MANAGEMENT SCIENCE
Vol 38, No 5, May 1992
Printed in U S A

UNDERSTANDING THE EFFICIENCY OF MULTI-SERVER SERVICE SYSTEMS*

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

In the design and operation of service systems, it is important to determine an appropriate level of server utilization (the proportion of time each server should be working). In a multi-server queue with unlimited waiting space, the appropriate server utilization typically increases as the number of servers (and the arrival rate) increases. We explain this economy of scale and give a rough quantitative characterization. We also show how increased variability in the arrival and service processes tends to reduce server utilization with a given grade of service. As part of this analysis, we develop simple approximations for the mean steady-state waiting time and the full steady-state waiting-time distribution. These approximations exploit an infinite-server approximation for the probability of delay and a single-server approximation for the conditional waiting-time distribution given that waiting occurs. The emphasis is on simple formulas that directly convey understanding.

(QUEUES; MULTI-SERVER QUEUES; SERVICE SYSTEMS; UTILIZATION; ECONOMY OF SCALE; APPROXIMATIONS; PROBABILITY OF DELAY; PEAKEDNESS; HAYWARD'S APPROXIMATION)

1. Introduction

This paper emerged from a question posed by a factory manager. He was contemplating a new production line that would have four machines in one work area instead of just one. From experience, he had learned that it was desirable to run the previous line so that the single machine was utilized (i.e., busy processing) about 80% of the time. He wisely speculated that he might expect to have a higher server utilization with more machines. He asked if a higher utilization would indeed be appropriate and what it might be.

When this question was posed to me, my first reaction was the cautious (and obviously correct) response "it depends," but I also strongly believed that the answer is "yes" with appropriate qualifications. Moreover, I believed that queueing theory provides the basis for a simple quantitative answer.

Consider a service system with unlimited waiting space and the first-come first-served discipline in steady state. Let s be the number of servers, ρ the server utilization (the proportion of time each server is busy); and γ a constant giving a rough indication of the grade of service. As a rough rule of thumb, I proposed the *utilization equation*

$$(1 - \rho)\sqrt{s} = \gamma. \tag{1}$$

More specifically, I suggested that if the number of servers increased from s_1 to s_2 then the utilization should increase from ρ_1 to at least ρ_2 where $(1 - \rho_i)\sqrt{s_i} = \gamma$ for $i = 1, 2$. In other words, I suggested that (1) should serve as an approximation, and an approximate lower bound, for the way ρ should increase with s.

Qualitatively, (1) is consistent with an important well-known principle in the design and operation of service systems: *The appropriate level of server utilization typically increases with capacity*. (Supporting stochastic comparisons for this efficiency principle appear in Smith and Whitt 1981.) Quantitatively, (1) states that $1 - \rho$ should be approximately inversely proportional to $\sqrt{s}$. For example, if $\rho = 0.9$ is deemed appropriate

* Accepted by Linda Green; received March 19, 1990. This paper has been with the author 2 months for 1 revision.

0025-1909/92/3805/0708$01.25

for $s = 1$, then $\rho = 0.99$ should be roughly appropriate for $s = 100$. For the production line example above, if $\rho = 0.8$ was appropriate for $s = 1$, then the grade of service is $\gamma = 0.2$ and $\rho = 0.9$ (or something slightly higher) should be appropriate for $s = 4$.

This economy of scale associated with larger service systems is illustrated in Table 1, which shows the number of servers with individual service rate $\mu = 1$ required to yield at least a grade of service $\gamma = 0.2$ when the total arrival rate is $\lambda = 100$ and there are n separate facilities each with arrival rate λ/n, assuming that (1) is valid. Of course part of the advantage of larger systems is due to the fact that servers must be provided in integer quantities.

The first purpose of this paper is to provide theoretical, heuristic and empirical support for the utilization equation (1). Formula (1) was chosen to give the value of ρ as a function of s that tends to keep a measure of congestion fixed. The particular measure of congestion is the probability of delay, i.e., $P(W > 0)$ where W is the steady-state waiting time before beginning service. We claim that if s and ρ are changed with (1) holding for some fixed γ then $P(W > 0)$ should remain approximately unchanged. We provide support for this idea in §2.

The problem stated above is finding the utilization ρ as a function of s for given grade of service γ. This problem is closely related to the design problem of finding the number of servers s as a function of the arrival rate λ for given service-time distribution and given grade of service γ. For this essentially equivalent problem, our reasoning yields the equation

$$s = \lambda + \gamma\sqrt{\lambda}. \tag{2}$$

In §2 we develop (2) from an infinite-server approximation and show that (1) follows from (2) when λ and s are large. This line of reasoning suggests that (1) and (2) should be more accurate for large s, and this is so, but we believe they are useful more generally.

We hasten to point out that (1) and (2) are not new, but they do not seem to be nearly as well known as they should be. For example, (1) is discussed by Halfin and Whitt (1981) and (2) is discussed by Newell (1973, 1982), Grassman (1986, 1988) and Kolesar (1986). The scaling in (1) and (2) has also been identified as important for multi-server loss systems and stochastic networks; see Jagerman (1974), McKenna, Mitra and Ramakrishnan (1981), Whitt (1984), Hunt and Kelly (1989) and Reiman (1989, 1990). Hence, our discussion of (1) and (2) should be regarded as a review and an elaboration.

A second purpose of the present paper is to go beyond (1) and (2). First, we want to determine the effect of variability in the arrival and service processes. We also investigate this in §2. Our analysis leads again to (2) but with the grade of service parameter γ being inversely proportional to $\sqrt{z}$, where z is a measure of variability called the peakedness; see Eckberg (1983, 1985) and Whitt (1984).

TABLE 1

The Number of Servers with Individual Service Rate $\mu = 1$ Required to Yield a Grade of Service $\gamma = 0.2$ for a Total Arrival Rate of 100 When There Are n Separate Service Facilities Each with Arrival Rate 100/n, Based on (1)

Number of Separate Facilities	Arrival Rate Per Facility	s from Equation (2)	Number of Servers Per Facility	Total Number of Servers
1	100	102	102	102
4	25	26	26	104
25	4	4.4	5	125
100	1	1.2	2	200

We also want to consider other measures of congestion besides the probability of delay $P(W > 0)$, such as the mean waiting time EW and the tail probability $P(W > t)$ for $t > 0$. We do this in §3. Our strategy is to use the infinite-server approximation for the probability of delay $P(W > 0)$ and a single-server approximation for the conditional probability $P(W > t \mid W > 0)$. (There is some history for this too; e.g., see Newell 1973 and Hokstad 1978.) This leads to an interesting new simple approximation for EW that uses both the peakedness mentioned above and the limiting value of the index of dispersion for counts (IDC) in Sriram and Whitt (1986) and Fendick and Whitt (1989). This seems to be the first time that these approximation tools have been used together.

