HOW MULTISERVER QUEUES SCALE WITH GROWING CONGESTION-DEPENDENT DEMAND

by

Ward Whitt

AT&T Labs – Research
Room A117, Shannon Laboratory
180 Park Avenue
Florham Park, NJ 07932-00971
email: wow@research.att.com

October 12, 2001

Abstract

We investigate how performance scales in the standard $M/M/n$ queue in the presence of growing congestion-dependent customer demand. We scale the queue by increasing the number of servers. We let the arrival rate depend on the steady-state congestion, considering several alternative congestion measures. We assume that the arrival rate is equal to the maximum possible output rate (the individual service rate times the number of servers) multiplied by a decreasing function of the congestion measure. Under minor regularity conditions, there is a unique equilibrium arrival rate for each n and, as n increases, the queue is brought into heavy traffic. The three different heavy-traffic regimes for multiserver queues identified by Halfin and Whitt (1981) each can arise depending on the congestion measure used. In considerable generality, there is asymptotic service efficency – the server utilization approaches one – even if there is significant uncertainty about the arrival rate.

Keywords: queues, state-dependent queues, multiserver queues, heavy traffic, equilibrium, congestion-dependent demand, asymptotic service efficiency, uncertainty about demand in queues, economics of queues

We are interested in the way performance scales in service systems that grow in response to increasing customer demand. There are several ways that service systems can grow. The service system could grow as a collection of an increasing number of separate service facilities of approximately fixed size, e.g., as with a chain of gas stations. Here we shall consider a single sevice facility modeled as a queue with either more servers (homogeneous servers working in parallel) or a faster service rate. The case of an increasing number of servers is natural for a telephone call (or more general contact) center or a modem bank maintained at a given location by an Internet service provider. The case of increasing service rate is natural in communication services with increasing bandwidth.

The growth of a service system depends upon many factors, including the nature of the service being provided, the price of the service, customer preferences and competition from alternative service providers. We will focus on the influence of congestion (thinking of the price as fixed) and the structure of the service system as captured by a basic queueing model. We will not consider customer preferences or competition directly. We will not consider specific services; we seek general principles.

We consider the classical $M/M/n$ queue with n servers, unlimited waiting space and the first-come first-served service discipline, without customer abandonment, where the arrival rate is allowed to be a function of the steady-state congestion. We primarily consider the asymptotic behavior as the number of servers is allowed to increase. For fixed arrival rate, the congestion decreases as n increases. Hence, the congestion-dependent demand causes the arrival rate to increase as n increases. Under regularity conditions, the congestion-dependence brings the queue into heavy traffic as $n \to \infty$. This phenomenon occurs because of the economy of scale associated with a larger system with more servers; e.g., see Smith and Whitt (1981) and Whitt (1992). However, as shown by Halfin and Whitt (1981), there are actually three different heavy-traffic regimes for the $M/M/n$ queue as $n \to \infty$. We show that each of these three heavy-traffic regimes can result, depending on the way in which "congestion" is defined.

The three heavy-traffic regimes can be characterized by the asymptotic behavior of the delay probability (the probability of having to wait before beginning service) as the arrival rate and number of servers both approach infinity. In the standard heavy-traffic regime usually associated with a fixed number of servers, the delay probability approaches 1. In the standard heavy-traffic regime, properly scaled versions of the queue-length and waiting-time processes converge to reflected Brownian motion (RBM), as we show in Theorem 2.2 below. In the infinite-server heavy-traffic regime, the delay probability approaches 0; this regime is called heavy traffic too because the number of customers in the system approaches infinity along with the arrival rate and the number of servers, even though the server utilization (the proportion of time each server is busy) approaches a limit strictly between 0 and 1.

In the intermediate heavy-traffic regime identified by Halfin and Whitt, the delay probability approaches a limit α with $0 < \alpha < 1$. As indicated by the delay probability asymptotics, the intermediate regime is often the most realistic for multiserver queues. We provide additional support for the intermediate heavy-traffic regime here, but our analysis shows that it is not automatic.

The standard and intermediate heavy-traffic regimes imply *asymptotic service efficiency*: As $n \to \infty$ the server utilization approaches one. The efficiency principle for multiserver queues implies that we can satisfy given exogenous demand very efficiently in a multiserver system by choosing an appropriate number of servers, n, when the arrival rate is very high, provided that we actually know the arrival rate. However, if there is significant uncertainty about the arrival rate, then it may be necessary to have extra servers to hedge against that uncertainty. The need to hedge against uncertainty is especially strong if the service facility in question is the sole service provider, and there is determination to provide very good quality of service. Indeed, in many large multiserver service systems it is the uncertainty about the arrival rate that holds back efficiency. For example, with large n such as $n \geq 400$, the target server utilization may be set at about 92% when the queueing performance measures would directly indicate 98% or more, because of uncertainty about the arrival rate. (At very high efficiency, the need for server idle time may also play a role, even with non-human servers.)

However, in many cases the arrival rate is not actually determined exogenously. If the arrival rate can be regarded as being a function of the congestion as we have postulated here, then our analysis shows that the service system can be asymptotically efficient as the number of servers increases even when there is significant uncertainty about demand, provided that the initial "congestion-free" demand exceeds supply.

This paper was initially motivated by work on customer contact centers by Armony and Maglaras (2001). For related work on contact (or telephone call) systems, see Whitt (1999), Garnett, Mandelbaum and Reiman (1999), and references therein.

More generally, this investigation is intended to contribute to a better understanding of the economics of queues, multiserver queues, state-dependent queues and heavy-traffic theory for queues. For background on these topics, see Mendelson and Whang (1990), Mandelbaum and Pats (1995), Hassin and Haviv (1997), Mandelbaum and Shimkin (2000), Whitt (1990, 1992,

2001) and references therein.

1. The Model with an Increasing Number of Servers

We consider the $M/M/n$ queue, first letting n increase with a fixed service rate μ. Without loss of generality, let the individual service rate be 1. Let the arrival rate be λ_n and the traffic intensity be ρ_n, with the subscript n indicating the number of servers. Since the service rate has been set at 1, $\rho_n = \lambda_n/n$. The traffic intensity coincides with the utilization of each server.

