ vectors must coincide, because both (2.2) and (2.3) imply a weak law of large numbers (WLLN).

Remark 2.2. The discontinuity condition in (2.3) is always satisfied if $\tilde{Z}$ has continuous paths which covers the standard case in which $\tilde{Z}$ is an m-dimensional Brownian motion. The discontinuity condition in (2.3) is important to cover cases in which $\tilde{Z}$ does not have continuous paths. In particular, it is easy to see that the discontinuity condition is satisfied if the limit process $\tilde{Z}$ has independent marginals $\tilde{Z}_i$ in D_1 and if each marginal process $\tilde{Z}_i$ has independent increments with $\tilde{Z}_i(t)$ continuous at t almost surely for each t. For example, this covers standard stable process limits (which are composed of independent one-dimensional marginal stable processes).

Given (2.3) we can easily obtain FCLTs for $\vec{Q}_n(t)$ and $Q_n(t)$. Throughout this paper, we adopt the convention that all processes are extended to $(-\infty, 0)$ by setting them identically equal to zero over that interval.

Theorem 1. If (2.3) holds, then

$$n^{\gamma}(n^{-1}\vec{Q}_n(t) - \vec{m}_Q(t)) \Rightarrow \hat{Q}(t) \qquad \text{in } D_m \text{ as } n \rightarrow \infty$$

and

$$n^{\gamma}(n^{-1}Q_n(t) - m_Q(t)) \Rightarrow \tilde{Q}(t) \qquad \text{in } D_1 \text{ as } n \rightarrow \infty,$$

where

$$\hat{Q}(t) = (\hat{Q}_1(t), \cdots, \hat{Q}_m(t)),\ \tilde{Q}(t) = \hat{Q}_1(t) + \cdots + \hat{Q}_m(t)$$

and

$$\hat{Q}_i(t) = \tilde{Z}_i(t) - \tilde{Z}_i(t-x_i), \qquad t \geqq 0.$$

Of course, typically (2.3) holds with $\gamma = \frac{1}{2}$ and $\tilde{Z} = B$, where $B \equiv (B(t)\colon t \geqq 0)$ is an m-dimensional Brownian motion, so that $\tilde{Z}$ is a Gaussian process. Then it is immediate that the limit processes $\hat{Q}$ and $\tilde{Q}$ are also Gaussian, e.g., see p. 87 of Feller (1971). Hence, when $Z = B$, Theorem 1 supplies Gaussian approximations for the distributions of $\vec{Q}(t)$ and $Q(t)$. To use this approximation in a practical setting, recall that if the processes $N^1, \cdots, N^m$ are appropriately uniformly integrable, then the covariance matrix C of the limiting Brownian motion B can be related to that of the counting process $\vec{N}$ as follows:

$$\lim_{t\to\infty} t^{-1} \operatorname{cov}(N^i(t), N^j(t)) = C_{ij} \qquad \text{as } t \to \infty.$$

For $t \geqq \max\{x_1, \cdots, x_m\}$, the ordinary-CLT consequence of Theorem 1 with $\tilde{Z} = B$ and $\gamma = \frac{1}{2}$ can be interpreted as stating that if the arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_m$ are large, then

$$\text{(2.4)} \qquad \vec{Q}(t) \equiv (Q^1(t), \cdots, Q^m(t)) \overset{\mathscr{D}}{\approx} \mathrm{N}(\vec{m}_Q, C^Q) \qquad \text{in } \mathbb{R}^m$$

with

$$\vec{m}_Q \equiv (\lambda_1 x_1, \cdots, \lambda_m x_m), \qquad C^Q_{ij} \equiv C_{ij}(x_i \wedge x_j),$$

and

$$\text{(2.5)} \qquad Q(t) \overset{\mathscr{D}}{\approx} \mathrm{N}\Bigl(\sum_{i=1}^{m} \lambda_i x_i, \sum_{i=1}^{m}\sum_{j=1}^{m} C^Q_{ij}\Bigr),$$

where $\overset{\mathscr{D}}{\approx}$ denotes 'approximate equality in distribution', $x_i \wedge x_j = \min\{x_i, x_j\}$ and $\mathrm{N}(\mu, C)$ denotes a normally distributed random vector with mean vector μ and covariance matrix C (variance in $\mathbb{R}^1$).

We turn next to the study of the departure process $\vec{D}(t) = (D^1(t), \cdots, D^m(t))$ which records the cumulative number of customers of each type to depart the station during the interval $[0, t]$ and the overall departure process $D(t) = D^1(t) + \cdots + D^m(t)$. We note that if the service times are all of duration x almost surely, then the number of departures prior to time t is precisely equal to the number of arrivals prior to time $t - x$. As a consequence, we conclude that with m service types the number of departures of type i prior to time t in the nth queueing system is given by

$$D^i_n(t) = N^i(n(t - x_i)), \qquad t \geqq 0.$$

Theorem 2. (a) If (2.2) is satisfied, then $n^{-1}\vec{D}_n(t) \to \vec{m}_D(t)$ and $n^{-1}D_n(t) \to m_D(t)$ a.s. as $n \to \infty$, where $\vec{m}_D(t) = (m^1_D(t), \cdots, m^m_D(t))$, $m_D(t) = m^1_D(t) + \cdots + m^m_D(t)$ and $m^i_D(t) = \lambda_i[t - x_i]^+$, $t \geqq 0$.

(b) If (2.3) holds, then

$$n^\gamma(n^{-1}\vec{D}_n(t) - \vec{m}_D(t)) \Rightarrow \hat{D}(t) \qquad \text{in } D_m$$

and

$$n^\gamma(n^{-1}D_n(t) - m_D(t)) \Rightarrow \tilde{D}(t) \qquad \text{in } D_1 \text{ as } n \to \infty,$$

where $\hat{D}(t) = (\hat{D}_1(t), \cdots, \hat{D}_m(t))$, $\tilde{D}(t) = \hat{D}_1(t) + \cdots + \hat{D}_m(t)$ and $\hat{D}_i(t) = \tilde{Z}_i(t - x_i)$, $t \geqq 0$.

The proof of Theorem 2 mirrors that of Proposition 1 and Theorem 1 and is therefore omitted. Indeed, the proof of Theorem 1 can easily be modified to obtain a joint FCLT for $[\vec{Q}_n(t), \vec{D}_n(t)]$ with limit $[\hat{Q}, \hat{D}]$ in D_{2m}.

Theorem 2 suggests that with $\tilde{Z} = B$, $\gamma = \frac{1}{2}$ and $t \geqq \max\{x_1, \cdots, x_m\}$, the departure processes $\vec{D}(t)$ and $D(t)$ may be approximated as

$$\vec{D}(t+s) - D(t) \overset{\mathscr{D}}{\approx} \mathrm{N}(\vec{m}_D, C^D) \tag{2.6}$$

with

$$\vec{m}_D \equiv (\lambda_1 s, \cdots, \lambda_m s), \qquad C_{ij}^D = C_{ij}([s - |x_i - x_j|]^+),$$

and

$$D(t+s) - D(t) \overset{\mathscr{D}}{\approx} \mathrm{N}\Big(\sum_{i=1}^m \lambda_i s, \sum_{i=1}^m \sum_{j=1}^m C_{ij}^D\Big) \tag{2.7}$$

for $s \geqq 0$, provided that the arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_m$ are large.

