AT&T Bell Laboratories Technical Journal
Vol. 63, No. 5, May–June 1984
Printed in U.S.A.

Heavy-Traffic Approximations for Service Systems With Blocking

By W. WHITT*

(Manuscript received October 27,1983)

This paper develops approximations for the blocking probability and related congestion measures in service systems with s servers, r extra waiting spaces, blocked customers lost, and independent and identically distributed service times that are independent of a general stationary arrival process (the G/GI/s/r model). The approximations are expressed in terms of the normal distribution and the peakedness of the arrival process. They are obtained by applying previous heavy-traffic limit theorems and a conditioning heuristic. There are interesting connections to Hayward's approximation, generalized peakedness, asymptotic expansions for the Erlang loss function, the normal-distribution method, and bounds for the blocking probability. For the case of no extra waiting space, a renewal arrival process and an exponential service-time distribution (the GI/M/s/0 model), a heavy-traffic local limit theorem by A. A. Borovkov implies that the blocking depends on the arrival process only through the first two moments of the renewal interval as the offered load increases. Moreover, in this situation Hayward's approximation is asymptotically correct.

I. INTRODUCTION AND SUMMARY

In this paper we introduce and investigate approximations for congestion measures in G/GI/s/r service systems, which have s servers, r extra waiting spaces, the first-come first-served discipline, and independent and identically distributed (i.i.d.) service times with a general distribution that are independent of a general stationary

*AT&T Bell Laboratories.

arrival process. Customers arriving when all s servers are busy and all r waiting spaces are full are blocked; they leave without receiving service and without affecting future arrivals (no retrials). We primarily focus on the case $r = 0$ (except for Section VII). We present approximate expressions for the proportion of arriving customers that are blocked (call congestion) and the proportion of time that the system is full (time congestion). We also approximate the distributions of the number of customers in the system at arrival epochs and at arbitrary times.

We obtain our approximations by applying previous heavy-traffic limit theorems[1-4] and a conditioning heuristic (Section III). As with much of the earlier work on this problem, we are not able to present a completely rigorous development, but we believe that we have a novel perspective that provides additional insight. There are interesting connections to earlier work, including Hayward's approximation,[5] generalized peakedness,[6] asymptotic expansions for the Erlang loss function,[7-9] the normal distribution method,[10-14] and bounds for the blocking probability.[15-17]

Perhaps our most important contribution is to point out the significance of a heavy-traffic local limit theorem by Borovkov [Theorem 15(2) of Ref. 2], which was first published in Russian in 1972. ("Local" means that the limit is for the probability mass function instead of the cumulative distribution function.) For GI/M/s/0 models (no extra waiting space, renewal arrival process, and exponential service-time distribution), this theorem provides a rigorous basis for the approximations under heavy loads. For example, this theorem implies that Hayward's approximation [(22) in Section 6.2] is asymptotically correct as the offered load increases. Of course, this property is consistent with extensive numerical evidence, but apparently no mathematical proof has been given before.

Here is how this paper is organized. In Section II we review a heavy-traffic limit theorem for G/GI/∞ models that we will apply, which is also due to Borovkov.[1] In Section III we introduce a conditioning heuristic and apply it with the limit theorem in Section II to obtain an approximation for the distribution of the number of busy servers at an arbitrary time in the associated G/GI/s/0 system. In Section IV we use a conservation law plus the approximation in Section III to generate an approximation for the blocking probability in G/GI/s/0 systems. We also discuss an approximation for the distribution of the number of busy servers at arrival epochs. In Section V we state Borovkov's local heavy-traffic limit theorem for GI/M/s/0 models that supports the approximations. We also make several conjectures about related theorems.