Let π_n be the steady-state delay probability, w_n the mean steady-state waiting time (before beginning service), π_n^x the probability that the steady-state waiting time exceeds x, and τ_n the mean steady-state response (or sojourn) time – the mean service time plus the mean waiting time. As a function of n and the traffic intensity ρ, the (steady-state) delay probability is

$$\pi_n \equiv \pi_n(\rho) = [(n\rho)^n/n!(1-\rho)]\nu_n(\rho) \tag{1.1}$$

with

$$\nu_n(\rho) \equiv [(n\rho)^n/n!(1-\rho)) + \sum_{k=0}^{n-1}(n\rho)^k/k!]^{-1} \ . \tag{1.2}$$

In terms of the delay probability, the other congestion measures can be expressed as

$$\begin{aligned} w_n \equiv w_n(\rho) &= \frac{\pi_n}{n(1-\rho)}, \\ \pi_n^x \equiv \pi_n^x(\rho) &= \pi_n \exp\{-n(1-\rho)x\}, \quad x \geq 0, \\ \tau_n \equiv \tau_n(\rho) &= 1 + w_n \ . \end{aligned} \tag{1.3}$$

In order to obtain congestion-dependent demand, we regard the arrival rate as a function of a real-valued steady-state congestion measure. We will consider seven different cases for the demand function. The first five have the general form

$$\lambda_n \equiv \lambda_n(\xi) \equiv f(\xi)n \ , \tag{1.4}$$

where ξ is the steady-state congestion measure and f is a strictly-decreasing continuous real-valued function of a real variable with

$$f(0) = \gamma > 1 \quad \text{and} \quad \lim_{x \to \xi_{lim}} f(x) = 0 \ , \tag{1.5}$$

where ξ_{lim} is the upper limit for the congestion measure ξ. The assumption (1.5) ensures that demand would exceed supply if there were no congestion and that demand dissipates completely as congestion approaches its upper limit. It will be elementary that, for each n, there is an equilibrium arrival rate λ_n^* producing an equilibrium level of congestion ξ_n^* such that $\lambda_n^* = \lambda_n(\xi_n^*)$. We will be interested in the way the equilibrium performance behaves as $n \to \infty$.

The specific five steady-state congestion measures that will play the role of ξ above are: (1) the mean steady-state waiting time, w, (2) the steady-state waiting-time tail probability, π^x for $x > 0$, (3) the steady-state delay probability, π, (4) the (per-server) server utilization, ρ, and (5) the mean steady-state response time, τ.

These one-dimensional summary statistics are plausible characterizations of the entire steady-state performance. Since the steady-state waiting-time distribution is exponential with an atom (positive probability) at the origin, it is fully described by two parameters. Similarly, the steady state queue length (not counting customers in service) has a geometric distribution with a modified mass at the origin, so it too depends on only two parameters.

In equation (1.4) the arrival rate is made proportional to the number of servers, n. We should explain why we have the n. We assume that the number of servers is being increased in response to a growing customer base. Thus we are thinking of a growing system with the number of potential customers being proportional to n. With that framework, it seems natural to regard the potential arrival rate with n servers, given no congestion, as being γn for $\gamma > 1$. Then individual customer preferences within the customer population of size proportional to n makes it natural for the demand rate to assume the form in (1.4).

The sixth case is designed to reflect increasing customer expectations in the presence of a growing service system. Specifically, we assume that customers learn to expect a better quality of service as n increases. In the intermediate heavy-traffic regime of Halfin and Whitt the mean waiting time is of order $1/\sqrt{n}$ as n increases. Thus, in the sixth case, we let

$$\lambda_n \equiv \lambda_n(w) \equiv f(\sqrt{n}w)n \ . \tag{1.6}$$

Motivated by Armony and Maglaras (2001), the final seventh case has the potential demand differing from the capacity n by order $\sqrt{n}$. Specifically, we assume that

$$\lambda_n \equiv \lambda_n(w) \equiv f(w)(n - \delta_n\sqrt{n}) \ , \tag{1.7}$$

where

$$\delta_n \to \delta \ , \tag{1.8}$$

as $n \to \infty$, δ is a constant and f is a strictly-decreasing function of w with $f(0) = 1$ and $f(w) \to 0$ as $w \to \infty$.

In all seven cases above the function f is independent of n. A more general model would have functions f_n depending upon n. However, the more general model is essentially equivalent to the model above if we assume that $f_n \to f$ as $n \to \infty$, with the convergence being uniform over bounded intervals, and that f and f_n for $n \geq 1$ all have the properties assumed for the single function f above. Hence, we consider only a single function f.

In the sections below we analyze the seven cases of congestion-dependent demand specified above. Afterwards we briefly consider the case of the $M/M/n$ model with increasing service rate μ (and fixed n) and make concluding remarks. We place all proofs in the final section.

2. Function of the Mean Waiting Time

In this section we let the congestion-dependent arrival rate be as in (1.4) with $\xi = w$, the mean-steady-state waiting time. We first convert the congestion-dependent arrival rate into an associated congestion-dependent traffic intensity: Equation (1.4) is obviously equivalent to

$$\rho_n(w) = f(w) \ . \tag{2.1}$$

On the other hand, we have the formula for the mean steady-state waiting time as a function of the traffic intensity ρ in model n (with n servers) given in (1.3).

Given the two equations, we seek an equilibrium: we seek values w_n^* and ρ_n^* such that both equations are satisfied simultaneously. We then want to describe the asymptotic behavior as $n \to \infty$. These results are contained in the following theorem, whose proof draws heavily upon Proposition 1 of Halfin and Whitt. Let f^{-1} be the inverse of the strictly-decreasing function f.

Theorem 2.1. *Suppose that* (2.1) *holds. For each* n*, there exist unique numbers* ρ_n^* *and* w_n^* *such that* $0 < \rho_n^* < 1$, $0 < f^{-1}(1) < w_n^* < \infty$,

$$w_n(\rho_n^*) = w_n^* \tag{2.2}$$

and

$$\rho_n(w_n^*) = \rho_n^* \ . \tag{2.3}$$

Moreover, $\rho_{n+1}^* > \rho_n^*$ *and* $w_{n+1}^* < w_n^*$ *for all* n*. As* $n \to \infty$,

$$\begin{aligned} \rho_n^* &\to 1 \ , \\ n(1-\rho_n^*) &\to \frac{1}{f^{-1}(1)} \ , \\ \pi_n(\rho_n^*) &\to 1 \ , \\ w_n^* &\to f^{-1}(1) \ . \end{aligned} \tag{2.4}$$

The case of heavy traffic produced by Theorem 2.1 has the equilibrium traffic intensities ρ_n^* approach the limit 1 so rapidly that the heavy-traffic behavior is essentially the same as for the M/M/n queue with fixed n, which in turn is the same as for the $M/M/1$ queue. To state the results, let $Q_n(t)$ be the queue length (not counting the customers in service) at time t and let $W_{n,k}$ be the waiting time until beginning service for the k^{th} customer to arrive after time 0 for the n^{th} model with the equilibrium arrival rate determined by Theorem 2.1. For simplicity, assume that the n^{th} system starts with n customers present (all in service) at time 0. (Other initial conditions can easily be treated.) Let $Q_n(\infty)$ and $W_{n,\infty}$ be random variables with the steady-state (limiting and stationary) distributions for these processes.

Let $\{R(t;m,\sigma^2) : t \geq 0\}$ be reflected Brownian motion (RBM) starting at the origin with drift m and diffusion coefficient σ^2. Let $R(\infty;m,\sigma^2)$ be a random variable with the steady-state (limiting and stationary) distribution of RBM, which has an exponential distribution with mean $\sigma^2/2m$. Let $\Rightarrow$ denote convergence in distribution with the context as indicated. For convergence of stochastic processes, we can use the function space $D \equiv D([0,\infty),\mathbb{R})$ with the usual Skorohod J_1 topology; see Billingsley (1999) or Whitt (2001). Since the limit process has continuous sample paths, the mode of convergence on D is equivalent to uniform convergence over bounded intervals.