Our final limit theorem in this section is for the work remaining in the system at time t. As a first step in obtaining such a limit theorem, we consider the process $Q^i(t, y)$, defined as the number of type-i customers at the station at time t having a remaining service time greater than y. Note that if the service times are equal to x almost surely, then the customers in the system at time t having remaining service time greater than $y (y < x)$ are precisely those customers that arrived in the interval $(t - x + y, t]$. Hence, if $Q_n^i(t, y)$ is the number of type-i customers in the nth system having a remaining service time greater than y, then

$$Q_n^i(t, y) = N^i(nt) - N^i(n(t - [x_i - y]^+)).$$

Note that for each customer in the system at time t, the remaining work for that customer may be expressed as

$$\int_0^\infty I(\text{remaining service time} > y)\, dy,$$

where $I(A)$ is 1 if A is true and 0 otherwise. Hence, we conclude that the type-i work remaining in the nth system at time t can be expressed as

$$V_n^i(t) = \int_0^{x_i} Q_n^i(t, y)\, dy.$$

We are now ready to state SLLNs and FCLTs for $\vec{Q}_n(t, y) \equiv (Q_n^1(t, y), \cdots, Q_n^m(t, y))$, $Q_n(t, y) \equiv Q_n^1(t, y) + \cdots + Q_n^m(t, y)$, $\vec{V}_n(t) \equiv (V_n^1(t), \cdots, V_n^m(t))$ and $V_n(t) \equiv V_n^1(t) + \cdots + V_n^m(t)$.

Theorem 3. (a) If (2.2) is satisfied, then (for $y \geqq 0$)

$$n^{-1}\vec{Q}_n(t, y) \to \vec{m}_y(t), \qquad n^{-1}Q_n(t, y) \to m_y(t),$$
$$n^{-1}\vec{V}_n(t) \to \vec{m}_V(t) \quad \text{and} \quad n^{-1}V_n(t) \to m_V(t) \qquad \text{a.s. as } n \to \infty,$$

where

$$\vec{m}_y(t) = (m_y^1(t), \cdots, m_y^m(t)), \qquad m_y^i(t) = \lambda_i(t - [t - [x_i - y]^+]^+)$$

$$m_V^i(t) = \int_0^{x_i} m_y^i(t)\,dy \quad \text{and} \quad m_V(t) = m_V^1(t) + \cdots + m_V^m(t).$$

(b) If (2.3) holds, then (for $y \geqq 0$)

$$\begin{aligned} n^\gamma(n^{-1}\vec{Q}_n(t, y) - \vec{m}_y(t)) &\Rightarrow \hat{Q}_y(t) && \text{in } D_m, \\ n^\gamma(n^{-1}Q_n(t, y) - m_y(t)) &\Rightarrow \tilde{Q}_y(t) && \text{in } D_1, \\ n^\gamma(n^{-1}\vec{V}_n(t) - \vec{m}_V(t)) &\Rightarrow \hat{V}(t) && \text{in } D_m, \\ n^\gamma(n^{-1}V_n(t) - m_V(t)) &\Rightarrow \tilde{V}(t) && \text{in } D_1 \text{ as } n \to \infty, \end{aligned}$$

where

$$\hat{Q}_y(t) = (\hat{Q}_{y1}(t), \cdots, \hat{Q}_{ym}(t)), \ \tilde{Q}_y(t) = (\hat{Q}_{y1}(t) + \cdots + \hat{Q}_{ym}(t)),$$
$$\hat{V}(t) = (\hat{V}_1(t), \cdots, \hat{V}_m(t)), \ \tilde{V}(t) = \hat{V}_1(t) + \cdots + \hat{V}_m(t),$$
$$\hat{V}_i(t) = \int_0^{x_i} \hat{Q}_{yi}(t)\,dy \quad \text{and} \quad \hat{Q}_{yi}(t) = \tilde{Z}_i(t) - \tilde{Z}_i(t - [x_i - y]^+).$$

Theorem 3 suggests that if $\tilde{Z} = B$, $\gamma = \frac{1}{2}$, $t \geqq \max\{x_i\colon 1 \leqq i \leqq m\}$ and the arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_m$ are large, then we can use the following approximation for the distribution of the vector of workloads in the system at time t:

(2.8)
$$\vec{V}(t) \overset{\mathscr{D}}{\approx} \mathrm{N}(\vec{m}_V, C^V)$$

with

$$\vec{m}_V = \left(\frac{\lambda_1 x_1^2}{2}, \cdots, \frac{\lambda_m x_m^2}{2}\right), \qquad C_{ij}^V = C_{ij}\left[(x_i \vee x_j)\frac{(x_i \wedge x_j)}{2} - \frac{(x_i \wedge x_j)^3}{6}\right]$$

and

(2.9)
$$\tilde{V}(t) \overset{\mathscr{D}}{\approx} \mathrm{N}\left(\sum_{i=1}^m \frac{\lambda_i x_i^2}{2}, \sum_{i=1}^m \sum_{j=1}^m C_{ij}^V\right).$$

To obtain (2.8) and (2.9), note that $\vec{V}(t)$ and $\tilde{V}(t)$, involving the integral and sum of Gaussian processes, are themselves Gaussian. Furthermore,

$$C_{ij}^V = \int_0^{x_i} \int_0^{x_j} \operatorname{cov}[B_i(t) - B_i(t - x_i + y), B_j(t) - B_j(t - x_j + z)]\,dz\,dy.$$

A straightforward calculation then yields C_{ij}^V in (2.8).

Remark 2.3. It is worth noting that when $\tilde{Z} = B$ the process $\theta_s X = (X(t + s)\colon t \geqq 0)$ is a stationary Gaussian process when $s \geqq \max\{x_i\colon 1 \leqq i \leqq m\}$ and X is any of the following limit processes considered in this section: $\tilde{Z}(u + t) - \tilde{Z}(u)$ for $u \geqq 0$, $\hat{Q}(t)$, $\tilde{Q}(t)$, $\hat{D}(u + t) - \hat{D}(u)$ and $\tilde{D}(u + t) - \tilde{D}(u)$ for $u \geqq 0$, $\hat{Q}_y(t)$, $\tilde{Q}_y(t)$, $\hat{V}(t)$ and $\tilde{V}(t)$. Hence, these limit processes obtained here have the interesting property that they exhibit steady-state behavior in finite time (as the original processes do with a Poisson arrival process).

Remark 2.4. From a practical standpoint, it is easy to see that the approximations (2.4)–(2.9) actually hold whenever at least one of the λ_i's is large. (In other words, we do not need all the λ_i's to be simultaneously large.) For example, suppose λ_1 is large and $\lambda_2, \cdots, \lambda_m$ are small by comparison. Then, Gaussian approximations hold for the type 1 customers. On the other hand, the type i customers for $i \geqq 2$ contribute very little to the limit. For $i \geqq 2$, we conclude that the content relative to type 1 is negligible.

Remark 2.5. The FCLTs in Theorems 1–3 can be combined to obtain a joint FCLT for all the processes considered.

3. Verification of the basic assumptions

In this section, we discuss in greater detail the assumptions (2.2) and (2.3) that were critical to the analysis of Section 2. We start by noting that assumptions (2.2) and (2.3) work with the counting processes $N^1, N^2, \cdots, N^m$ as the primitive modelling elements, i.e., the model is directly formulated in terms of these m counting processes. In certain applications, this may be a reasonable starting point for the analysis. For example, in some manufacturing applications, one may have m different products being processed, each with its own characteristic (deterministic) processing time. The only stochastic elements that enter the picture are the arrival instants of the individual jobs to be processed. In such a setting, using the counting processes $N^1, N^2, \cdots, N^m$ as primitive modelling elements may be quite natural.