The analysis of $E[W \mid W > 0]$ and $P(W > t \mid W > 0)$ in §3 indicates that these measures of congestion actually decrease as s increases under (1). This is the reason we regard (1) as a lower bound for the way ρ should increase with s for a given grade of service γ. The way ρ should increase with s depends on the performance measure of interest.

We end this paper in §4 with our conclusions. For related material in much the same spirit, see Newell (1973, 1982).

2. The Probability of Delay

In this section we give support for formulas (1) and (2). In §2.1 we show how to derive (1) from (2). In §2.2 we review a heavy-traffic limit theorem supporting (1) and (2) in a special case. In §2.3 we introduce an infinite-server (IS) approximation to heuristically derive (2). In the next four subsections, we apply the IS approximation to treat $M/D/s$, $M/G/s$, $G/D/s$ and $G/G/s$ models, respectively. We also consider numerical examples to provide empirical validation. We include separate treatment of special cases, because the entire analysis is much easier to present and understand for the special cases. In §2.8 we discuss other simple approximation formulas following from this analysis of the probability of delay. Finally, in §2.9 we provide theoretical support for our IS approximation.

2.1. *The Connection Between* (1) *and* (2)

Henceforth in this paper let the service rate of each server be 1, which corresponds to measuring time in the scale of mean service times. As before, let the arrival rate be λ. By Little's law ($L = \lambda W$) applied to the servers, the server utilization ρ coincides with the traffic intensity, i.e., $\rho = \lambda/s$; e.g., see Example 11-8, p. 400, of Heyman and Sobel (1982). Consequently, (1) is equivalent to

$$\lambda = s - \gamma\sqrt{s}. \tag{3}$$

Formula (3) is not quite the same as (2), but it nearly is when γ is small compared to $\sqrt{s}$.

PROPOSITION 2.1. *Formula* (2) *implies that*

$$\lambda = s + \frac{\gamma^2}{2} - \gamma\sqrt{s + \frac{\gamma^2}{4}}, \tag{4}$$

so that

$$s - \gamma\sqrt{s} \le \lambda \le s - \gamma\sqrt{s} + \frac{\gamma^2}{2}. \tag{5}$$

PROOF. Solve (2) for λ to obtain (4). Since $\sqrt{x + y} \le \sqrt{x} + \sqrt{y}$ for positive x and y, (4) implies the lower bound in (5) as well as the obvious upper bound. □

2.2. *Heavy-Traffic Limits*

Direct support for the utilization equation (1) is provided by a heavy-traffic limit theorem in Halfin and Whitt (1981). They consider a $GI/M/s$ queue with different possible values of s and ρ (e.g., obtained by changing the service rate as well as s). (One would expect that similar results hold for more general models, but they evidently have not yet been established.) Let $W_{s\rho}$ be a random variable with the steady-state waiting-time distribution for each positive integer s and each ρ, $0 < \rho < 1$.

PROPOSITION 2.2 (Halfin and Whitt 1981). *The limit* $P(W_{s\rho} > 0) \to \alpha$ *as* $s \to \infty$, *where* $0 < \alpha < 1$, *holds if and only if*

$$(1 - \rho)\sqrt{s} \to \gamma \qquad as \qquad s \to \infty,$$

where $0 < \gamma < \infty$.

Halfin and Whitt (1981) also derive an explicit, but somewhat complicated, expression for the limiting probability of delay α in Proposition 2.2, which has been found to be a good approximation; see Whitt (1985), where other approximations are also developed and evaluated. In particular,

$$\alpha = [1 + \sqrt{2\pi}\beta\Phi(\beta)\exp(\beta^2/2)]^{-1}, \tag{6}$$

where $\Phi(x) = P(N(0, 1) \le x)$ with $N(m, \sigma^2)$ being a normal random variable having mean m and variance σ^2, and $\beta = 2\gamma/(1 + c_a^2)$ with c_a^2 being the squared coefficient of variation (SCV, i.e., the variance divided by the square of the mean) of an interarrival time; see Proposition 1 and Theorem 4 of Halfin and Whitt (1981). The limits are stated for the continuous-time queue length process, but also apply to the embedded sequence at arrival epochs, as shown in the proof of Theorem 3 there.

Proposition 2.2 says that (1) is asymptotically correct as s gets large. Below we present heuristic arguments that are intended to provide an intuitive explanation. Similar asymptotic results have been obtained for the blocking probability in loss models; see Jagerman (1974) and Whitt (1984).

2.3. *The Infinite-Server (IS) Approximation*

To understand the basic equation (2), we believe that it is helpful to consider an associated model with infinitely many servers, in which every customer begins service immediately upon arrival. The associated IS model should have the same arrival process and the same service-time distribution as the s-server model of interest. The idea is that we should be able to deduce (2) by analyzing the much-easier-to-analyze IS model.

What we want to determine is the function $s(\lambda)$ such that the probability of delay in the $s(\lambda)$-server queue is approximately some prescribed value, independent of the arrival rate λ. We contend that in some rough sense the level of congestion experienced by the first s servers in the IS model should be about the same as the level of congestion in the s-server model. For example, if all s servers are essentially never busy in the IS model, then the same will be true in the s-server model. As an approximation, it seems reasonable to conclude that the probability of delay $P(W > 0)$ in an s-server model should remain roughly constant as a function of s and λ when the probability of having more busy servers than s in the IS model remains constant. Let N be the steady-state number of busy servers in the IS model. We do not claim that $P(N \ge s)$ actually equals $P(W > 0)$ in the s-server model, but that if we increase λ as we increase s so that $P(N \ge s)$ remains fixed, then we contend that $P(W > 0)$ should approximately remain fixed too.

In summary, *our basic hypothesis is that the probability of delay in the* $s(\lambda)$*-server model should be approximately some prescribed value independent of* λ *if the probability of having* $s(\lambda)$ *busy servers in the associated IS model is approximately some prescribed*

value independent of λ. We do not claim that these two prescribed probabilities are necessarily equal, which would determine the grade of service γ in (2), but only that we can use the IS model to support formula (2), without identifying γ. We consider this basic IS approximation hypothesis as an intuitively reasonable starting point, without requiring proof. We substantiate it *empirically* below when we evaluate the resulting approximations. In §2.9 we also provide theoretical support for the IS approximation.