Theorem 2.2. *Suppose that* (2.1) *holds for the $M/M/n$ queue initially with n customers in the system. For each n, let the arrival rate be the equilibrium arrival rate $\lambda_n^* = n\rho_n^*$ whose existence is established in Theorem* 2.1. *Let $\zeta = 1/f^{-1}(1)$. Then, as $n \to \infty$,*

$$\{\zeta n^{-1} Q_n(\zeta^{-2} nt) : t \geq 0\} \Rightarrow \{R(t;-1,2) : t \geq 0\} \tag{2.5}$$

and

$$\{\zeta W_{n,\lfloor \zeta^{-2} n^2 t \rfloor} : t \geq 0\} \Rightarrow \{R(t;-1,2) : t \geq 0\} \tag{2.6}$$

in D, and

$$\zeta n^{-1} Q_n(\infty) \Rightarrow R(\infty;-1,2) \tag{2.7}$$

and

$$\zeta W_{n,\infty} \Rightarrow R(\infty;-1,2) \tag{2.8}$$

*in $\mathbb{R}$, where $R(\infty;-1,2)$ is a mean-*1 *exponential random variable.*

For background on heavy-traffic limits such as Theorem 2.2, see Chapters 5, 9 and 10 in Whitt (2001). Note that (2.8) (with appropriate uniform integrability [p. 31 of Billingsley], which can be established here) implies that $\pi_n^* \to 1$ and $w_n^* \to f^{-1}(1)$, as concluded in (2.4). In turn, since the distribution of $W_{n,\infty}$ is exponential plus an atom at the origin, the limits for $\pi_n(\rho_n^*)$ and $w_n(\rho_n^*)$ in (2.4) directly imply the limit in (2.8).

Remark 2.1. *The waiting-time tail probability.* By the same reasoning, essentially the same heavy-traffic regime is obtained if the congestion measure is the steady-state waiting-time tail probability π^x. If $\xi = \pi^x$, the demand function f in (1.4) maps the interval $[0,1]$ onto the interval $[0,\gamma]$. We obtain analogs of Theorems 2.1 and 2.2, again with $\rho_{n+1}^* > \rho_n^*$ for all n and $\rho_n^* \to 1$ and $\pi_n(\rho_n^*) \to 1$ as $n \to \infty$, but now with $\pi_{n+1}^{x*} < \pi_n^{x*}$ for all n and

$$\begin{aligned} \pi_n^{x*} &\to f^{-1}(1), \\ n(1-\rho_n^*) &\to -x \log f^{-1}(1), \\ w_n(\rho_n^*) &\to 1/(-x \log f^{-1}(1)) \ . \end{aligned} \tag{2.9}$$

3. Function of the Delay Probability

In this section we let the congestion-dependent arrival rate be as in (1.4) with $\xi = \pi$, the steady-state delay probability. As in the previous section, we work with the traffic intensity instead of the arrival rate. Thus, instead of (2.1), here we have

$$\rho_n(\pi) = f(\pi) \ . \tag{3.1}$$

On the other hand, the steady-state delay probability with n servers is a function of the traffic intensity, $\pi_n(\rho)$ as given in (1.1).

To state the analog of Theorem 2.1, let $\alpha \equiv \alpha(\beta)$ be the (strictly-decreasing continuous) function from (2.3) of Halfin and Whitt describing the nondegenerate limit for the delay probability in the intermediate heavy-traffic regime,

$$\alpha(\beta) \equiv [1 + \sqrt{2\pi}\beta\Phi(\beta)\exp(\beta^2/2]^{-1} \ , \tag{3.2}$$

where Φ is the standard normal cdf, i.e., $\Phi(x) \equiv P(N(0,1) \le x)$. Halfin and Whitt show that $\sqrt{n}(1-\rho_n) \to \beta$ with $0 < \beta < \infty$ if and only if $\pi_n(\rho_n) \to \alpha$ as $n \to \infty$ with $0 < \alpha < 1$ for α in (3.2).

Theorem 3.1. *Suppose that* (3.1) *holds. For each* n*, there exist unique numbers* ρ_n^* *and* π_n^* *such that* $0 < \rho_n^* < 1$, $0 < \pi_n^* < 1$,

$$\pi_n(\rho_n^*) = \pi_n^* \tag{3.3}$$

and

$$\rho_n(\pi_n^*) = \rho_n^* . \tag{3.4}$$

Moreover, $\rho_{n+1}^* > \rho_n^*$ *and* $\pi_{n+1}^* < \pi_n^*$ *for all* n. *As* $n \to \infty$,

$$\begin{aligned} \rho_n^* &\to 1 , \\ \sqrt{n}(1-\rho_n^*) &\to \beta \equiv \alpha^{-1}(f^{-1}(1)) , \\ \pi_n^* &\to f^{-1}(1) = \alpha(\beta) , \\ \sqrt{n} w_n^* &\to \alpha(\beta)/\beta . \end{aligned} \tag{3.5}$$

Since this case produces the intermediate many-server heavy-traffic regime considered by Halfin and Whitt, we can apply that paper to obtain all desired associated heavy-traffic limits. The congestion-dependent demand serves to determine the one asymptotic parameter β from the functions f and α via $\beta = \alpha^{-1}(f^{-1}(1))$ in (3.5).

4. Function of the Server Utilization

In this section we let the congestion-dependent arrival rate be as in (1.4) with $\xi = \rho$, the (per-server) server utilization. When we convert the congestion-dependent arrival rate into the congestion-dependent traffic intensity, we obtain the equation

$$\rho_n(\rho) = f(\rho) . \tag{4.1}$$

On the other hand, trivially, $\rho_n(\rho) = \rho$. Thus we have the following elementary result.

Theorem 4.1. *Suppose that* (4.1) *holds. There exists a unique number* ρ^* *with* $0 < \rho^* < 1$ *such that*

$$f(\rho^*) = \rho^* . \tag{4.2}$$

For all n,

$$\rho_n(\rho^*) = \rho^* . \tag{4.3}$$

As $n \to \infty$,

$$\begin{aligned} \pi_n(\rho_n^*) &\downarrow 0 \\ w_n(\rho_n^*) &\downarrow 0 . \end{aligned} \tag{4.4}$$

This third case is the infinite-server heavy-traffic regime in which the arrival rate is kept directly proportional to the number of servers, n, as n increases. This heavy-traffic regime was first considered by Iglehart (1965); see also Glynn and Whitt (1991) and Chapter 10 of Whitt (2001).

5. Function of the Mean Response Time

In this section we let the congestion-dependent arrival rate be as in (1.4) with $\xi = \tau$, the mean steady-state response time. When we convert the congestion-dependent arrival rate into the congestion-dependent traffic intensity, we obtain the equation

$$\rho_n(\tau) = f(\tau) \ . \tag{5.1}$$

The second equation involving $\tau_n(\rho)$ is given in (1.3).

Since $\tau_n(\rho) \geq 1$ for all n and ρ, the performance depends critically on $f(1)$. It turns out that we obtain all three previous cases depending on whether $f(1)$ is less than, equal to, or greater than 1. To treat the case in which $f(1) = 1$, we assume that f has a continuous derivative $f'(1)$ in the neighborhood of 1. To state the result, let h be the (strictly decreasing) function

$$g(\beta) \equiv \alpha(\beta)/\beta^2 \tag{5.2}$$

for α in (3.2).