Even when it is natural to consider the m counting processes $N^1, N^2, \cdots, N^m$, the data may naturally consist of arrival times for each of the m streams of customers. In particular, suppose that $A(i, n)$ is the instant at which the nth customer of type i arrives to the queue. If $\vec{A}(n) = (A(1, n), \cdots, A(m, n))$ are the data, then we need to obtain properties of $\vec{N}$ from properties of $\vec{A}$. Fortunately, it is known that in great generality SLLNs and FCLTs hold for $\vec{N}$ if and only if they do for $\vec{A}$, and the limits are directly related; see Iglehart and Whitt (1971), Vervaat (1972) and Section 7 of Whitt (1980). We state a specific consequence of this theory here without proof. Let $\lfloor x \rfloor$ be the greatest integer less than or equal to x. Let $\vec{\lambda}^{-1} = (\lambda_1^{-1}, \cdots, \lambda_m^{-1})$.

Proposition 2. Suppose that $0 < \lambda_i < \infty$ for $1 \leqq i \leqq m$.

(a) $t^{-1}N^i(t) \to \lambda_i$ a.s. as $t \to \infty$ if and only if $n^{-1}A(i, n) \to \lambda_i^{-1}$ a.s. as $n \to \infty$;

(b) $\tilde{Z}_n(t) \equiv n^{\frac{1}{2}}(n^{-1}\vec{N}(nt) - \vec{\lambda}t) \Rightarrow \tilde{Z}(t)$ in D_m as $n \to \infty$ with $\tilde{Z}$ having continuous paths almost surely if and only if

$$Z_n^*(t) \equiv n^{\frac{1}{2}}(n^{-1}\vec{A}(\lfloor nt \rfloor) - \vec{\lambda}^{-1}t) \Rightarrow Z^*(t) \qquad \text{in } D_m \text{ as } n \to \infty$$

with Z^* having continuous paths almost surely. If the limits exist, then $\tilde{Z}_i(t) = -\lambda_i Z_i^*(\lambda_i t)$, $1 \leqq i \leqq m$. In addition, if the limits exist, then Z^* is a centered Brownian motion if and only if $\tilde{Z}$ is a centered Brownian motion, in which case the covariance matrices C^* and $\tilde{C}$ are related by $\tilde{C}_{ij} = (\lambda_i \lambda_j)^{\frac{3}{2}} C_{ij}^*$.

Remark 3.1. The limit processes $\tilde{Z}$ and Z^* in Proposition 2(b) need not have continuous paths if we switch to the Skorohod (1956) M_1 topology; see Theorem 7.5 of Whitt (1980).

In some applications contexts, it is rather unnatural to work directly in terms of the counting processes $N^1, \cdots, N^m$. For example, the standard queueing modelling approach is to consider a single stream of homogeneous customers that are handed out service times $S_1, S_2, \cdots$ stochastically. Assuming that the range of the S_i's consists of the finite set $\{x_1, \cdots, x_m\}$, we can set up counting processes $I_1(n), \cdots, I_m(n)$ defined as

$$I_j(n) = \sum_{k=1}^{n} I(S_k = x_j).$$

Thus, $I_j(n)$ counts the number of customers, out of the first n to arrive, that are assigned service time x_j.

The process $\vec{N} \equiv (N^1, \cdots, N^m)$ may be described in terms of the process $\vec{I}(n) = (I_1(n), \cdots, I_m(n))$ and the arrival time sequence $A(n)$, where $A(n)$ denotes the arrival time of the nth customer to the station. We now relate the SLLN and FCLT behavior of the processes $(\vec{I}(n), A(n))$ and $\vec{N}$.

Proposition 3. (a) Suppose that there exists λ, $0<\lambda<\infty$, such that $n^{-1}A(n) \to \lambda^{-1}$ a.s. as $n\to\infty$. If $n^{-1}\vec{I}(n)\to\vec{p}$ a.s. as $n\to\infty$, then $t^{-1}\vec{N}(t)\to\lambda\vec{p}$ a.s. as $t\to\infty$.

(b) Suppose that λ, $0<\lambda<\infty$, and the vector $\vec{p}$ are deterministic. If

$$\tilde{Y}_n(t) \equiv n^{\frac{1}{2}}\left(\frac{\vec{I}(\lfloor nt\rfloor)}{n} - \vec{p}t, \frac{A(\lfloor nt\rfloor)}{n} - \lambda^{-1}t\right) \Rightarrow \tilde{Y}(t) \quad \text{in } D_{m+1} \text{ as } n\to\infty,$$

then

$$\tilde{Z}_n(t) \equiv n^{\frac{1}{2}}(n^{-1}\vec{N}(nt) - \vec{\lambda}t) \Rightarrow \tilde{Z}(t) \quad \text{in } D_m \text{ as } n\to\infty,$$

where

$$\tilde{Z}_i(t) = \tilde{Y}_i(\lambda t) - p_i\lambda\tilde{Y}_{m+1}(\lambda t), \qquad t \geqq 0.$$

In addition, if $\tilde{Y}(t) = B$, a centered $(m+1)$-dimensional Brownian motion with covariance matrix $\tilde{C}$, then $\tilde{Z}$ is a centered m-dimensional Brownian motion with covariance matrix C, where

$$C_{ij} = \lambda\tilde{C}_{ij} - \lambda^2 p_i\tilde{C}_{i,m+1} - \lambda^2 p_j\tilde{C}_{j,m+1} + \lambda^3 p_i p_j\tilde{C}_{m+1,m+1}.$$

Proposition 3 provides a general basis for approximations. It also indicates the relevant data, namely, λ, p_i $(1\leqq i\leqq m)$ and the $(m+1)\times(m+1)$ covariance matrix C of the limiting Brownian motion. As an application of Proposition 3, suppose that the service times $(S_n : n\geqq 1)$ are i.i.d. (with probability mass function $P(S_n = x_i) = p_i$) and independent of the arrival-time sequence $(A(n)$: $n\geqq 1)$. If we further assume that $A(n)$ satisfies a FCLT of the form

$$n^{\frac{1}{2}}\left(\frac{A(\lfloor nt\rfloor)}{n} - \lambda^{-1}t\right) \Rightarrow \sigma B_0(t), \tag{3.1}$$

where B_0 is a standard (one-dimensional) Brownian motion, then the assumptions of Proposition 3 are satisfied, so that $\tilde{Y}_n(t) \Rightarrow \tilde{Y}(t)$, where $\tilde{Y}$ is a zero-drift Brownian motion with covariance matrix elements $\tilde{C}_{ii} = p_i\,(1-p_i)\;(1 \leqq i \leqq m)$, $\tilde{C}_{m+1,m+1} = \sigma^2$, $\tilde{C}_{ij} = -p_i p_j\;(1 \leqq i,\ j \leqq m)$, and $\tilde{C}_{i,m+1} = \tilde{C}_{m+1,i} = 0$. Then, Proposition 3 states that the covariance matrix C of the Brownian motion limit process $\tilde{Z}$ appearing in (2.3) is given by:

$$(3.2)\qquad \begin{aligned} C_{ij} &= \begin{cases} \lambda p_i(1-p_i) + \lambda^3 p_i^2 \sigma^2, & i = j \\ -\lambda p_i p_j + \lambda^3 p_i p_j \sigma^2, & i \neq j \end{cases} \\ &= \lambda p_i \delta_{ij} + (\lambda^2\sigma^2 - 1)\lambda p_i p_j, \end{aligned}$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise.

Assuming that the covariance matrix C takes the form (3.2), the covariance function of the limiting Gaussian process $\tilde{Q}$ in Theorem 1 may be represented as

$$(3.3)\quad \operatorname{cov}(\tilde{Q}(s), \tilde{Q}(s+t)) = \lambda \int_0^s H(u)\bar{H}(t+u)\,du + \sigma^2\lambda^3 \int_0^s \bar{H}(t+u)\bar{H}(u)\,du$$

for $s,\ t \geqq 0$, where $H(u) = \sum_{x_k : x_k > \mu} p_k$ and $\bar{H}(u) = 1 - H(u)$, which agrees with previous results; see for example p. 176 of Whitt (1982).