In Section VI we discuss connections to other work. We indicate that the normal approximation for the blocking probability in M/M/s/0 systems has a long history, going all the way back to Erlang.[18] In the Appendix we also give a simple proof of the heavy-traffic local limit theorem for M/M/s/0 systems, using the elementary central limit theorem and Stirling's formula.[19] In Section VI we also discuss connections to Hayward's approximation,[5,6] bounds for the blocking probability,[15-17] and previous normal approximations.[10-14]

In Section VII we indicate how the approach can be extended to systems with finite waiting rooms, drawing on Halfin and Whitt.[4] In doing so, we obtain a modification of Hayward's approximation for the case of a finite waiting room [see (42)]. Finally, in Section VIII we give the results of experiments testing the approximations for G/M/s/0 systems. As observed by Rahko,[10-12] Hertzberg,[13] and Delbrouck,[14] the normal approximation tends to work quite well except in low loads.

We close this introduction by noting that the general blocking problem discussed here continues to generate considerable attention; several related papers were presented at the Tenth International Teletraffic Congress at Montreal.[6,12,20-24] Another recent related contribution is Newell.[25]

II. THE INFINITE-SERVER MODEL IN HEAVY TRAFFIC

Consider the G/GI/∞ service system, which has infinitely many servers and independent and identically distributed service times that are independent of a general stationary arrival process. Let $A(t)$ count the number of arrivals in the interval $[0, t]$ for $t \geq 0$. We assume that $A(t)$ satisfies a central limit theorem; i.e.,

$$[A(t) - \lambda t]/(\lambda c^2 t)^{1/2} \rightarrow N(0, 1) \tag{1}$$

as $t \rightarrow \infty$, where $\rightarrow$ denotes convergence in distribution, $N(a, b)$ denotes a random variable with the normal distribution having mean a and variance b, and λ is the arrival rate. When $A(t)$ is a renewal process, c^2 in (1) is the squared coefficient of variation (variance divided by the square of the mean) of the renewal interval. More generally, the denominator in (1) typically is asymptotically equivalent to the standard deviation of $A(t)$, so that

$$c^2 = \lim_{t\rightarrow\infty} \operatorname{var}[A(t)]/\lambda t = \lim_{t\rightarrow\infty} \operatorname{var}[A(t)]/EA(t). \tag{2}$$

The parameter c^2 in (1) and (2) is the basis for approximating the arrival process by a renewal process via the asymptotic method in Ref. 26.

Let μ^{-1} be the mean and $G(t)$ the cumulative distribution function (cdf) of the service-time distribution. Let $\alpha = \lambda/\mu$ be the offered load.

Our approximations are developed by considering limits as the offered load α increases. We fix the service-time cdf $G(t)$ and change α by changing λ.

Let X_α be the equilibrium number of busy servers in the G/GI/∞ system at an arbitrary time, as a function of α. (We assume that a unique equilibrium distribution exists—see Section 2.3 of Franken et al.[27]) Borovkov[1] proved the following heavy-traffic limit theorem for X_α. He showed that if (1) holds [actually a slightly stronger functional limit theorem for $A(t)$], then

$$(X_\alpha - \alpha)/\sqrt{\alpha z} \rightarrow N(0, 1) \tag{3}$$

as $\alpha \longrightarrow \infty$, where

$$z = 1 + (c^2 - 1)\eta \tag{4}$$

and

$$\eta = \mu \int_0^\infty [1 - G(t)]^2 dt; \tag{5}$$

also see Refs. 3 and 6. Actually, Borovkov did not directly treat the equilibrium variable X_α, so that there is a further interchange of limits to get (3). For practical purposes, we can regard (3) as a consequence of (1).

To interpret (3) through (5), recall that $EX_\alpha = \alpha$ for all α, even with a general stationary arrival process. The parameter z in (4) is the asymptotic peakedness of the arrival process $A(t)$ with respect to the service-time cdf $G(t)$ because

$$z = \lim_{\alpha \to \infty} \operatorname{var}(X_\alpha)/EX_\alpha = \lim_{\alpha \to \infty} \operatorname{var}(X_\alpha)/\alpha; \tag{6}$$

see Ref. 6 and references there. In particular, z in (4) is $z_G(0+)$ in Section 3.3.1 of Ref. 6. Note that η in (5) has the maximum value of 1 attained by a deterministic service-time distribution (unit mass on μ^{-1}) and can assume any value in the interval (0, 1]. For example, $\eta =$ 2/3, 1/2, and $1/n$, respectively, when the distribution is uniform, exponential and concentrated on two points with mass n^{-1} on 0.