Theorem 5.1. *Suppose that* (5.1) *holds. For each n, there exist unique numbers ρ_n^* and τ_n^* such that* $0 < \rho_n^* < 1$, $\max\{1, f^{-1}(1)\} < \tau_n^* < \infty$,

$$\tau_n(\rho_n^*) = \tau_n^* \tag{5.3}$$

and

$$\rho_n(\tau_n^*) = \rho_n^* \ . \tag{5.4}$$

Moreover, $\rho_{n+1}^* > \rho_n^*$ and $\tau_{n+1}^* < \tau_n^*$ for all n. There are three cases to describe the behavior as $n \to \infty$:

(a) If $f(1) > 1$, then

$$\begin{aligned} \rho_n^* &\to 1, \\ \tau_n^* &\to f^{-1}(1), \\ \pi_n(\rho_n^*) &\to 1, \\ w_n(\rho_n^*) &\to f^{-1}(1) - 1 \\ n(1-\rho_n^*) &\to 1/(f^{-1}(1) - 1); \end{aligned} \tag{5.5}$$

(b) If $f(1) = 1$ and f has a continuous derivative in the neighborhood of 1, then

$$\rho_n^* \to 1,$$

$$\begin{aligned}\sqrt{n}(1-\rho_n^*) &\rightarrow \beta \equiv g^{-1}(-f'(1)),\\ \tau_n^* &\rightarrow 1,\\ w_n(\rho_n^*) &\rightarrow 0\\ \pi_n(\rho_n^*) &\rightarrow \alpha(\beta);\end{aligned} \tag{5.6}$$

for g in (5.2).

(c) *If $f(1) < 1$, then*

$$\begin{aligned}\rho_n^* \rightarrow f(1) < 1,\\ \tau_n^* \rightarrow 1,\\ \pi_n(\rho_n^*) \rightarrow 0,\\ w_n(\rho_n^*) \rightarrow 0\ .\end{aligned} \tag{5.7}$$

6. Function of the Scaled Waiting Time

We can obtain different limits by letting the argument of the function f in (1.4) depend upon n. We illustrate in this section by obtaining the intermediate Halfin-Whitt many-server heavy-traffic regime when the congestion measure in the congestion-dependent demand function is the appropriately scaled mean steady-state waiting time, in particular, when (1.6) holds.

For $\alpha \equiv \alpha(\beta)$ in definition (3.2), let $h \equiv h(\beta)$ be the (strictly decreasing) function

$$h(\beta) \equiv \alpha(\beta)/\beta\ . \tag{6.1}$$

Theorem 6.1. *Suppose that* (1.6) *holds. For each n, there exist unique numbers ρ_n^* and w_n^* such that $0 < \rho_n^* < 1$, $0 < f^{-1}(1) < w_n^* < \infty$,*

$$w_n(\rho_n^*) = w_n^* \tag{6.2}$$

and

$$\rho_n(w_n^*) \equiv f(\sqrt{n} w_n^*) = \rho_n^*\ . \tag{6.3}$$

Moreover, as $n \rightarrow \infty$,

$$\begin{aligned}\rho_n^* &\rightarrow 1\ ,\\ \sqrt{n}(1-\rho_n^*) &\rightarrow \beta \equiv h^{-1}(f^{-1}(1))\ ,\\ \pi_n(\rho_n^*) &\rightarrow \alpha(\beta)\ ,\\ \sqrt{n} w_n^* &\rightarrow f^{-1}(1) = \alpha(\beta)/\beta\ ,\end{aligned} \tag{6.4}$$

where h and α are the functions in (6.1) *and* (3.2).

7. Closely Matching Potential Demand

We also can obtain the intermediate Halfin-Whitt heavy-traffic regime when the congestion measure is the mean steady-state waiting time if we assume that the potential demand for model n differs from n by order $\sqrt{n}$ for all n sufficiently large. This parallels the assumption made by Armony and Maglaras (2001) for their two-class model. However, it may be less realistic to assume that potential demand matches capacity so closely.

Specifically, Instead of (1.4) with $\xi = w$, we now assume that (1.7) and (1.8) hold. Then, instead of (2.1), we obtain

$$\rho_n(w) = f(w)(1 - \delta_n/\sqrt{n}) \tag{7.1}$$

for the same f and δ_n.

Closely paralleling Theorem 5.1 (b), we obtain an asymptotic result under a smoothness assumption by exploiting a Taylor series expansion of f about 0.

Theorem 7.1. *Suppose that* (7.1) *and* (1.8) *hold. For each* n*, there exist unique numbers* ρ_n^* *and* w_n^* *such that* $0 < \rho_n^* < \min\{1, 1 - \delta_n/\sqrt{n}\}$, $0 < w_n^* < \infty$,

$$w_n(\rho_n^*) = w_n^* \tag{7.2}$$

and

$$\rho_n(w_n^*) = \rho_n^* \ . \tag{7.3}$$

If in addition f *has a continuous derivative in the neighborhood of* 0*, then*

$$\begin{aligned} \rho_n^* &\to 1 \ , \\ \sqrt{n}(1 - \rho_n^*) &\to \beta^* \ , \\ \pi_n(\rho_n^*) &\to \alpha(\beta^*) \ , \\ \sqrt{n} w_n^* &\to \alpha(\beta^*)/\beta^* \ , \end{aligned} \tag{7.4}$$

as $n \to \infty$*, where* α *is as in* (3.2) *and* β^* *is the unique (necessarily positive) solution to the equation*

$$\beta^* - \delta = -f'(0)\alpha(\beta^*)/\beta^* \ . \tag{7.5}$$

8. Increasing Service Rate

In this section we suppose that the capacity of the $M/M/n$ queue increases by increasing the service rate instead of by increasing the number of servers, n. Hence, here let the service rate be μ. The traffic intensity becomes $\rho = \lambda/n\mu$.

We now think of the arrival rate growing with μ. With a fixed number of servers, it is natural to evaluate the waiting time relative to the service time. Thus we propose the demand function

$$\lambda \equiv \lambda(w, \mu) = f(w\mu)n\mu \; , \tag{8.1}$$

where f has the same structure as before and w is again the mean steady-state waiting time. Note that if μ is increased solely by changing the time units, then $w\mu$ should remain unchanged and λ should increase directly proportional to μ, just as in (8.1). That observation provides additional motivation for (8.1). We will consider alternative expressions for the demand function in the next section.

In terms of the traffic intensity ρ, (8.1) can be expressed equivalently as

$$\rho \equiv \rho(w, \mu) = f(w\mu) \; . \tag{8.2}$$

On the other hand, w can be expressed as

$$w \equiv w(\rho, \mu) = \frac{\pi_n(\rho)}{n\mu(1-\rho)} \; . \tag{8.3}$$

For simplicity, let $x \equiv w\mu$. From (8.2) and (8.3), we obtain the two equations

$$\rho(x) = f(x) \tag{8.4}$$

and

$$x(\rho) = \frac{\pi_n(\rho)}{n(1-\rho)} \; , \tag{8.5}$$

with μ suppressed in the notation. The following theorem is immediate.