From (3.2), we see that the covariance matrix C is diagonal when $\sigma^2\lambda^2 = 1$, implying that each of the arrival streams $N^1, \cdots, N^m$ can be approximated by independent Brownian motions. As might be expected, this leads to enormous simplification in all the approximations derived in Section 2. Since the condition $\sigma^2\lambda^2 = 1$ is always satisfied when N is a Poisson process, the condition $\sigma^2\lambda^2 = 1$ may be interpreted as stating that the arrival process, when appropriately rescaled, behaves like a Poisson process. For the $M/G/\infty$ system, it is well known that the number of busy servers and the cumulative number of departures at time t are independent Poisson random variables for each t; see pp. 18, 29 of Ross (1970). The mean (and variance) of the number of busy servers at time t is then $\lambda \int_0^t \bar{H}(u)\,du$; i.e., the time-dependent mean and variance in the Gaussian approximation are then exact.

The set-up described above, in which the service times are i.i.d. and independent of an arrival process satisfying (3.1), is the context in which all previous limits for the infinite-server queue have been established. Historically, the first contribution to this area was that of Iglehart (1965), in which the case of exponential interarrival-time and service-time distributions was considered. Borovkov (1967) (1984) then extended Iglehart's result to the above 'independent' framework, thereby permitting interarrival and service time distributions to be non-exponential. Later, Whitt (1982) provided a simplified proof of an important special case of Borovkov's result, namely that in which the interarrival times are i.i.d. (rather than merely satisfying (3.1)) and the service times are of phase type. The latter argument gives a clear explanation as to why the infinite-server Gaussian limit (see the definition of $\tilde{Q}$ in Theorem 1) is typically non-Markov. It arises from the fact that one needs to keep track of the residual service times of the customers in the system in order to obtain a limit process which is Markov; see also Glynn (1982). One characteristic of the above proofs is that they are all highly analytical, in the sense that they all involve reasonably explicit computations involving convergence of the finite-dimensional distributions (for example, both Iglehart (1965) and Whitt (1982) calculate certain infinitesimal means and covariances for related Markov processes).

By contrast, the approach taken here is entirely probabilistic and involves only standard weak convergence tools, such as continuous mapping ideas and converging-together arguments. While it is limited to service times that are discrete-valued, this is not a significant limitation from a practical standpoint. Furthermore, the approach taken in this paper allows for significantly more complicated dependencies between interarrival times and service times. Our analysis clearly shows that, in contrast to the heavy-traffic limit theory for a single-server station, an FCLT for the service time sequence does not provide adequate information in the current setting. Rather, as suggested by Theorem 3, one needs to assume that an invariance principle holds for the 'empirical density' process $(I_1(n), \cdots, I_m(n))$. Thus, in predicting the performance of a system having a large number of servers, one needs to gather significantly more statistical information than in the heavy-traffic single-server context. The theory that we have developed here points to the type of information that needs to be collected (namely, estimates for λ, $\vec{p}$, and $\tilde{\bar{C}}$).

We conclude this section by describing the covariance matrix C in one special dependent setting. Suppose that $\{(S_n, A(n)-A(n-1))\colon n \geqq 1\}$ is regenerative. For example, the service times and interarrival times may jointly form a finite-state irreducible Markov chain. If the first regeneration occurs at $n=1$ and T_1 denotes the time of the next regeneration, then it is easy to show that

$$C_{ij} = E\left[\left(\sum_{n=1}^{T_1-1} I(S_n = x_i) - p_i\lambda A(T_1-1)\right)\left(\sum_{n=1}^{T_1-1} I(S_n = x_j) - p_j\lambda A(T_1-1)\right)\right]\Big/E[A(T_1-1)]$$

$$p_i = E\left[\sum_{n=1}^{T_1-1} I(S_n = x_i)\right]\Big/E(T_1-1)$$

$$\lambda = E[A(T_1-1)]/E(T_1-1),$$

if $E[T_1^2 + A^2(T_1)] < \infty$. In addition to providing a basis for calculation of λ, $\vec{p}$ and C given model structure, these formulas provide a basis for estimating them from data.

4. More on the heavy-traffic limit for a single station

In this section, we return to the heavy-traffic limit theory of Section 2, and further investigate its structure. Our first result is a rate-of-convergence theorem for the 'fully independent' $GI/GI/\infty$ single-station model. Specifically, we assume that the customer interrival-time sequence $\{A(n)-A(n-1)\colon n \geqq 1\}$ is i.i.d. (with $A(0)=0$) and independent of an i.i.d. service-time sequence $\{S_n\colon n \geqq 1\}$, where as before the S_n's take values in the finite set $\{x_1, \cdots, x_m\}$.

Theorem 4. Suppose that $EA(1)^p < \infty$ for some $p > 2$ in the $GI/GI/\infty$ model. Then, for each $t \geqq 0$,

$$P(n^{\frac{1}{2}}(n^{-1}Q_n(t) - m_Q(t)) \leqq x) = P(\tilde{Q}(t) \leqq x) + o(n^{-(p-1)/2(p+1)}),$$
$$P(n^{\frac{1}{2}}(n^{-1}D_n(t) - m_D(t)) \leqq x) = P(\tilde{D}(t) \leqq x) + o(n^{-(p-1)/2(p+1)})$$

and

$$P(n^{\frac{1}{2}}(n^{-1}V_n(t) - m_V(t)) \leqq x) = P(\tilde{V}(t) \leqq x) + o(n^{-(p-1)/2(p+1)}) \qquad \text{as } n \to \infty,$$

where $\tilde{Q}(t)$, $\tilde{D}(t)$ and $\tilde{V}(t)$ have the Gaussian distributions in (2.5), (2.7) and (2.9) arising when $\tilde{Z}$ in (2.3) is Brownian motion.

The proof of Theorem 4 is based on the theory of strong approximation as in Csörgő and Révész (1981)); see Csörgő et al. (1987a) for related results in queueing.

Remark 4.1. The rates of convergence in Theorem 4 are expressed for the one-dimensional marginal distributions, which are typically of principal interest in applications. However, as can be seen from the proof, the rates also apply to the random elements of $D([0, T])$ using the uniform metric on $D([0, T])$ and the Prohorov metric on the space of probability measures.

Our next result combines strong approximation theory and the fluctuation theory for Brownian motion to obtain an approximation for the maximum of the queue-length process of a $GI/D/\infty$ queue in heavy traffic.

Theorem 5. Consider a $GI/D/\infty$ queue in which $E\exp(\delta A(1)) < \infty$ for some $\delta > 0$. If $t_n \to \infty$ satisfies $t_n = O(\exp(n^r))$ for some $0 \leqq r < 1$, then

$$\sup_{0 \leqq t \leqq t_n} \left(\frac{Q_n(t) - nm_Q(t)}{n^{\frac{1}{2}}\sigma\lambda^{\frac{3}{2}}}\right)(2x \log(t_n/x))^{-\frac{1}{2}} \Rightarrow 1 \qquad \text{as } n \to \infty,$$

where $\lambda^{-1} = EA(1)$ and $\sigma^2 = \operatorname{var} A(1)$.