Of course, IS approximations have been considered before, e.g., by Newell (1973, 1982) and Grassman (1988). However, note that these were direct approximations, specifying γ as well as supporting (2).

2.4. *Deterministic Service Times and Poisson Arrivals* ($M/D/\infty$)

We now apply the basic hypothesis in §2.3 and consider an IS model. To start with, suppose that all customers have deterministic service times of length 1. Let $A(t)$ count the number of arrivals in $(0, t]$. Here is the key property.

PROPOSITION 2.3. *If all the service times are* 1 *in an IS model, then* $N(t) = A(t) - A(t-1)$.

PROOF. We assume that a departure at time t is not counted, while an arrival at time t is. Hence the interval includes t but not $t-1$. Each customer arriving before time $t-1$ will be gone by time t, while each customer arriving in the interval $(t-1, t]$ will still be there at time t. □

To be even more concrete, we suppose that the arrival process is a Poisson process with rate λ.

PROPOSITION 2.4. *In an* $M/D/\infty$ *model with service times of length* 1, $N(t)$ *has a Poisson distribution with mean* λ *for each* $t > 1$.

PROOF. Apply Proposition 2.3, noting that $A(t) - A(t-1)$ has the same distribution as $A(1)$. □

This Poisson property leads directly to (2). First, it is well known that the variance equals the mean for a Poisson distribution. Second, it is known that the Poisson distribution can be approximated by a normal distribution when λ is suitably large, by virtue of the central limit theorem; see Chapters VI and X of Feller (1968). (For this model, large λ is used only at this step.) Hence, the number of busy servers is approximately normally distributed with mean and variance λ for suitably large λ. Since the mean and variance of $N(t)$ are both λ, the standard deviation is $\sqrt{\lambda}$ and the SCV is $1/\lambda$. Hence, as λ increases, the distribution of $N(t)$ becomes more concentrated about its mean (in a relative sense).

We consider this asymptotic concentration of the distribution about its mean as λ increases the primary explanation for the economy of scale associated with more servers. Since the steady-state number of busy servers tend to concentrate around λ as λ increases, we can choose s to be very close to λ (and obtain high ρ), i.e., we can achieve (1) and (2) for a given grade of service γ.

Since the steady-state number of busy servers, N, is approximately normally distributed with mean λ and standard deviation $\sqrt{\lambda}$,

$$\pi \equiv P(N \geq s) = P\left(\frac{N-\lambda}{\sqrt{\lambda}} \geq \frac{s-\lambda}{\sqrt{\lambda}}\right) \approx 1 - \Phi\left(\frac{s-\lambda}{\sqrt{\lambda}}\right), \tag{7}$$

where Φ is again the standard normal cdf. Now we define γ in terms of π in (7) by solving

$$1 - \Phi(\gamma) = \pi. \tag{8}$$

TABLE 2

Descriptive Characteristics of an M/D/s Queue (Poisson Arrival Process, $c_a^2 = 1$) as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W > 0)$	0.900	0.884	0.874	0.870
	$\sqrt{s}E[W \mid W > 0]$	5.00	5.07	5.10	5.12
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W > 0)$	0.800	0.775	0.759	0.754
	$\sqrt{s}E[W \mid W > 0]$	2.50	2.58	2.61	2.62
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W > 0)$	0.500	0.490	0.479	0.475
	$\sqrt{s}E[W \mid W > 0]$	1.00	1.09	1.12	1.13
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W > 0)$		0.163	0.190	0.195
	$\sqrt{s}E[W \mid W > 0]$		0.61	0.63	0.63

Finally, from (7) and (8) we obtain $(s - \lambda)/\sqrt{\lambda} = \gamma$, which coincides with (2). However, we have yet to give a definite interpretation for γ in the s-server model.

To show the accuracy of (1), we consider some numerical examples. First, Table 2 displays descriptive characteristics of the $M/D/s$ queue as a function of the grade of service γ and the number of servers s under the assumption that (1) holds. All numerical values were obtained from the tables of Kühn (1976) and Seelen, Tijms and van Hoorn (1985). (Results for the conditional mean $E[W \mid W > 0]$ also appear in the tables; they will be discussed in §3.)

We claimed that the probability of delay $P(W > 0)$ should be nearly constant as we change s in this situation. This is borne out by Table 2, but the relationship weakens as the load decreases (as γ increases and thus λ decreases). This is consistent with the heuristic argument here and the limit theorem in §2.2, which indicate that (1) should become more appropriate as λ increases.

2.5. *General Service-Time Distributions* ($M/G/\infty$)

The distribution of the (steady-state) number of busy servers in an IS model with Poisson arrivals actually depends on the service-time distribution only through its mean. Consequently, the analysis of §2.4 extends to general service-time distributions. To see this directly, suppose that each service time assumes the values d_i with probability p_i, $1 \le i \le n$, where $\sum_{i=1}^{n} p_i d_i = 1$, so that the mean service time is again 1. Since the arrival process is Poisson, this model is equivalent to an IS model with the superposition of n independent Poisson arrival processes, where the ith arrival process has rate λp_i and deterministic service times of length d_i. Hence, for $t > \max\{d_i : 1 \le i \le n\}$, the number of busy servers has the distribution of the sum of n independent Poisson variables where the ith variable has mean $\lambda p_i d_i$. Since the sum of independent Poisson variables has a Poisson distribution with a mean equal to the sum of the means, we see that indeed the steady-state number of busy servers depends on the service-time distribution only through its mean.

The distribution of the number of customers in the system when there is a Poisson arrival process and only s servers is not independent of the service-time distribution beyond the mean, but this analysis shows that it will tend to be nearly so when s is large.