Theorem 8.1. *Consider the $M/M/n$ queue with individual service rate μ. Suppose that* (8.1) *holds. Then, for all $\mu > 0$, there exist unique numbers x^* and ρ^* with $0 < \rho^* < 1$ and $0 < x^* < \infty$ such that*

$$\begin{aligned} \rho(x^*) &= \rho^*, \\ x(\rho^*) &= x^*, \\ w^* \equiv w(\rho^*, \mu) &= x^*/\mu, \\ \tau^* \equiv \tau(\rho^*, \mu) &= (1 + x^*)/\mu \; . \end{aligned} \tag{8.6}$$

9. Changing Expectations with Increasing Service Rate

If the service rate does increase substantially, it is evident that customer expectations about their quality of service might not change correspondingly. Customer expectations for quality of service might grow more slowly or more quickly. To illustrate these possibilities, we now assume that, instead of (8.1), the arrival rate is given by

$$\lambda \equiv \lambda(w, \mu) = f(w\mu^{\delta})n\mu \tag{9.1}$$

for $0 < \delta < \infty$ with $\delta \neq 1$. As before, we have (8.3).

Now let $x = w\mu^{\delta}$. From (9.1) and (8.3), we obtain the two equations

$$\rho(x) \equiv \rho(x, \mu) = f(x) \tag{9.2}$$

and

$$x(\rho) \equiv x(\rho, \mu) = \frac{\mu^{\delta-1}\pi_n(\rho)}{n(1-\rho)} \ . \tag{9.3}$$

The following result shows that the performance scales very differently from the scaling in Theorem 8.1 when $\delta \neq 1$.

Theorem 9.1. *Consider the $M/M/n$ queue with individual service rate μ. Suppose that (9.1) holds. Then, for all $\mu > 0$, there exist unique numbers $\rho^*(\mu)$, $x^*(\mu)$ and $w^*(\mu)$ with $0 < \rho^*(\mu) < 1$, $0 < x^*(\mu) < \infty$ and $0 < w^*(\mu) < \infty$ such that*

$$\begin{aligned} \rho(x^*, \mu) &= \rho^*(\mu), \\ x(\rho^*, \mu) &= x^*(\mu), \\ w(\rho^*, \mu) &= w^*(\mu) = \mu^{-\delta}x^*(\mu) \ . \end{aligned} \tag{9.4}$$

There are two cases to describe the behavior as $\mu \to \infty$.

(a) If $\delta < 1$, then

$$\begin{aligned} \rho^*(\mu) &\to 1, \\ x^*(\mu) &\to f^{-1}(1), \\ w^*(\mu) &\to 0, \\ \mu w^*(\mu) &\to \infty \ . \end{aligned} \tag{9.5}$$

(b) If $\delta > 1$, then

$$\rho^*(\mu) \to 0,$$

$$\begin{aligned} x^*(\mu) &\to \infty, \\ w^*(\mu) &\to 0, \\ \mu w^*(\mu) &\to 0 \ . \end{aligned} \tag{9.6}$$

Note that there is asymptotic service efficiency (here meaning that $\rho^*(\mu) \to 1$) if $0 < \delta < 1$, but not otherwise.

10. Concluding Remarks

The analysis in Sections 2–7 supports the established notion of service efficiency of multi-server queueing systems as the number of servers increases, as discussed in Whitt (1992), with the exception of Section 4 and Theorem 5.1 (c), but these exceptions are understandable. The server utilization used as the congestion measure in Section 4 does not actually represent congestion as experienced by the arriving customers. The condition in Theorem 5.1 (c) artificially constrains the traffic intensity, keeping it bounded away from 1.

More importantly, the analysis here *extends* the notion of asymptotic service efficiency of multiserver queues to the commonly arising situation in which there is significant uncertainty about the arrival rate. Even in the presence of such uncertainty, asymptotic service efficiency persists, provided that there is appropriate congestion-dependent demand, as postulated here. With appropriate congestion-dependent demand, it is not necessary to hedge against the uncertain demand by adding many extra servers.

Our results in this paper have been restricted to the $M/M/n$ queue. From Halfin and Whitt, it is clear that the results extend to $GI/M/n$ queues. Even though only part of Proposition 1 for the $M/M/n$ queue in Halfin and Whitt (the implication in only one direction) is established for $GI/M/n$ queues in Theorem 4 there, it is not difficult to obtain the full analog of Proposition 1.

It is evident that the results here also extend to more general $GI/PH/n$ queues with phase-type service-time distributions, due to results established by Puhalskii and Reiman (2000), but there remain some technical details to provide a complete demonstration. The conjectured result is: $\sqrt{n}(1-\rho_n^*) \to \beta$ as $n \to \infty$ for some β with $0 < \beta < \infty$ if and only if $\pi_n(\rho_n^*) \to \alpha$ as $n \to \infty$ for some α with $0 < \alpha < \infty$. A remaining challenge when the PH service-time distribution is not M is to determine the functional form of $\alpha(\beta)$.

Heuristic approximations for the delay probability π_n in $GI/G/n$ queues are proposed and investigated in Section 3 of Whitt (1993). By letting $n = m\beta^2$ and $\rho_n = 1 - 1/\sqrt{m}$ for some m

(e.g., $m = 100$), these heuristic approximations can be regarded as approximations for $\alpha(\beta)$. The steady-state distribution of the limiting diffusion process in Puhalsii and Reiman (2000) evidently produces the exact value for $GI/PH/n$ systems, though.

11. Proofs

Proof of Theorem 2.1. As a consequence of the assumptions, the composite function $\rho_n(w_n(\rho))$ is a non-increasing continuous function mapping the closed interval $[0, \gamma]$ into itself, assuming the values γ and 0 at the left and right endpoints. Moreover, this function assumes the value 0 throughout the interval $[1, \gamma]$ and it is strictly decreasing on the interval $[0, 1]$. Thus, there exists a unique ρ_n^* with $0 < \rho_n^* < 1$ satisfying

$$\rho_n(w_n(\rho_n^*)) = \rho_n^* \ . \tag{11.1}$$

Similarly, the composite function $w_n(\rho_n(w))$ is a non-increasing function mapping the extended interval $[0, \infty] \equiv [0, \infty) \cup \{\infty\}$ into itself. It assumes the value ∞ on the interval $[0, f^{-1}(1)]$; it is strictly decreasing and continuous on the interval $(f^{-1}(1), \infty)$ with right limit ∞ at the left endpoint and limit 0 at ∞. Thus there exists a unique number w_n^* with $0 < f^{-1}(1) < w_n^* < \infty$ such that

$$w_n(\rho_n(w_n^*)) = w_n^* \ . \tag{11.2}$$

Finally, it is easy to see that equations (2.2) and (2.3) hold. To see that, suppose that they fail and deduce a contradiction. For example, if $w_n(\rho_n^*) < w_n^*$, then

$$\rho_n^* = \rho_n(w_n(\rho_n^*)) > \rho_n(w_n^*) \tag{11.3}$$

and

$$w_n(\rho_n^*) > w_n(\rho_n(w_n^*)) = w_n^* \ , \tag{11.4}$$

which is a contradiction.