Our next results pertain to approximations for steady-state versions of the queue-length, departure, and work-in-system processes. Our goal is to show that these steady-state processes converge to the steady-states of the corresponding limiting processes. To define the steady-state versions of these various processes, we consider the shifted processes $\theta_s(Q_n)$, $\theta_s(V_n)$, $\theta_s(\Delta D_n)$ and $\theta_s(\Delta N_n)$, where $\Delta X(t) = X(t) - X(0)$, e.g.,

$$\theta_s(Q_n)(t) = Q_n(s+t), \qquad t \geqq 0,$$
$$\theta_s(\Delta N_n)(t) = N_n(s+t) - N_n(s), \ t \geqq 0.$$

Proposition 4. Let S_k take values in $\{x_1, \cdots, x_m\}$ and assume that $(S_k : k \geqq 1)$ is a stationary sequence independent of the arrival process N. If N is a renewal process in which $A(1)$ has a continuous c.d.f with $EA(1) = \lambda^{-1} < \infty$, then

$$[\theta_s(\Delta N_n), \theta_s(\Delta D_n), \theta_s(Q_n), \theta_s(V_n)] \Rightarrow [\bar{N}_n, \bar{D}_n, \bar{Q}_n, \bar{V}_n] \text{ in } D_1^4 \qquad \text{as } s \to \infty,$$

where $\bar{D}_n$, $\bar{Q}_n$ and $\bar{V}_n$ are the corresponding queueing processes associated with $(S_k : k \geqq 1)$ and $\bar{N}_n$. Moreover, $\bar{Q}_n$ and $\bar{V}_n$ are stationary processes, while $\bar{N}_n$ and $\bar{D}_n$ have stationary increments.

We now state our heavy-traffic limit theorem for the steady-state processes $\bar{Q}_n$, $\bar{D}_n$ and $\bar{V}_n$. Let Q^*, D^* and V^* be the stationary versions of the limit processes from Section 2, i.e.,

$$Q^* = \theta_{m^*}(\tilde{Q}),\ D^* = \theta_{m^*}(\Delta \tilde{D})$$

and $V^* = \theta_{m^*}(\tilde{V})$, where $m^* = \max\{x_i : 1 \leqq i \leqq m\}$. (Under the following assumptions, the limits in Section 2 will exist.) As noted earlier, Q^* and V^* are stationary Gaussian processes, whereas D^* is Gaussian with stationary increments.

Theorem 6. In addition to the assumptions of Proposition 4, assume that $EA(1)^2 < \infty$ and

$$\beta_n(t) \equiv n^{\frac{1}{2}}(n^{-1}\vec{I}(\lfloor nt \rfloor) - \vec{p}t) \Rightarrow B(t) \text{ in } D_m \qquad \text{as } n \to \infty,$$

where B is m-dimensional Brownian motion. Then Proposition 3(b) holds for $\bar{N}$ as well as N with $\tilde{Y}$ the common Brownian motion limit, so that

$$\begin{aligned}
&n^{-\frac{1}{2}}\Big(\bar{Q}_n(t) - n\lambda \sum_{i=1}^{m} p_i x_i\Big) \Rightarrow Q^*(t) && \text{in } D_1,\\
&n^{-1/2}(\bar{D}_n(t) - n\lambda t) \Rightarrow D^*(t) && \text{in } D_1,\\
&n^{-1/2}\Big(\bar{V}_n(t) - \frac{n\lambda}{2} \sum_{i=1}^{m} p_i x_i^2\Big) \Rightarrow V^*(t) && \text{in } D_1 \text{ as } n \to \infty,
\end{aligned}$$

where $\lambda^{-1} = EA(1)$.

Theorem 6 supplements Theorem 2 of Whitt (1982), which applied a stochastic-order argument to prove that a normalized version of $\bar{Q}_n(t)$ converges to $Q^*(t)$ when the service times are i.i.d. and exponential. Our result for $\bar{D}_n$ verifies a conjecture stated on p. 540 of Whitt (1984); this conjecture was used in several of the results stated there.

Our final theorem of this section is a rate of convergence result for the steady-state distributions described above. Specifically, we shall obtain a Berry–Esséen theorem for the central limit approximation to the distribution of $\bar{Q}_n(t)$ where the system under consideration is the $GI/D/\infty$ model. It shows that the error in the Gaussian approximation for $\bar{Q}_n(t)$ is roughly of the order of the reciprocal of the square root of the arrival rate.

Theorem 7. Consider a $GI/D/\infty$ queue in which $A(1)$ has a continuous c.d.f. and satisfies $EA(1)^4 < \infty$. If $S_n \equiv x$, $\lambda^{-1} = EA(1)$ and $\sigma = (\text{var}\, A(1))^{1/2} > 0$, then

$$P\{n^{-\frac{1}{2}}(\bar{Q}_n(t) - n\lambda x) \leqq \sigma \lambda^{\frac{3}{2}} x^{\frac{1}{2}} y\} = P\{\mathrm{N}(0,1) \leqq y\} + O(n^{-\frac{1}{2}}) \qquad \text{as } n \to \infty.$$

5. Extension to networks

Our purpose now is to show how the theory developed for single stations can be applied to networks. We assume that the network consists of d stations and m customer classes ($d<\infty$ and $m<\infty$). We assume that the routing is deterministic for each class. Customers in class i visit $l(i)<\infty$ stations, not necessarily distinct, before leaving the network. Let $s_i(j)$ be the jth station visited by customers in class i. We also assume that class-i customers receive deterministic service time s_{ij} at the jth station along their route. Thus, a customer of the ith class that enters the network at time t leaves station $s_i(j)$ at time $t+x_{ij}$, where $x_{ij}=s_{i1}+\cdots+s_{ij}$.

Let $N^i(t)$ be the total number of class-i customers to arrive to the network in the interval $[0,t]$. We say that a customer of class i is at stage j of his route when the customer is at the jth station of his route, which is station $s_i(j)$. Then the number of class-i customers at stage j of their route at time t is given by $N^i(t-x_{i,j-1})-N^i(t-x_{ij})$. Hence, the total number of customers at station k at time t is just

$$Q^k(t)=\sum_{i=1}^{m}\sum_{j=1}^{l(i)}\delta(s_i(j),k)[N^i(t-x_{i,j-1})-N^i(t-x_{ij})],$$

where $\delta(j,k)=1$ if $j=k$ and 0 otherwise. We accelerate the arrival rate by a factor of n to construct the nth queueing system. Hence, the vector queue-length process $\vec{Q}_n(t)=(Q_n^1(t),\cdots,Q_n^d(t))$ associated with the nth system has components given by

$$Q_n^k(t)=\sum_{i=1}^{m}\sum_{j=1}^{l(i)}\delta(s_i(j),k)[N^i(n(t-x_{i,j-1}))-N^i(n(t-x_{ij}))].$$

We can now argue precisely as in Theorem 1 to obtain the following limit theorem for $\vec{Q}_n(t)$.

Theorem 8. If (2.3) holds, then

$$n^{\gamma}(n^{-1}\vec{Q}_n(t)-\vec{m}(t))\Rightarrow Q^*(\cdot) \qquad \text{in } D_d \text{ as } n\to\infty,$$

where $\vec{m}(t)=(m^1(t),\cdots,m^d(t))$ and $Q^*(t)=(Q_1^*(t),\cdots,Q_d^*(t))$, with

$$m^k(t)=\sum_{i=1}^{m}\sum_{j=1}^{l(i)}\delta(s_i(j),k)\lambda_i([t-x_{i,j-1}]^+-[t-x_{ij}]^+)$$

$$Q_k^*(t)=\sum_{i=1}^{m}\sum_{j=1}^{l(i)}\delta(s_i(j),k)[\tilde{Z}_i(t-x_{i,j-1})-\tilde{Z}_i(t-x_{ij})].$$

Similar limit theorems can be obtained for other processes, such as the cumulative departure processes, both internal and external to the network, and the vector workload process.