TABLE 3

Descriptive Characteristics of the M/M/s Queue as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W > 0)$	0.900	0.891	0.885	0.883
	$\sqrt{s}E[W \mid W > 0]$	10.00	10.00	10.00	10.00
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W > 0)$	0.800	0.788	0.779	0.755
	$\sqrt{s}E[W \mid W > 0]$	5.00	5.00	5.00	5.00
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W > 0)$	0.500	0.509	0.508	0.507
	$\sqrt{s}E[W \mid W > 0]$	2.00	2.00	2.00	2.00
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W > 0)$		0.174	0.209	0.217
	$\sqrt{s}E[W \mid W > 0]$		1.00	1.00	1.00
$\gamma = 2.0$	ρ		0.00	0.60	0.80
	$P(W > 0)$			0.0124	0.0196
	$\sqrt{s}E[W \mid W > 0]$			0.50	0.50

Table 3 displays descriptive characteristics for the $M/M/s$ model, assuming that (1) holds. As in Table 2, the probability of delay $P(W > 0)$ is indeed nearly constant in s when γ is suitably small. For $\gamma \geq 1.0$, formula (1) differs significantly from (4), which is the direct consequence of (2) derived in §2.1. For $\gamma = 1.0$, (4) yields $\rho = 0.905$, 0.82 and 0.61 for $s = 100$, 25 and 4. For $\gamma = 2.0$, (4) yields $\rho = 0.82$ and 0.67 for $s = 100$ and 25. Under (4), $P(W > 0)$ actually decreases slightly as s increases when $\gamma = 1.0$ or 2.0. If γ, λ and s are all large, then it is evident that the probability of delay is approximately constant as λ increases under (4).

2.6. *General Arrival Processes* ($G/D/\infty$)

Now suppose that the arrival process is general (a stationary point process) with $A(t)$ again representing the number of arrivals in the interval $(0, t]$. As in §2.4, let the service times be deterministic with mean 1. By Proposition 2.3, the steady-state number of busy servers in the IS model is equal to the number of arrivals in $(t - 1, t]$, which here is distributed as $A(\lambda)$.

Now we want to consider what happens when the arrival rate λ gets large. We need to be careful about how the arrival process changes as λ increases. For example, we could just scale time or we could consider the superposition of many slower component processes. There is no difficulty when the arrival process is a Poisson process as in §2.4 and §2.5, because then both methods produce a Poisson process, but this is not the case with non-Poisson processes. *We assume that the arrival rate λ is increased simply by scaling time in a given general arrival process*, so that $A(\lambda)$ represents the number of busy servers as a function of λ. (That is, we start with $A(1)$ for $\lambda = 1$ and obtain $A(\lambda)$ for general λ.)

For large λ, we characterize the distribution of the number N of busy servers by assuming a central limit theorem for the general counting process $A(t)$. In particular, we assume that

$$\lim_{t\to\infty} P\left(\frac{A(t)-t}{\sqrt{c_a^2 t}} \le x\right) = \Phi(x) \tag{9}$$

for all x, where

$$c_a^2 = \lim_{t\to\infty} \frac{\text{Var } A(t)}{EA(t)}. \tag{10}$$

For a renewal process, c_a^2 is the SCV of an interarrival time.

Assumption (9) can be expected to hold for any stationary point process arising in practice. In particular, (9) holds whenever a corresponding central limit theorem holds for the associated partial sums of the interarrival times; see Theorem 6 of Glynn and Whitt (1988). Moreover, many sufficient conditions exist for the central limit theorems for partial sums of dependent random variables; e.g., see Theorem 20.1 of Billingsley (1968). (See §4.6 of Newell 1982 for related discussion.) For complicated stationary point processes, the only difficulty is identifying the constant c_a^2 in (10).

Paralleling (7), from (9) we obtain

$$\begin{aligned}\pi \equiv P(N \ge s) &= P(A(\lambda) \ge s) \\ &= P\left(\frac{A(\lambda)-\lambda}{\sqrt{c_a^2\lambda}} \ge \frac{s-\lambda}{\sqrt{c_a^2\lambda}}\right) \approx 1 - \Phi\left(\frac{s-\lambda}{\sqrt{c_a^2\lambda}}\right).\end{aligned} \tag{11}$$

Combining (8) and (11), we obtain the generalization of (2)

$$s = \lambda + \gamma\sqrt{z\lambda}, \tag{12}$$

where $z = c_a^2$ in (10). In practice it remains to identify the parameter c_a^2.

Tables 4 and 5 display the probability of delay $P(W > 0)$ as a function of s assuming (1) holds in $GI/D/s$ models. Table 4 illustrates a low-variable arrival process with $c_a^2 = 0.25$, while Table 5 illustrates a high-variability arrival process with $c_a^2 = 4.0$. These numerical results are obtained from the tables of Seelen, Tijms and van Hoorn (1985). These tables also support (1), but show that the quality of the approximation deteriorates as the variability increases.

TABLE 4

Descriptive Characteristics of a GI/D/s Queue (Renewal Arrival Process) with Interarrival-Time Squared Coefficient of Variation $c_a^2 = 0.25$ as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W>0)$	0.775	0.762	0.754	0.751
	$\sqrt{s}E[W \mid W>0]$	1.29	1.30	1.31	1.31
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W>0)$	0.578	0.565	0.557	0.554
	$\sqrt{s}E[W \mid W>0]$	0.67	0.68	0.68	0.69
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W>0)$	0.163	0.187	0.195	0.198
	$\sqrt{s}E[W \mid W>0]$	0.30	0.31	0.32	0.32
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W>0)$		0.008	0.018	0.021
	$\sqrt{s}E[W \mid W>0]$		0.19	0.19	0.19

TABLE 5

Descriptive Characteristics of a GI/D/s Queue (Renewal Arrival Process) with Interarrival-Time Squared Coefficient of Variation $c_a^2 = 4.0$ as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W > 0)$	0.969	0.958	0.943	0.937
	$\sqrt{s}E[W \mid W > 0]$	18.2	19.2	19.1	20.1
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W > 0)$	0.931	0.915	0.888	0.877
	$\sqrt{s}E[W \mid W > 0]$	8.0	9.2	9.9	10.1
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W > 0)$	0.718	0.769	0.730	0.712
	$\sqrt{s}E[W \mid W > 0]$	1.84	3.12	3.92	4.10
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W > 0)$		0.421	0.496	0.484
	$\sqrt{s}E[W \mid W > 0]$		1.04	1.87	2.10

2.7. *General Arrival and Service Processes* ($G/G/\infty$)

The situation is more complicated when both the service times and the arrival process are general, so we will not try to give a detailed supporting argument. However, it is known that the steady-state number of busy servers in the IS model is typically asymptotically normally distributed as λ increases. (The first proof seems to be by Borovkov 1967.) This is relatively easy to see when the service-time distribution is concentrated on only finitely many points as in §2.5, then the argument is essentially the same as in §2.6; see Glynn and Whitt (1991).