We now want to show that $\rho_{n+1}^* > \rho_n^*$ and $w_{n+1}^* < w_n^*$ for all n. That follows because the composite function $\rho_n(w_n(\rho))$ is strictly increasing in n on the interval $(0, 1)$, while the composite function $w_n(\rho_n(w))$ is strictly decreasing in n on the interval $(f^{-1}(1), \infty)$.

Now we turn to the limits in (2.4). Clearly, the first limit is implied by the second, so we focus on the second. Given that $\{n(1 - \rho_n^*) : n \geq 1\}$ is a sequence of positive numbers, there must exist a subsequence $\{n_k(1 - \rho_{n_k}^*) : k \geq 1\}$, which approaches one of 0, $+\infty$ or ζ for some

ζ with $0 < \zeta < \infty$. At this point we invoke Proposition 1 of Halfin and Whitt (1981), which applies to subsequences as well as the full sequence. First suppose that $n_k(1 - \rho^*_{n_k}) \to 0$ as $k \to \infty$, which implies that $\rho^*_{n_k} \to 1$ and $\sqrt{n_k}(1 - \rho^*_{n_k}) \to 0$ as $k \to \infty$. Then Proposition 1 of Halfin and Whitt implies that $w^*_{n_k} \to \infty$, which in turn implies that $\rho^*_{n_k} = \rho_{n_k}(w^*_{n_k}) \to 0$, which is impossible.

Next suppose that $n_k(1 - \rho^*_{n_k}) \to \infty$ as $k \to \infty$. By (1.3), that implies that $w_{n_k}(\rho^*_{n_k}) \to 0$, which in turn implies that $\rho^*_{n_k} = \rho_{n_k}(w_{n_k}(\rho^*_{n_k})) \to \gamma > 1$, which is impossible.

Suppose that $n_k(1 - \rho^*_{n_k}) \to \zeta$ for $0 < \zeta < \infty$. By Proposition 1 of Halfin and Whitt, that forces the limit $\pi_{n_k}(\rho^*_{n_k}) \to 1$ and $w^*_{n_k} = w_{n_k}(\rho^*_{n_k}) \to 1/\zeta$. However, since $\rho^*_{n_k} \to 1$, we must have $w^*_{n_k} = w_{n_k}(\rho^*_{n_k}) \to f^{-1}(1)$. Hence, we must have $\zeta = 1/f^{-1}(1)$. Since all convergent subsequences must have the same limits, the entire sequences must themselves converge to the indicated limits in (2.4).

Proof of Theorem 2.2. First focus on the queue-length process. There are two parts to the argument. One part is to show that, asymptotically, the state in which the queue is empty but all servers are busy can be treated as a lower reflecting barrier. The second part is to establish the limits for the queue-length process with this added lower barrier. We do this second part first.

With the inserted lower barrier, we can relate the equilibrium queue-length process to the queue-length process in the $M/M/1$ queue. Specifically, with the lower barrier in place, the process $\{Q_n(t/n) : t \geq 0\}$ is distributed as the queue-length process of an $M/M/1$ queue with service rate 1 and arrival rate ρ^*_n. Thus, we obtain the limit

$$\{(1 - \rho^*_n)Q_n((1 - \rho^*_n)^{-2}t/n) : t \geq 0\} \Rightarrow \{R(t; -1, 2) : t \geq 0\} . \tag{11.5}$$

We obtain the limit in (2.5) by multiplying and dividing by n in two places in (11.5) and using the fact that $n(1 - \rho^*_n) \to \zeta \equiv 1/f^{-1}(1)$ by the second limit in (2.4).

Similarly, assuming the lower barrier, the process $\{nW_{n,\lfloor t \rfloor} : t \geq 0\}$ behaves like the process $\{W_{\lfloor t \rfloor} : t \geq 0\}$ in the $M/M/1$ queue with service rate 1 and arrival rate ρ^*_n. Hence, we have the limit

$$\{n(1 - \rho^*_n)W_{n,\lfloor (1-\rho^*_n)^{-2}t \rfloor} : t \geq 0\} \Rightarrow \{R(t; -1, 2) : t \geq 0\} \tag{11.6}$$

in D as $n \to \infty$. Since $n(1 - \rho^*_n) \to \zeta$ as $n \to \infty$, the limit in (11.6) implies the limit in (2.6).

It remains to complete the first part of the argument, i.e., to show that the given processes are asymptotically equivalent to the processes with the inserted lower barrier, with the indicated scaling. We can show that by bounding the given processes above and below by processes that converge to the same limit after scaling; e.g., see Corollary 12.11.6 in Whitt (2001); the M_2 topology there is equivalent to the standard J_1 topology because the limit process almost surely has continuous sample paths.

Let $N_n \equiv \{N_n(t) : t \geq 0\}$ be the given birth-and-death process representing the number of customers in the system in model n with traffic intensity ρ_n^*. Then the upper bound process N_n^u is the process N_n modified by inserting a reflecting lower barrier at n. For each ϵ, a lower bound process N_n^ϵ is constructed from N_n by increasing the service rate in certain states, while leaving the arrival rate unchanged. Specifically, for states k with $k \geq n$, the service remains fixed at n; for states k with $n(1-\epsilon) \leq k < n$, let the lower-bound service rate be n instead of k; for states k with $0 \leq k \leq n(1-\epsilon)$, let the lower-bound service rate be $n(1-\epsilon)$ instead of k. We thus have lower-bound birth-and-death processes for each ϵ and n. We can order these birth-and-death processes by a strong stochastic ordering, yielding

$$N_n^{l,\epsilon_1} \leq_{st} N_n^{l,\epsilon_2} \leq_{st} N_n \leq_{st} N_n^u \tag{11.7}$$

for all $n, \epsilon_1, \epsilon_2$ with $\epsilon_1 > \epsilon_2$, where $\leq_{st}$ means that there exists versions of the two stochastic processes defined on the same underlying probability space such that the sample paths are ordered with probability one; e.g., see Whitt (1981). Indeed, the processes can be constructed for each n so the the sample paths are ordered for all the processes (for all ϵ) by constructing the transitions from thinned versions of common Poisson processes.

The process we considered above in (11.5) is the queue-length process associated with the upper-bound process N_n^u, i.e.,

$$Q_n^u(t) \equiv N_n^u(t) - n, t \geq 0 \ . \tag{11.8}$$

The given queue-length process is

$$Q_n(t) \equiv [N_n(t) - n]^+, t \geq 0 \ , \tag{11.9}$$

where $[x]^+ \equiv \max\{x, 0\}$. The associated lower-bound processes are

$$Q_n^{l,\epsilon}(t) \equiv N_n^u(t) - n, t \geq 0 \ . \tag{11.10}$$

As a consequence of (11.7), we have

$$Q_n^{l,\epsilon_1} \leq_{st} Q_n^{l,\epsilon_2} \leq_{st} Q_n \leq_{st} Q_n^u \tag{11.11}$$

for all $n, \epsilon_1, \epsilon_2$ with $\epsilon_1 > \epsilon_2$. We will show that, after scaling, $Q_n^{l,\epsilon}$ has the same limit as Q_n^u as first $n \to \infty$ and then $\epsilon \to 0$.