Network limit theorems for the vector queue-length process are also described in Whitt (1982). The approach taken there is to assume that all but one of the arrival streams for the various customer classes are Poisson (with the remaining arrival stream being a general renewal process); the service times are independent random variables of phase type. While the number of customer classes m in our set-up would typically grow rapidly as a function of d, similar growth appears in the phase-type analysis. Furthermore, the limit processes that appear in Whitt (1982) are multivariate Ornstein–Uhlenbeck processes, whereas ours is a sum of increments of Brownian motion. The covariance function of an Ornstein–Uhlenbeck process is obtained by solving a system of linear differential equations. On the other hand, the covariance structure of Q^* above can be easily analyzed by using the stationary independent increments property of the vector-valued Brownian motion $\vec{B}$. As a consequence, it seems likely that the approach taken in Theorem 8 is algorithmically more attractive for calculating the covariance structure of the Gaussian limit.

It is important to note that we can extend the approach beyond deterministic service times and deterministic routing. First, suppose that we wish to assume for a given customer class that the service times experienced along the route are random variables. This can be done by approximating the joint distribution of the service time sequence by a distribution with finite support on the appropriate product space. We then split the original customer class according to the number of point masses in the discrete approximation to the joint distribution. Each sub-class thus created then experiences only deterministic service times. Hence, it can be captured within our current framework. Similarly, we can split the original customer classes to represent stochastic routing. Thus, the approach in this paper permits significant modelling flexibility.

Remark 5.1. It is important to note that in the case of Poisson arrivals with general routing (independent of the arrival process), the exact distribution has a relatively simple product-form for each t; i.e., the queue lengths are independent, Poisson distributed and independent of the Poisson number of departures by time t; see Harrison and Lemoine (1981). Then the means and variances in the Gaussian approximations are exact, so that the approximation reduces to familiar Gaussian approximations for the Poisson distributions.

6. Proofs

Proof of Proposition 1. Under (2.2), $\vec{N}(nt)/n = t\vec{N}(nt)/(nt) \to t\vec{\lambda}$ a.s. as $n \to \infty$. Hence, for each $s \geqq 0$, $N^i(ns)/n \to \lambda_i s$ a.s. as $n \to \infty$. Plugging these limits into (2.1) yields the result.

Proof of Theorem 1. Observe that $n^\gamma(n^{-1}\vec{Q}_n(t) - \vec{m}_Q(t)) = g(\tilde{Z}_n)(t)$, where $g: D_m \to D_m$ is defined by $g(z)_i(t) = z_i(t) - z_i(t - x_i)$. Hence, we can apply the continuous mapping theorem, Theorem 5.1 of Billingsley (1968), to conclude that $g(\tilde{Z}_n) \Rightarrow g(\tilde{Z})$ in D_m. The map g is measurable and continuous almost surely with respect to $\tilde{Z}$ by Theorem 4.1 of Whitt (1980), the discontinuity condition in (2.3) and the continuity of the shift operator θ_s for $s < 0$. The discontinuity condition in

(2.3) implies that

$$P\left(\bigcup_{i=1}^{m}[\mathrm{Disc}\,(\tilde{Z}_i)\cap \mathrm{Disc}\,(\theta_{-x_i}\tilde{Z}_i)]\right)=0,$$

which is what we need here. Next, we obtain $n^{\gamma}(n^{-1}Q_n(t)-m_Q(t))\Rightarrow\tilde{Q}(t)$ by applying the continuity map with addition, again invoking Theorem 4.1 of Whitt (1980) and the discontinuity condition in (2.3). Here we need the union of all $\binom{2m}{2}$ pairwise intersections of the sets $\mathrm{Disc}\,(\tilde{Z}_i)$ and $\mathrm{Disc}\,(\theta_{-x_i}\tilde{Z}_i)$ to have P-measure 0, which is implied by the continuity condition.

Proof of Theorem 3. The results of $\vec{Q}_n(t, y)$ and $Q_n(t, y)$ follow as in the proof of Theorem 1. To obtain the limits for $\vec{V}$, we apply the continuous mapping theorem once again with the functions $h: D_m\rightarrow D_m$ defined by

$$h(z)_i(t)=\int_0^{x_i}[z_i(t)-z_i(t-x_i+y)]\,dy,$$

which is continuous (see Lemma 1 below), and simple addition, which is also continuous because $h(z)$ has continuous paths for all $z\in D_m$. To do part (a), note that (2.2) implies the corresponding FSLLN, i.e., that $n^{-1}\vec{N}(nt)\rightarrow\vec{\lambda}t$ in D_m as $n\rightarrow\infty$ a.s.; see Theorem 4 of Glynn and Whitt (1988).

To show that the function h above is indeed continuous, we apply the following lemma.

Lemma 1. The function g defined by $g(z)(t)=\int_a^b z(t-y)\,dy$, $c\leqq t\leqq d$, is a continuous mapping of $D([a-d, b-c], R)$ into $C([c, d], R)$, where D is endowed with any of the Skorohod (1956) topologies and C is endowed with the uniform topology.

Proof. Suppose that $z_n\rightarrow z$ as $n\rightarrow\infty$ in $D([a-d, b-c], R)$ endowed with one of the Skorohod (1956) topologies. This implies convergence in Skorohod's weakest M_2 topology, which is equivalent to convergence in the Hausdorff metric m applied to the completed graphs in $[a-d, b-c]\times R$, where the completed graph is the set

$$\Gamma(z)=\{(t, x): z(t-)\leqq x\leqq z(t),\ a-d\leqq t\leqq b-c\};$$

see Pomarede (1976). It is easy to see that $|g(z_n)(t)-g(z)(t)|$ is bounded above by the area of the ε-neighborhood of z in the Hausdorff metric when $m(z_n, z)\leqq\varepsilon$. It is easy to see, using Lemma 1 on p. 110 of Billingsley (1968), that for each z this area converges to 0 as $m(z_n, z)\rightarrow 0$. To see that $g(z)$ is continuous, note that

$$\begin{aligned}|g(z)(t+\varepsilon)-g(z)(t)| &= \left|\int_a^b z(t+\varepsilon-y)-z(t-y)\,dy\right|\\ &\leqq\left|\int_{a-\varepsilon}^{a} z(t-y)-\int_{b-\varepsilon}^{b} z(t-y)\,dy\right|\\ &\leqq 2\varepsilon\sup_{a-\varepsilon\leqq t\leqq b}|z(t)|.\end{aligned}$$

Proof of Proposition 3. (a) Let $N(t)$ count the number of arrivals during $[0, t]$. As in part (i) of Proposition 2, $N(t)/t \to \lambda$ a.s. as $t \to \infty$. Then, $N^i(t) = \sum_{j=1}^{N(t)} I(S_j = x_i) = I_i(N(t))$, so $N^i(t)/t = I_i(N(t))/(N(t)) \cdot (N(t)/t) \to p_i\lambda$ a.s. as $t \to \infty$.