By Little's law, the mean number of busy servers is always λ. In great generality, the variance is $z\lambda$ where z is a constant called the peakedness reflecting the variability of the arrival and service processes; see Eckberg (1983), (1985) and Whitt (1984). Indeed, an asymptotic expression for z for large λ when the service times are i.i.d. and independent of the arrival process is

$$z = 1 + (c_a^2 - 1)\int_0^\infty [1 - G(x)]^2 dx, \tag{13}$$

where G is the cdf of the service-time distribution (assumed to have mean 1) and c_a^2 is given by (9). For example, if G is exponential, then $\int_0^\infty [1 - G(x)]^2 dx = \frac{1}{2}$, so that $z = (c_a^2 + 1)/2$. The maximum value of $\int_0^\infty [1 - G(x)]^2 dx$ associated with mean 1 occurs with a deterministic service-time distribution, yielding $z = c_a^2$. (Hence, (13) reduces to (10) for the $G/D/\infty$ case.)

In summary, we use the following result. To state it, let N_λ be the steady-state number of busy servers in the $G/G/\infty$ model with arrival rate λ. As in §2.6 we assume that the arrival process changes with λ by simple scaling.

PROPOSITION 2.5 (Borovkov 1967). *For a $G/G/\infty$ model,*

$$P(N_\lambda \geq s) \to 1 - \Phi\left(\frac{s - \lambda}{\sqrt{z\lambda}}\right) \quad \text{as} \quad \lambda \to \infty$$

for z in (13).

TABLE 6

Descriptive Characteristics of a GI/M/s Queue (Renewal Arrival Process) with Interarrival-Time Squared Coefficient of Variation $c_a^2 = 0.25$ as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W > 0)$	0.843	0.849	0.850	0.850
	$\sqrt{s}E[W \mid W > 0]$	6.38	6.31	6.28	6.26
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W > 0)$	0.694	0.710	0.716	0.717
	$\sqrt{s}E[W \mid W > 0]$	3.27	3.19	3.15	3.14
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W > 0)$	0.302	0.371	0.398	0.406
	$\sqrt{s}E[W \mid W > 0]$	1.43	1.32	1.28	1.26
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W > 0)$		0.066	0.112	0.125
	$\sqrt{s}E[W \mid W > 0]$		0.72	0.65	0.64

The main point is that the analysis of §2.4 applies once again, with the sole modification that the standard deviation $\sqrt{\lambda}$ should be replaced by $\sqrt{z\lambda}$. Instead of (2), we obtain (12) with z in (13).

Tables 6 and 7 display the probability of delay $P(W > 0)$ as a function of s assuming that (1) holds in $GI/M/s$ models. Table 6 illustrates a low-variability arrival process with $c_a^2 = 0.25$, while Table 7 illustrates a high-variability arrival process with $c_a^2 = 4.0$. The behavior is evidently consistent with Tables 2–5.

TABLE 7

Descriptive Characteristics of a GI/M/s Queue (Renewal Arrival Process) with Interarrival-Time Squared Coefficient of Variation $c_a^2 = 4.0$ as a Function of the Grade of Service γ and the Number of Servers s, Assuming That the Utilization Equation (1) Holds

Grade of Service	Congestion Measures	Number of Servers, s			
		1	4	25	100
$\gamma = 0.1$	ρ	0.90	0.95	0.98	0.99
	$P(W > 0)$	0.957	0.947	0.935	0.929
	$\sqrt{s}E[W \mid W > 0]$	23.4	24.2	24.7	24.9
$\gamma = 0.2$	ρ	0.80	0.90	0.96	0.98
	$P(W > 0)$	0.908	0.892	0.871	0.861
	$\sqrt{s}E[W \mid W > 0]$	10.8	11.7	12.2	12.4
$\gamma = 0.5$	ρ	0.50	0.75	0.90	0.95
	$P(W > 0)$	0.684	0.715	0.690	0.675
	$\sqrt{s}E[W \mid W > 0]$	3.16	4.14	4.68	4.85
$\gamma = 1.0$	ρ	0.00	0.50	0.80	0.90
	$P(W > 0)$		0.373	0.427	0.422
	$\sqrt{s}E[W \mid W > 0]$		0.79	2.16	2.34

TABLE 8

The Values of $\sqrt{z}P(W = 0)$ as a Function of z and s When $\gamma = 0.1$ in Tables 2–7

Arrival and Service Characteristics			Number of Servers, s			
c_a^2	c_s^2	z	1	4	25	100
4	0	4.0	0.086	0.106	0.130	0.142
4	1	2.5	0.068	0.084	0.103	0.112
1	1	1.0	0.100	0.109	0.115	0.117
1	0	1.0	0.100	0.116	0.126	0.130
0.25	1	0.625	0.124	0.119	0.119	0.119
0.25	0	0.25	0.113	0.119	0.123	0.124

2.8. *Other Implications*

Formula (12) leads to a simple rough approximation for the probability of no delay, $P(W = 0)$. We assume, as a rough approximation, that $P(W = 0)$ depends only on the grade of service γ when (12) holds. Then we apply the exact result $P(W = 0) = 1 - \rho$ for the $M/G/1$ queue. Hence, for the general $G/G/s$ queue satisfying (12) with $\gamma < 1$, we propose the rough approximation

$$P(W = 0) \approx \gamma = \frac{(1 - \rho)\sqrt{s}}{\sqrt{z}}. \tag{14}$$

The simple approximation (14) corresponds to using the value of $P(W > 0)$ for $s = 1$ for higher s in Tables 2–7. This approximation is not too bad when γ is small, but dramatically deteriorates as $\gamma \to 1$.

Given that (1) holds, formulas (12) and (14) lead us to predict that $P(W = 0)$ should be approximately inversely proportional to $\sqrt{z}$ where z is the peakedness in (13). With exponential service times, $z = (c_a^2 + 1)/2$; with deterministic service times, $z = c_a^2$. To test this hypothesis, Tables 8 and 9 display the values of $\sqrt{z}P(W = 0)$ for the cases $\gamma = 0.1$ and $\gamma = 0.2$ in Tables 2–7. From Tables 8 and 9, we see that formulas (12) and (14) evidently capture the variability effect relatively well for the probability of delay, although the quality of the approximation evidently deteriorates in the case of high variability ($c_a^2 = 4.0$) and few servers ($s \leq 4$).