Note that $\{Q_n^{l,\epsilon}(t/n) : t \geq 0\}$ is a birth-and-death process on the integers k with $k \geq -n$. For $k \geq -n\epsilon$, the birth rate is ρ_n^* and the death rate is 1; for $0 \leq k < -n\epsilon$, the birth rate is ρ_n^* and the death rate is $1 - \epsilon$. We will show that this structure implies that

$$\{(1-\rho_n^*)Q_n^{l,\epsilon}((1-\rho_n^*)^{-2}t/n) : t \geq 0\} \Rightarrow \{R_{\zeta\epsilon}(t; -1, 2) : t \geq 0\} \ , \qquad (11.12)$$

where R_ϵ is RBM with reflecting barrier at $-\epsilon$ starting at 0 and $\zeta \equiv \lim_{n\to\infty} n(1-\rho_n^*)$.

We obtain (11.12) because $\{(1-\rho_n^*)Q_n^{l,\epsilon}((1-\rho_n^*)^{-1}t/n) : t \geq 0\}$ is asymptotically equivalent to $\{(1-\rho_n^*)Q_n^{l,\epsilon,*}((1-\rho_n^*)^{-1}t/n) : t \geq 0\}$, where $Q_n^{l,\epsilon,*}$ is the process $Q_n^{l,\epsilon}$ modified by adding a reflecting lower barrier at $-n\epsilon$. To see why that is so, observe that $Q_n^{l,\epsilon,*}$ can be obtained from $Q_n^{l,\epsilon}$ by deleting all the excursions below the level $-n\epsilon$, i.e., portions of the sample path going below $-n\epsilon$ until the process again goes above the level $-n\epsilon$. The negative values of these segments correspond to the much-studied busy periods in the $M/M/1$ queue with traffic intensity bounded above by $1-\epsilon$. In the presence of the heavy-traffic scaling, the cumulative effect of these excursions over time and the maximum extent over space (over bounded time intervals) are asymptotically negligible.

To show that the maximum extent over space is asymptotically negligible under heavy-traffic scaling, let the lower-bound process $Q_n^{l,\epsilon}$ be bounded above by the process $Q_n^{l,\epsilon,u}$, which is defined to be $Q_n^{l,\epsilon}$ modified by adding an upper reflecting barrier at $-n\epsilon$. Then $Z_n \equiv -Q_n^{l,\epsilon,u} - n\epsilon$ behaves like the queue-length process in an $M/M/1$ queue with traffic intensity $\rho_n^*(1-\epsilon)$. Hence

$$\{(1-\rho_n^*)Z_n((1-\rho_n^*)^{-2}t/n) : t \geq 0\} \Rightarrow \{\theta(t) : t \geq 0\} \qquad (11.13)$$

in D, where $\theta(t) = 0$ for all $t \geq 0$, reasoning as in Section 5.3.2 of Whitt (2001). As a consequence,

$$\inf_{0 \leq t \leq T} \{(1-\rho_n^*)Q_n^{l,\epsilon}((1-\rho_n^*)^{-2}t/n) \Rightarrow -\zeta\epsilon \qquad (11.14)$$

for any $T > 0$ and $\epsilon > 0$, where $\zeta \equiv \lim_{n\to\infty} n(1-\rho_n^*)$.

To show that the cumulative extent of the excursions over time in the process $\{Q_n^{l,\epsilon}(t/n) : t \geq 0\}$ is asymptotically negligible under heavy-traffic scaling, note that (by virtue of $M/M/1$ busy-period properties) the length of an excursion below $-n\epsilon$ is stochastically bounded above by a random variable with mean $(1-\epsilon)^{-1}$ for all n, while the interval between excursions is a random variable with mean $(1-\rho_n^*)^{-1}$, which explodes as $n \to \infty$.

As a consequence of these stochastic bounds, the distance between the scaled versions of the two processes $Q_n^{l,\epsilon}$ and $Q_n^{l,\epsilon *}$ converges in distribution to 0. The demonstration is easily done by applying the triangle inequality after comparing these two processes to the scaled version of the process that remains constant at the lower barrier during each excursion.

Since

$$\{(1-\rho_n^*)Q_n^{l,\epsilon,*}((1-\rho_n^*)^{-2}t/n) : t \geq 0\} \Rightarrow \{R_{\zeta\epsilon}(t;-1,2) : t \geq 0\} \;, \tag{11.15}$$

where R_ϵ is RBM with reflecting barrier at $-\epsilon$ starting at 0, by the argument used to treat Q_n^u, we can apply the convergence-together theorem, e.g., Theorem 11.4.7 of Whitt (2001), to obtain (11.12).

Next, since R_ϵ is distributed as $R-\epsilon$ started at 0, we have the ordering

$$R - \epsilon \leq_{st} R_\epsilon \leq_{st} R \tag{11.16}$$

for all ϵ, from which we can deduce that

$$R_\epsilon \Rightarrow R \quad \text{in} \quad D \tag{11.17}$$

as $\epsilon \to 0$. Hence we do obtain (11.5) and thus (2.5).

Next note that the workload (or virtual waiting-time) process is easily seen to be the first passage time to emptiness in the process Q_n, ignoring future arrivals. Thus the bounding and sandwiching argument extends directly to the workload process. The two bounding processes correspond to the workload processes in the $M/M/1$ queue, so that they can be treated directly. The waiting-time process can then be treated as a random time change of the workload process, invoking the continuous mapping theorem with the composition map; e.g., see Section 13.2 of Whitt (2001).

Finally, it remains to establish limits for the steady-state distributions. Since the steady-state queue-length distribution is geometric except for a modified mass at the origin and the steady-state waiting time is exponential except for a modified mass at the origin, the limits can be obtained from the limits for $\pi_n(\rho_n^*)$ and $w_n(\rho_n^*)$. We use PASTA to deduce that the probability distribution upon arrival is the same as at an arbitrary time; we use Little's law to relate $w_n(\rho_n^*)$ to the mean queue length; we use well-known conditions for the convergence of scaled geometric distributions to the exponential distribution; see p. 2 of Feller (1971).

Proof of Theorem 3.1. The proof closely parallels the proof of Theorem 2.1. Having obtained the unique ρ_n^* and π_n^*, we observe that the sequence $\{\sqrt{n}(1-\rho_n^*) : n \geq 1\}$ must have a subsequence converging to one of 0, ∞ or β for $0 < \beta < \infty$. Again applying Halfin and Whitt, we find that the limits 0 and ∞ are not possible, because they lead to contradictions. Thus there must be a subsequence with a finite positive limit. As a consequence, $\rho_n^* \to 1$ through this subsequence. Thus we must have the corresponding subsequence of $\{\pi_n^* : n \geq 1\}$ converge to $f^{-1}(1)$. Since all subsequences must have this same limit, the entire sequence must converge. And all full sequences must converge. In this case we obtain directly the many-server heavy-traffic regime considered by Halfin and Whitt.