(b) Observe that

$$\begin{aligned}\tilde{Z}_{ni}(t) = \tilde{Y}_{ni}(n^{-1}N(nt)) - p_i\lambda\tilde{Y}_{n,m+1}(n^{-1}N(nt)) \\ + n^{-\frac{1}{2}}p_i\lambda(A(N(nt)) - nt), \qquad t \geqq 0.\end{aligned} \tag{6.1}$$

By Proposition 2,

$$n^{\frac{1}{2}}\left(\frac{A(nt)}{n} - \lambda^{-1}t, \frac{N(nt)}{n} - \lambda t\right) \Rightarrow (\tilde{Y}_{m+1}(t), -\lambda\tilde{Y}_{m+1}(\lambda t)) \qquad \text{in } D_2 \text{ as } n \to \infty.$$

Then, by applying composition plus addition, Theorem 5.1 of Whitt (1980),

$$n^{-\frac{1}{2}}(A(N(nt)) - nt) \Rightarrow 0 \qquad \text{in } D_1 \text{ as } n \to \infty. \tag{6.2}$$

Combining (6.1) and (6.2), plus Theorem 4.1 of Billingsley, we find that

$$(Z_{ni}(t): 1 \leqq i \leqq m) \Rightarrow (\tilde{Y}_i(\lambda t) - p_i\lambda\tilde{Y}_{m+1}(\lambda t): 1 \leqq i \leqq m) \qquad \text{in } D_m \text{ as } n \to \infty.$$

Proof of Theorem 4. The idea is to use the theory of strong approximation as in Csörgö and Révesz (1981). Note that

$$\vec{N}(t) = \left(\sum_{n=1}^{N(t)} I(S_n = x_1), \cdots, \sum_{n=1}^{N(t)} I(S_n = x_m)\right).$$

We can combine Theorem 4 of Einmahl (1989) and Theorem B of Csörgö et al. (1987a) to obtain a multivariate analog of Theorem 1.1 (ii) of Csörgö et al. (1987b). (We use the same argument as given for the scalar case.)

This yields the inequality

$$P\left(\sup_{0 \leqq s \leqq nt} \|\vec{N}(ns) - n\vec{\lambda}s - \vec{B}(ns)\| > x\right) \leqq \beta(n)nx^{-p}$$

for an appropriately defined probability space, where $\beta(n) \to 0$ as $n \to \infty$. As a consequence, we have

$$P(|n^{-\frac{1}{2}}(Q_n(t) - nm_Q(t)) - \tilde{Q}_n(t)| > n^{-\frac{1}{2}}x) \leqq \beta(n)nx^{-p}$$

where $\tilde{Q}_n(t) = n^{-\frac{1}{2}}\sum_{i=1}^{m} [B_i(nt) - B_i(n(t - x_i))]$ has a distribution independent of n. Suppose that x_n is chosen so that $n^{-\frac{1}{2}}x_n = \beta(n)nx_n^{-p}$; then $n^{-\frac{1}{2}}x_n = o(n^{-(p-1)/2(p+1)})$. It follows from Theorem 1.2, p. 96, of Ethier and Kurtz (1986), that the Prohorov distance ε_n between the distributions of $n^{-\frac{1}{2}}(Q_n(t) - nm_Q(t))$ and $\tilde{Q}(t)$ is $o(n^{-(p-1)/2(p+1)})$. Hence, it follows, by definition of the Prohorov metric, that

$$P(n^{-\frac{1}{2}}(Q_n(t) - nm_Q(t)) \leqq x) \leqq P(\tilde{Q}(t) \leqq x + 2\varepsilon_n) + 2\varepsilon_n.$$

Since $\tilde{Q}(t)$ is Gaussian, it has a continuous density, so that $P(\tilde{Q}(t) \leqq x + 2\varepsilon_n) = P(\tilde{Q}(t) \leqq x) + O(\varepsilon_n)$. So,

$$P(n^{-\frac{1}{2}}(Q_n(t) - nm_Q(t)) \leqq x) \leqq P(\tilde{Q}(t) \leqq x) + o(n^{-(p-1)/2(p+1)}).$$

We can similarly show that

$$P(n^{-\frac{1}{2}}(Q_n(t) - nm_Q(t)) \geqq x) \leqq P(\tilde{Q}(t) \geqq x) + o(n^{-(p-1)/2(p+1)}),$$

proving the result for the queue-length process. Similar arguments work for the other two processes.

Proof of Theorem 5. We apply a strong approximation argument due to Csörgö et al. (1987b) to conclude that on an appropriately defined probability space,

$$\sup_{0 \leqq s \leqq t} |N(s) - \lambda s - \sigma\lambda^{\frac{3}{2}} B(s)| = O(\log t) \quad \text{a.s.},$$

where B is a standard Brownian motion; see (1.10) of Csörgö et al. (1987a). Then, after inserting the scaling by n,

$$\sup_{0 \leqq t \leqq t_n} [Q_n(t) - nm_Q(t)] = \sigma\lambda^{\frac{3}{2}} \sup_{0 \leqq t \leqq t_n} [B(nt) - B(n(t-x))] + O(\log n) + O(\log t_n) \quad \text{a.s.}$$

Hence,

$$\begin{aligned} &n^{-\frac{1}{2}} \sup_{0 \leqq t \leqq t_n} [Q_n(t) - nm_Q(t)] - \sigma\lambda^{\frac{3}{2}} \sup_{0 \leqq t \leqq t_n} [n^{-\frac{1}{2}}(B(nt) - B(n(t-x)))] \\ &\quad = \sqrt{\log t_n}\, O(\sqrt{\log t_n / n}) + o(1) = o(\sqrt{\log t_n}) \qquad \text{a.s. as } n \to \infty. \end{aligned} \tag{6.3}$$

(We use $t_n = O(\exp n^r)$ to get $O(\sqrt{\log t_n/n}) = o(1)$.) But

$$(2x \log (t_n/x))^{-\frac{1}{2}} \sup_{0 \leqq t \leqq t_n} [n^{-\frac{1}{2}} B(nt) - n^{-\frac{1}{2}} B(n(t-x))]$$

$$\overset{\mathcal{D}}{=} (2x \log (t_n/x))^{-\frac{1}{2}} \sup_{0 \leqq t \leqq t_n} [B(t) - B(t-x)] \to 1 \qquad \text{a.s. as } n \to \infty,$$

from (1.2.2) of Csörgő and Révész (1981) with $a_T = x$. ($\overset{\mathcal{D}}{=}$ denotes equality in distribution.) Combining this with (6.3) yields our result.

Proof of Proposition 4. First, the convergence $\theta_s(\Delta N_n) \Rightarrow \bar{N}_n$ as $s \to \infty$ is just the familiar convergence to the equilibrium renewal process associated with N_n. By the standard renewal argument, the interval until the first point in $\theta_s(\Delta N_n)$ converges weakly to the stationary forward recurrence time associated with the renewal process $\bar{N}_n$, which has interarrival times $n^{-1}A(1)$. The limiting stationary forward recurrence time is proper because $EA(1) < \infty$. Consequently, by Theorem 3.2 of Billingsley, there is convergence of the entire interarrival-time sequences in R^∞. Finally, since $A(1)$ has a continuous c.d.f., all interarrival times are strictly positive almost surely so that there is weak convergence of the associated counting processes, as in Section 2 of Whitt (1974). The other limits follow by applying more continuous mappings, as in Whitt (1974). The assumptions that N is renewal and $A(1)$ has a continuous c.d.f. guarantee that no two jumps (arrivals or departures) occur at the same time in the limit process (almost surely). As we have seen before, the assumption that S_k takes values in the finite set $\{x_1, \cdots, x_m\}$ enables us to easily express the processes explicitly. To illustrate, we treat the queue length process for $n=1$. (The other processes can be treated similarly. Setting $n \geqq 2$ just rescales the interarrival times.)

Note that

$$\begin{aligned}\theta_s(Q_1)(t) \equiv Q_1(s+t) &= \sum_{i=1}^{m} [N^i(t+s) - N^i(t+s-x_i)] \\ &= \sum_{i=1}^{m} \sum_{n=N(t+s-x_i)+1}^{N(t+s)} I(S_n = x_i) \\ &\overset{\mathscr{D}}{=} \sum_{i=1}^{m} \sum_{n=N(t+s-x_i)-N(s)+1}^{N(t+s)-N(s)} I(S_n = x_i),\end{aligned}$$

which implies that

$$\theta_s(Q_1)(t) \equiv Q_1(t+s) \Rightarrow \sum_{i=1}^{m} \sum_{n=\bar{N}(t+x_m-x_i)+1}^{\bar{N}(t+x_m)} I(S_n = x_i) \qquad \text{in } D_1 \text{ as } s \to \infty$$

where, without loss of generality, we assume that $x_1 < \cdots < x_m$. (We introduce the x_m term to ensure that the time argument of $\bar{N}$ is non-negative for all $t \geqq 0$. To justify the weak convergence, use the almost sure representation of the weak convergence of $\theta_s(\Delta N_n)$ provided by the Skorohod representation theorem. Also use the continuity of the c.d.f. of $A(1)$ to ensure that in the limit, almost surely, no arrivals and departures occur simultaneously.)