Assume that (12) is valid, we see that we should obtain approximately the same grade of service γ with parameter triples (s, λ, z) and $(s/z, \lambda/z, 1)$; just divide both sides of (12) by z. In particular, let $P(W > 0; \lambda, s, z)$ represent the steady-state probability of

TABLE 9

The Values of $\sqrt{z}P(W = 0)$ as a Function of z and s When $\gamma = 0.2$ in Tables 2–7

Arrival and Service Characteristics			Number of Servers, s			
c_a^2	c_s^2	z	1	4	25	100
4	0	4.0	0.138	0.170	0.224	0.246
4	1	2.5	0.145	0.171	0.204	0.220
1	1	1.00	0.200	0.212	0.221	0.225
1	0	1.00	0.200	0.225	0.241	0.246
0.25	1	0.625	0.241	0.229	0.224	0.223
0.25	0	0.25	0.211	0.218	0.222	0.223

delay in an s-server system with arrival rate λ and peakedness z. (The mean service time has been fixed at 1.) Then (12) supports the approximation

$$P(W > 0; \lambda, s, z) = P(W > 0; \lambda/z, s/z, 1). \tag{15}$$

Since a Poisson process has $z = 1$, (15) means that we can use a model with a Poisson arrival process to approximate a process with more complicated variability for computing $P(W > 0)$. In other words, *we obtain a Hayward-type approximation for delay systems paralleling the Hayward approximation for loss systems*; see Fredericks (1980) and Whitt (1984). Indeed, our analysis here is closely related to the asymptotic analysis in Whitt (1984).

2.9. *Theoretical Support for the IS Approximation*

We have presented the IS approximation as an intuitively appealing starting point to develop and understand (1) and (2). The IS approximation thus serves as a substitute or complement to the more technical limiting result in Proposition 2.2. However, in turn, *Proposition* 2.2 *can be used to establish the asymptotic validity of the IS approximation in a special case*. In particular, for $GI/M/s$ models, Propositions 2.2 and 2.5 together show that the IS approximation in the form we presented it is asymptotically correct as $s \to \infty$.

To state the result, as a slight modification of §2.2, let $W_{s\lambda}$ be the steady-state waiting time in a $GI/M/s$ model with service rate 1, s servers and arrival rate λ. Let the renewal arrival counting process change with λ by having $A_\lambda(t) = A_1(\lambda t)$, $t \geq 0$. Let N_λ be the steady-state number of busy servers in an associated $GI/M/\infty$ model with infinitely many servers and the same arrival and service processes.

PROPOSITION 2.6. *For the $GI/M/s$ and $GI/M/\infty$ models above,*

$$P(W_{s\lambda} > 0) \to \alpha \qquad as \qquad \lambda \to \infty,$$

where $0 < \alpha < 1$, and

$$P(N_\lambda > s) \to \eta \qquad as \qquad \lambda \to \infty,$$

where $0 < \eta < 1$, both hold if and only if $(1 - \rho)\sqrt{s} \to \xi \geq 0$ as $\lambda \to \infty$. If the limits hold, then α is given by (6) with $\lambda = \xi$ and $\eta = 1 - \Phi(\xi/\sqrt{z})$.

PROOF. Apply Propositions 2.1 and 2.2 to treat $W_{s\lambda}$ and Proposition 2.5 to treat N_λ. □

It is significant that, even for the special case of a Poisson arrival process, $\alpha \neq \eta$ in Proposition 2.6, so that our particular form of the IS approximation in §2.3 with γ acting as a free parameter is important.

3. The Waiting-Time Distribution

To approximately analyze the full steady-state waiting-time distribution, we separate the probability of delay $P(W > 0)$ from the conditional distribution of the waiting time given that all servers are busy. We note that an $M/M/s$ model with utilization ρ and individual service rate 1 behaves exactly like an $M/M/1$ model with utilization ρ and service rate s whenever all servers are busy. We use a heuristic extension of this property for general $G/G/s$ models.

3.1. *The $M/M/s$ Queue*

For s-server queues, we first consider an $M/M/s$ model. Since the conditional waiting time given that all servers are busy is then exponential with mean $1/s(1 - \rho)$, we immediately obtain the following result.

PROPOSITION 3.1. *For an $M/M/s$ model,*

$$P(W > x \mid W > 0) = e^{-\gamma\sqrt{s}} \quad \textit{and} \tag{16}$$

$$E[W \mid W > 0] = 1/\gamma\sqrt{2}, \tag{17}$$

assuming that (1) *holds.*

From (16) and (17) we see that these measures of congestion actually *decrease* as s increases, assuming that (1) holds. Consistent with (17), Table 3 shows that $\sqrt{s}E[W \mid W > 0]$ is indeed constant as a function of s in the $M/M/s$ model when (1) holds.

3.2. *The $G/G/s$ Queue*

As in §2.6 and §2.7, experience indicates that (16) and (17) need modification when we do not have the $M/M/s$ model. Indeed, experience with the $M/G/1$ special case indicates that $E(W \mid W > 0)$ should depend on the service-time distribution beyond its mean even in the $M/G/s$ case. As in §2.7, the general model is more complicated, so that we do not intend to give a detailed supporting argument. Our main idea is to apply a heavy-traffic diffusion approximation for the $G/G/s$ queue to approximately characterize the conditional waiting-time distribution given that a customer must wait before beginning service; see Iglehart and Whitt (1970), Newell (1982), Whitt (1982) and Heyman and Sobel (1982) for more discussion. This leads to $(W \mid W > 0)$ being distributed nearly the same as in a $G/G/1$ system with the same arrival process and service-time distribution, except that the service rate is s instead of 1. Paralleling the exact results for $M/M/s$ queues, this is a natural direct rough approximation.

Furthermore, the heavy-traffic analysis leads to $(W \mid W > 0)$ having an exponential distribution. Thus all that remains is to specify the mean. The heavy-traffic analysis indicates that the mean should be proportional to the asymptotic value of the index of dispersion for work (IDW), $I_w(\infty)$, discussed in Fendick and Whitt (1989). In particular, if $X(t)$ represents the total work in service time to arrive in the time interval $(0, t]$, then the IDW is the function

$$I_w(t) = \frac{\text{Var } X(t)}{\tau EX(t)}, \qquad t > 0, \tag{18}$$

where τ is the mean service time. The asymptotic value $I_w(\infty)$ is the limit as $t \to \infty$. For the $G/D/s$ model, $I_w(\infty) = c_a^2$ for c_a^2 defined in (10). For the $GI/G/s$ special case (the interarrival times and service times come from independent sequences of i.i.d. random variables), $I_w(\infty) = c_a^2 + c_s^2$ where c_a^2 and c_s^2 are the SCVs of an interarrival time and a service time. The asymptotic IDW value $I_w(\infty)$ is useful when there is more complicated dependence among the interarrival times and service times; e.g., see Fendick, Saksena and Whitt (1989). Since the sequence of successive waiting times in the $G/G/1$ model can be regarded as a random walk with possibly dependent steps and an impenetrable barrier at the origin, the heavy-traffic diffusion approximation can be regarded as yet another application of the central limit theorem (in addition to (7), (9), (11) and Propositions 2.2 and 2.5); see Section 10 of Billingsley (1968) and Iglehart and Whitt (1970).