Proof of Theorem 5.1 (b). Note that $\rho_n^* = f(1 + w_n(\rho_n^*))$, so that, performing a Taylor series expansion of f about 1, we obtain

$$1 - \rho_n^* = 1 - f(1) - f'(\delta_n)\pi_n(\rho_n^*)/n(1 - \rho_n^*) \ , \tag{11.18}$$

where $\delta_n \to 1$ as $n \to \infty$. Hence we obtain

$$n(1 - \rho_n^*)^2 = -f'(\delta_n)\pi_n(\rho_n^*) \ . \tag{11.19}$$

We now apply Proposition 1 of Halfin and Whitt to subsequences. If the left side of (11.19) has a subsequence converging to 0, then the corresponding subsequence of the right side must converge to $-f'(1) \neq 0$, which is a contradiction. If the left side has a subsequence converging to infinity, then the corresponding subsequence of the right side must converge to 0, which is a contradiction. Suppose that the left side has a subsequence $\{n_k(1-\rho_n^*)^2\}$ with $n_k(1-\rho_n^*)^2 \to \beta^2$ as $k \to \infty$. Then

$$-f'(\delta_{n_k})\pi_{n_k}(\rho_{n_k}^*) \to -f'(1)\alpha(\beta) \ , \tag{11.20}$$

so that we must have

$$\beta = g^{-1}(-f'(1)) \tag{11.21}$$

for g in (5.2). Since all convergent subsequences must have the same limit, the full sequences converge to the indicated limits.

Proof of Theorem 6.1. Paralleling the proof of Theorem 2.1, start by observing that the sequence $\{\sqrt{n}w_n^* : n \geq 1\}$ must have a subsequence converging to one of 0, ∞ or η for $0 < \eta < \infty$. The first two cases lead to counterexamples, because then the corresponding subsequence of $\{\rho_n^* : n \geq 1\}$ must converge to $f(0) = \gamma > 1$ and $f(\infty) = 0$. The first case is clearly impossible since $\rho_n^* < 1$ for all n. The second case implies that the corresponding subsequence of $\{\sqrt{n}w_n^* : n \geq 1\}$ must converge to 0, which is a contradiction with the infinite limit in this case.

Given that we are in the third case, we must be in the intermediate Halfin-Whitt heavy-traffic regime, which has the subsequence of $\{\rho_n^* : n \geq 1\}$ converge to 1. Thus, the limit os the subsequence $\{\sqrt{n}w_n^* : n \geq 1\}$ must be $f^{-1}(1)$. Since all convergent subsequences of $\{\sqrt{n}w_n^* : n \geq 1\}$ and $\{\rho_n^* : n \geq 1\}$ must have these same two limits, the entire sequences converge to these limits. From Proposition 1 of Halfin and Whitt, we deduce that the sequence $\{\sqrt{n}(1-\rho_n^*) : n \geq 1\}$ must converge to some β satisfying $0 < \beta < \infty$. We then deduce that $\beta = h^{-1}(f^{-1}(1))$ for h in (6.1).

Proof of Theorem 7.1. The first parts are the same as in Theorem 2.1, so we only treat the asymptotics. First note that we must have $\rho_n^* \to 1$ and $w_n^* \to 0$ as $n \to \infty$: To see why, suppose that $\rho_{n_k}^* \to \rho < 1$ for some subsequence. Then $w_{n_k} \to 0$ by (1.3) and (7.1) implies that $\rho = 1$, which is a contradiction. Given that $\rho_n^* \to 1$, (7.1) implies that we must also have $w_n^* \to 0$, because $f(0) = 1$ and $f(w) < 1$ for all $w > 0$.

Given that $\rho_n^* \to 1$ and $w_n^* \to 0$ as $n \to \infty$, we focus on the possible limits for subsequences $\{\sqrt{n_k}(1-\rho_{n_k}^*)\}$ of the sequence $\{\sqrt{n}(1-\rho_n^*)\}$. Use (7.1), Taylor's theorem and the smoothness condition to write

$$\sqrt{n}(1-\rho_n^*) = \delta_n - f'(\zeta_n)\sqrt{n}w_n^* \tag{11.22}$$

for some ζ_n with $0 \leq \zeta_n \leq w_n^*$ for all n sufficiently large. Reasoning as before, we see that convergence of subsequences of the sequence $\{\sqrt{n}(1-\rho_n^*)\}$ to 0 or ∞ lead to a contradiction: First suppose that $\sqrt{n_k}(1-\rho_{n_k}^*) \to 0$. Then, by (1.3), $\sqrt{n_k}w_{n_k} \to \infty$, which implies that $\sqrt{n_k}(1-\rho_{n_k}^*) \to \infty$, which is a contradiction. Next suppose that $\sqrt{n_k}(1-\rho_{n_k}^*) \to \infty$ with $\rho_{n_k}^* \to 1$. Then $\sqrt{n_k}w_{n_k} \to 0$, which implies that $\sqrt{n_k}(1-\rho_{n_k}^*) \to \delta$ by (11.22), which again is a contradiction. On the other hand, convergence to a positive finite limit β^* leads to the associated limits in (7.4). Since all convergent subsequences must have the same limit, there is convergence of the full sequences. The relation (11.22) leads to (7.5). There is a unique solution to (7.5) because the right side of (7.5) is decreasing as a function of β^*, going from $+\infty$ to 0.

References

Armony, M. and C. Maglaras. 2001. On customer contact centers with a call-back option: customer decisions, sequencing rules, and system design. New York University.

Billingsley, P. 1999. *Convergence of Probability Measures*, second ed., Wiley, New York.

Feller, W. 1971. *An Introduction to Probability and its Applications*, vol. II, second edition, Wiley, New York.

Garnett, O., A. Mandelbaum and M. I. Reiman. 2000. Designing a call center with impatient customers. Bell Laboratories, Murray Hill, N. J.

Glynn, P. W. and W. Whitt. 1991. A new view of the heavy-traffic limit for infinite-server queues. *Adv. Appl. Prob.* 23, 188–209.

Halfin, S. and W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Res.* 29, 567–588.

Hassin, R. and M. Haviv. 1995. Equilibrium threshold strategies: the case of queues with priorities. *Operations Res.* 45, 966–973.

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

Mandelbaum, A. and G. Pats. 1995. State-dependent queues: approximations and applications. *Stochastic Networks*, IMA Volumes in Mathematics and its Applications, F. P. Kelly and R. J. Williams, eds., Springer, New York, 239–282.

Mandelbaum, A. and N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* 36, 141–173.

Mendelson, H. and S. Whang. 1990. optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Res.* 38, 870–883.

Puhalskii, A. and M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Adv. Appl. Prob.* 32, 564–595.

Smith, D. R. and W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Tech. J.* 60, 39–55.

Whitt, W. 1981. Comparing counting processes and queues. *Adv. Appl. Prob.* 13, 207–220.

Whitt, W. 1990. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* 6, 335–351.

Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* 38, 708–723.

Whitt, W. 1993. Approximations for the $GI/G/m$ queue. *Production and Opns. Mgmt.* 2, 114–161.

Whitt, W. 1999. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Res. Letters* 24, 205–212.

Whitt, W. 2001. *Stochastic-Process Limits*, Springer, New York, to appear. Preliminary draft available online: http://www.research.att.com/ wow/