Finally, to see that the claimed stationarity holds for the limit process $[\bar{N}_n, \bar{D}_n, \bar{Q}_n, \bar{V}_n]$, note that the same limit holds for the original processes shifted by θ_{s+u} as $s \to \infty$, but these shifted processes converge to the shifted limit process $[\theta_u(\Delta\bar{N}), \theta_u(\Delta\bar{D}), \theta_u(\bar{Q}_n), \theta_u(\vec{B}_n)]$ as $s \to \infty$.

Proof of Theorem 6. Proposition 3(b) holds for both counting processes $\bar{N}$ and N by Donsker's theorem for the interarrival times, the assumed convergence of β_n, and Theorem 3.2 of Billingsley. (The first interarrival time of $\bar{N}_n$ is asymptotically negligible, so it does not alter the result by Theorem 4.1 of Billingsley.) Proposition 3(b) implies that (2.3) holds with $\gamma = \frac{1}{2}$ and $\bar{Z}$ Brownian motion, and it characterizes the limiting covariances. The limits for $\bar{Q}_n$, $\bar{D}_n$ and $\bar{V}_n$ then follow from Theorems 1–3, using the counting process $\bar{N}_n$ and restricting attention to the space $D([m^*, \infty), R)$. For example, note that

$$\bar{Q}_n(t) = \sum_{i=1}^{m} \sum_{j=\bar{N}(n(t+x_m-x_1)+1}^{\bar{N}(n(t+x_m))} I(S_j = x_i), \tag{6.4}$$

which is a valid representation for $Q_n(t)$ for $t \geqq m^*$ using the arrival process $\bar{N}$.

Proof of Theorem 7. From (6.4), we see that $\bar{Q}_n(t) \overset{\mathscr{D}}{=} \bar{N}(nx)$ where $\bar{N}$ is the equilibrium renewal process associated with N. Let $\bar{A}(m)$ be the arrival times associated with $\bar{N}$. Of course, $\{\bar{A}(m) - \bar{A}(m-1)\colon\ m \geqq 2\} \overset{\mathscr{D}}{=} \{A(m) - A(m-1)\colon$

$m \geqq 2\}$ and $\bar{A}(1)$ is independent of $\{\bar{A}(m) - \bar{A}(m-1)\colon m \geqq 2\}$. Moreover,

$$E\bar{A}(1)^3 = \lambda \int_0^\infty x^3 P(A(1) > x)\, dx = \frac{\lambda E A(1)^4}{4} < \infty.$$

Hence, we can apply the Berry–Esséen theorem for non-identically distributed summands on p. 544 of Feller (1971) to show that

$$\sup_y |P((\bar{A}(n) - \lambda^{-1}n)/n^{\frac{1}{2}} \leqq \sigma y) - P(\mathrm{N}(0,1) \leqq y)| = O(n^{-\frac{1}{2}}).$$

Then,

$$\begin{aligned} P(\bar{N}(nx) - n\lambda x > zn^{\frac{1}{2}}) &= P(\bar{A}(\lfloor n\lambda x + zn^{\frac{1}{2}} \rfloor) < nx) \\ &= P\left(\mathrm{N}(0,1) < \frac{nx - \lambda^{-1}(\lfloor n\lambda x + zn^{\frac{1}{2}} \rfloor)}{\sigma(\lfloor n\lambda x + zn^{\frac{1}{2}} \rfloor)^{\frac{1}{2}}}\right) + O(n^{-\frac{1}{2}}) \\ &= P(\mathrm{N}(0,1) < -\lambda^{-\frac{3}{2}} x^{-\frac{1}{2}} z \sigma^{-1} + O(n^{-\frac{1}{2}})) + O(n^{-\frac{1}{2}}) \\ &= P(\mathrm{N}(0,1) < -\lambda^{-\frac{3}{2}} x^{-\frac{1}{2}} z \sigma^{-1}) + O(n^{-\frac{1}{2}}), \end{aligned}$$

where we used the fact that the normal distribution has a bounded continuous density.

Acknowledgement

The first author performed some of the research for this paper while visiting the Department of Mathematics and Statistics at Carleton University, Ottawa, Canada. He wishes to thank the Department for its hospitality.

References

Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

Borovkov, A. A. (1967) On limit laws for service processes in multi-channel systems. *Siberian Math. J.* **8,** 746–763.

Borovkov, A. A. (1984) *Asymptotic Methods in Queuing Theory.* Wiley, New York.

Csörgő, M. and Révész, P. (1981) *Strong Approximations in Probability and Statistics.* Academic Press, New York.

Csörgő, M., Deheuvels, P. and Horváth, L. (1987a) An approximation of stopped sums with applications in queueing theory. *Adv. Appl. Prob.* **19,** 674–690.

Csörgő, M., Horváth, L. and Steinebach, J. (1987b) Invariance principles for renewal processes. *Ann. Prob.* **15,** 1441–1460.

Einmahl, U. (1989) Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *J. Multivariate Anal.* **28,** 20–68.

Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes: Characterization and Convergence.* Wiley, New York.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications,* Vol. II, 2nd edn. Wiley, New York.

Glynn, P. W. (1982) On the Markov property of the $GI/G/\infty$ Gaussian limit. *Adv. Appl. Prob.* **14,** 191–194.

GLYNN, P. W. and WHITT, W. (1988) Ordinary CLT and WLLN versions of $L = \lambda W$. *Math. Operat. Res.* **13,** 674–692.

HARRISON, J. M. and LEMOINE, A. J. (1981) A note on networks of infinite-server queues. *J. Appl. Prob.* **18,** 561–567.

IGLEHART, D. L. (1965) Limit diffusion approximations for the many server queue and the repairman problem. *J. Appl. Prob.* **2,** 429–441.

IGLEHART, D. L. and WHITT, W. (1971) The equivalence of functional central limit theorems for counting processes and associated partial sums. *Ann. Math. Statist.* **42,** 1372–1378.

NEWELL, G. F. (1973) *Approximate Stochastic Behavior of n-Server Service Systems with Large n.* Lecture Notes in Economics and Mathematical Systems **87,** Springer-Verlag, Berlin.

POMAREDE, J.-M. L. (1976) A Unified Approach Via Graphs to Skorohod's Topologies on the Function Space D. Ph.D. dissertation, Department of Statistics, Yale University.

REIMAN, M. I. (1984) Open queueing networks in heavy traffic. *Math. Operat. Res.* **9,** 441–458.

ROSS, S. M. (1970) *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

SKOROHOD, A. V. (1956) Limit theorems for stochastic processes. *Theor. Probability Appl.* **1,** 261–290.

VERVAAT, W. (1972) Functional central limit theorems for processes with positive drift and their inverses. *Z. Wahrscheinlichkeitsth.* **23,** 245–253.

WHITT, W. (1974) The continuity of queues. *Adv. Appl. Prob.* **6,** 175–183.

WHITT, W. (1980) Some useful functions for functional limit theorems. *Math. Operat Res.* **5,** 67–85.

WHITT, W. (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Prob.* **14,** 171–190.

WHITT, W. (1984) Departures from a queue with many busy servers. *Math. Operat. Res.* **9,** 534–544.