Given (18), the heavy-traffic analysis suggests the approximations

$$P(W > x \mid W > 0) \approx e^{-\beta\sqrt{s}} \quad \text{and} \tag{19}$$

$$E[W \mid W > 0] \approx 1/\beta\sqrt{s} \tag{20}$$

when (1) holds, where

$$\beta = 2\gamma/I_w(\infty). \tag{21}$$

TABLE 10

The Normalized Conditional Mean Waiting Times from Tables 2–7 Assuming (1) *Holds for the Case* $s = 25$, *Supporting Approximation* (20) *and* (21)

Table	c_a^2	c_s^2	$\frac{c_a^2 + c_s^2}{2}$	$\left(\frac{2\gamma\sqrt{s}}{c_a^2 + c_s^2}\right)E[W \mid W > 0]$			
				$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.5$	$\gamma = 1$
2	1.00	1.00	1.000	1.00	1.00	1.00	1.00
3	0.25	1.00	0.625	1.00	1.01	1.02	1.04
4	4.00	1.00	2.500	0.99	0.98	0.94	0.86
5	1.00	0.00	0.500	1.02	1 04	1.12	1.26
6	0.25	0.00	0.125	1.05	1.09	1.28	1.52
7	4.00	0.00	2.000	1.00	0.99	0.98	0.94
	1.00	2.50	1.750	0.98	0.96	0.89	0.80
	0.25	2.50	1.375	0.98	0.96	0.89	0.80
	4.00	2 50	3.250	0.98	0.96	0.91	0.81

Note that (19) and (20) agree with (16) and (17) for the $M/M/s$ queue. For other $GI/G/s$ queues, approximation (20) with (21) is supported by the numerical results in Tables 2–7.

Tables 2–8 show that $\sqrt{s}E[W \mid W > 0]$ tends to be proportional to $(c_a^2 + c_s^2)/2\gamma$ as suggested in (20) and (21). (Recall that $I_w(\infty) = c_a^2 + c_s^2$ for a $GI/G/s$ model.) It is easy to see that $(c_a^2 + c_s^2)/2$ is more appropriate than the peakedness z in (13) as a variability factor in $E[W \mid W > 0]$ by comparing the $M/D/s$ and $M/M/s$ cases in Tables 2 and 3. For both these cases the peakedness is $z = 1$, but $(c_a^2 + c_s^2)/2$ is 1 for $M/M/s$ and $\frac{1}{2}$ for $M/D/s$.

Table 10 displays normalized conditional mean waiting times from Tables 2–7 for the cases in which $s = 25$. Also included are three cases in which $c_s^2 = 2.5$ to give a wider range of variability parameters. Consistent with approximations (20) and (21), these values are all nearly 1.00 for small γ, but the quality of the approximation tends to deteriorate as γ increases.

Finally, combining (14) and (20), we obtain a rough approximation for the mean when (12) holds with $\gamma < 1$ that reflects the variability, i.e.,

$$EW = P(W > 0)E[W \mid W > 0] = \left(1 - \frac{(1-\rho)\sqrt{s}}{\sqrt{z}}\right)\left(\frac{I_w(\infty)}{2(1-\rho)s}\right). \tag{22}$$

We believe that (22) is the first approximation for $GI/G/s$ queues to use *both* the peakedness z and the asymptotic value of the IDW, $I_w(\infty)$ (or the squared coefficients of variation c_a^2 and c_s^2 in the $GI/G/s$ case). For an $M/M/s$ queue, (22) reduces to

$$EW = \frac{(1 - (1-\rho)\sqrt{s})}{(1-\rho)s} = \frac{1-\gamma}{\gamma\sqrt{s}}, \tag{23}$$

assuming that (1) holds.

3.3. *Implications*

Formulas (20)–(21) indicate that formulas (1)–(4) actually tend to provide lower bounds on the economy of scale. If the average waiting time or a delay percentile is the criterion, then the utilization can increase as a function of s somewhat faster than indicated by (1) or (4). Indeed, if we are only concerned with the conditional waiting time given that a customer must wait before beginning service, then our analysis indicates that (1) should be replaced by

$$(1 - \rho)s = \gamma' \tag{24}$$

for some constant γ' which is equivalent to $s - \lambda = \gamma'$. Our numerical analysis confirms that (1) and (4) are quite accurate when the criterion is $P(W > 0)$, and (24) is quite accurate when the criterion is $E[W \mid W > 0]$. When the criterion is EW, the average of the utilization estimates provided by (1) and (24) is a reasonable rough estimate. For EW, a more accurate utilization equation might be $(1 - \rho)s^{3/4} = \gamma''$ for some constant γ''.

Formulas (4), (19) and (20) indicate that (1) should provide a lower bound on the way ρ should grow with s to maintain a fixed grade of service, with almost any waiting-time criterion. To see this, let the criterion be the mean EW and consider the case $\gamma = 0.5$ in Table 3. Suppose that we start with $s = 4$ and change to $s = 25$. For $s = 4$ and $\gamma = 0.5$, $\rho = 0.75$ and $EW = 0.51$; for $s = 25$, (1) yields $\rho = 0.90$ and $EW = 0.21$. From the tables, $EW = 0.30$ and 0.59 for $\rho = 0.92$ and 0.94, so that $EW = 0.51$ for $s = 25$ when ρ is between 0.93 and 0.94. If we use (24) instead of (1), which is exact for the conditional mean $E[W \mid W > 0]$, then we obtain $\gamma' = 1.0$ when $\rho = 0.75$ for $s = 4$. We also obtain $\rho = 0.96$ for $s = 25$ using (24). The appropriate utilization for $s = 25$ with the criterion EW is thus about halfway between what is suggested by (1) and (24).

Now suppose that we consider changing from $s = 25$ to $s = 100$ with $\gamma = 0.5$. For $\gamma = 0.5$ and $s = 25$, $\rho = 0.90$ and $EW = 0.21$; for $s = 100$, (1) yields $\rho = 0.95$ and $EW = 0.10$. From the tables, $EW = 0.19$ and 0.36 for $\rho = 0.96$ and 0.97. Hence, (1) predicts that we should go from $\rho = 0.90$ to $\rho = 0.95$ when we go from $s = 25$ to $s = 100$, but with the criterion of the mean the actual utilization should be about $\rho = 0.96$ when $s = 100$. If we use (24) instead of (1), then we obtain $\gamma' = 2.5$ when $\rho = 0.90$ and $s = 25$. Hence, (24) dictates $\rho = 0.975$ for $s = 100$. Again we see that the appropriate value of 0.96 is about half way between what is suggested by (1) and (24).

4. Conclusions

The approximation strategy here has been to approximate the behavior of the complicated s-server model by two more elementary models. We approximate the behavior when all s servers are not busy by the associated infinite-server model with the same arrival process and the same service-time distribution. We approximate the behavior when all s servers are busy by the associated single-server model with the same arrival process and a scaled version of the same service time distribution, i.e., each service time is divided by s. We then use established heavy-traffic approximations for the two more elementary models in the two regions. These heavy-traffic limit theorems essentially are applications of the central limit theorem.

The approximation formulas here are intended only as simple rough approximations. They are intended to quickly provide understanding, so that we can sensibly think about alternative designs and operating policies. Queueing formulas and tables are available for more precise results. There also are more accurate (and more involved) approximations intended for greater numerical accuracy, e.g., Whitt (1985). Nevertheless, the simple approximations seem remarkably accurate. However, the analysis and the numerical examples indicate that the quality of the approximations deteriorates as s and λ decrease and as the grade of service γ and the variability parameters z in (13) and $I_w(\infty)$ in (18) increase.

It is significant that the proper relation between the server utilization and the number of servers depends on the performance measure of interest. The simple formulas (1), (2) and (12) apply to the probability of delay, whereas the very different simple formula (24) applies to the conditional distribution $P(W > t \mid W > 0)$. The situation gets much more complicated when we consider a performance measure such as the mean EW,

which involves both. However, (1) and (24) seem to provide quick rough lower and upper bounds for what happens with the mean EW.[1]

[1] I am grateful to my colleague Bill Kahan for discussions that motivated this paper and to a referee for excellent suggestions about the presentation.

References

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

BOROVKOV, A. A., "On Limit Laws for Service Processes in Multi-Channel Systems," *Siberian Math. J.*, 8 (1967), 746–763.

ECKBERG, A. E., "Generalized Peakedness of Teletraffic Processes," *Proc. Tenth Internat. Teletraffic Congress*, Montreal, Canada, p. 4.4b.3, June 1983.

———, "Approximations for Bursty (and Smoothed) Arrival Queueing Delays Based on Generalized Peakedness," *Proc. Eleventh Internat. Teletraffic Congress*, Kyoto, Japan, 1985.

FELLER, W. *An Introduction to Probability Theory and Its Applications*, Vol. I, (3rd Ed.), Wiley, New York, 1968.

FENDICK, K. W., V. R. SAKSENA AND W. WHITT, "Dependence in Packet Queues," *IEEE Trans. Commun.*, 37 (1989), 1173–1183.

——— AND W. WHITT, "Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single Server Queue," *Proc. IEEE*, 77 (1989), 171–194.

FREDERICKS, A. A., "Congestion in Blocking Systems—A Simple Approximation Technique," *Bell System Tech. J.*, 59 (1980), 805–827.

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

——— AND ———, "A New View of the Heavy-Traffic Limit Theorem for Infinite-Server Queues," *Adv. in Appl. Prob.*, 23 (1991), 188–209.

GRASSMAN, W. K., "Is the Fact That the Emperor Wears No Clothes a Subject Worthy of Publication," *Interfaces*, 16 (1986), 43–51.

———, "Finding the Right Number of Servers in Real-World Queueing Systems," *Interfaces*, 18 (1988), 94–104.

HALFIN, S. AND W. WHITT, "Heavy-Traffic Limits for Queues with Many Exponential Servers," *Oper. Res.*, 29 (1981), 567–588.

HEYMAN, D. P. AND M. J. SOBEL, *Stochastic Models in Operations Research*. Vol. I, McGraw-Hill, New York, 1982.

HOKSTAD, P., "Approximations for the $M/G/m$ Queue," *Oper. Res.*, 26 (1978), 510–523.

HUNT, P. J. AND F. P. KELLY, "On Critically Loaded Loss Networks," *Adv. in Appl. Prob.*, 21 (1989) 831–841.

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

JAGERMAN, D. L., "Some Properties of the Erlang Loss Function," *Bell System Tech. J.*, 53 (1974), 525–551.

KOLESAR, P. J., "Comment to 'Is the Fact That the Emperor Wears No Clothes a Subject Worthy of Publication?'" *Interfaces*, 16 (1986), 50–51.

KÜHN, P., *Tables on Delay Systems*, Institute of Switching and Data Technics, University of Stuttgart, 1976.

MCKENNA, J., D. MITRA AND K. G. RAMAKRISHNAN, "A Class of Closed Markovian Queueing Networks: Integral Representations, Asymptotic Expansions, Generalizations," *Bell System Tech. J.*, 60 (1981), 599–641.

NEWELL, G. F., "Approximate Stochastic Behavior of n-Server Service Systems with Large n," *Lecture Notes in Economics and Math. Systems*, 87, Springer-Verlag, New York, 1973.

———, *Applications of Queueing Theory*, (Second ed.), Chapman and Hall, London, 1982.

REIMAN, M. I., "Asymptotically Optimal Trunk Reservation for Large Trunk Groups," *Proc.* 28th *IEEE Conf. Decision and Control*, 1989, 2536–2541.

———, "Some Allocation Problems for Critically Loaded Loss Systems with Independent Links," *Proc. Performance* '90, Edinburgh, Scotland, 1990, 145–158.

SEELEN, L. P., H. C. TIJMS AND M. H. VAN HOORN, *Tables for Multi-Server Queues*, North-Holland, Amsterdam, 1985.

SMITH, D. R. AND W. WHITT, "Resource Sharing for Efficiency in Traffic Systems," *Bell System Tech. J.*, 60 (1981), 39–55.

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

WHITT, W., "Refining Diffusion Approximations for Queues," *Oper. Res. Lett.*, 1 (1982), 165–169.

———, "Heavy-Traffic Approximations for Service Systems with Blocking," *AT&T Bell Lab. Tech. J.*, 63 (1984), 689–708.

———, "Approximations for $GI/G/m$ Queues," AT&T Bell Laboratories, 1985.