We have defined three parameters characterizing variability: c^2, η, and z. The parameter c^2 in (1) and (2) is a measure of the variability of the arrival process (over large time intervals). The parameter η in (5) is a measure of the variability of the service-time distribution and the parameter z in (4) is a second measure of variability of the arrival process (as measured by the G/GI/∞ model with service-time cdf G having variability parameter η in heavy traffic). The heavy-traffic peakedness z in (4) is increasing in c^2, but whether it is increasing or decreasing in η depends on the sign of $c^2 - 1$. For a Poisson process,

$c^2 = 1$. In general, increased service-time variability as expressed by decreasing η in (5) tends to make z in (4) closer to one. Hence, if the arrival process is more variable or bursty than a Poisson process in the sense that $c^2 > 1$, then z in (4) is increasing in η, which means that the peakedness increases as the variability of the service-time distribution decreases. When $c^2 > 1$, the deterministic service-time cdf gives the largest heavy-traffic peakedness among all service-time cdf's with the same mean. This phenomenon seems to have been first discussed by Wolff[28] (see also Section 3.3.1 of Ref. 6). Related results about GI/G/1 queues are contained in Whitt.[29]

III. THE CONDITIONING HEURISTIC

We now use the heavy-traffic limit theorem (3) for the G/GI/∞ system to produce approximations for the associated G/GI/s/0 loss system, which has s servers, no extra waiting space, and the same arrival process and service-time distribution. Our starting point is a basic property of the Markovian M/M/s/0 models in which the number of customers in the system evolves as a birth-and-death process. The equilibrium probability $p_k(s_1)$ that there are k customers in an M/M/s_1/0 model is just the conditional probability that there are k customers in an M/M/s_2/0 model given that there are no more than s_1 customers, for any s_2 with $s_1 \le s_2 \le \infty$; in other words, $p_k(s_1)$ is obtained by truncating and renormalizing $p_k(s_2)$:

$$p_k(s_1) = p_k(s_2) \Big/ \sum_{j=0}^{s_1} p_j(s_2) \tag{7}$$

for $0 \le k \le s_1 \le s_2$. Truncation formula (7) is an immediate consequence of the known formula for $p_k(s)$, but also is easily derived for more general birth-and-death processes (see p. 68 of Heyman and Sobel).[30]

Let Y_α be the equilibrium number of busy servers at an arbitrary time in the G/GI/s/0 model. (We also assume existence and uniqueness—see Section 2.3.2 of Ref. 27.) As an approximation, we assume that Y_α is related to X_α of Section II by the same conditioning formula (7). In particular, we suggest the heuristic approximation

$$P(Y_\alpha = k) \approx P(X_\alpha = k)/P(X_\alpha \le s) \tag{8}$$

for $0 \le k \le s$. This conditioning approximation is no doubt an old idea, which would be hard to trace; it was used previously by Jagerman to develop approximate blocking formulas in the case of nonstationary Poisson traffic.[31]

Next we obtain a further approximation by invoking the heavy-traffic limit theorem for X_α in (3). For this purpose, let $\Phi(t)$ be the standard normal cdf, i.e., of $N(0, 1)$, and $\phi(t)$ the associated density. We combine (3) and (8) to obtain

$$\begin{aligned} P(Y_\alpha = k) &\approx \frac{P[k - 0.5 \le N(\alpha, \alpha z) \le k + 0.5]}{P[N(\alpha, \alpha z) \le s + 0.5]} \\ &\approx \frac{\Phi\left(\dfrac{k - \alpha + 0.5}{\sqrt{\alpha z}}\right) - \Phi\left(\dfrac{k - \alpha - 0.5}{\sqrt{\alpha z}}\right)}{\Phi\left(\dfrac{s - \alpha + 0.5}{\sqrt{\alpha z}}\right)} \\ &\approx (1/\sqrt{\alpha z})\phi\left(\frac{k - \alpha}{\sqrt{\alpha z}}\right) \Big/ \phi\left(\frac{s - \alpha + 0.5}{\sqrt{\alpha z}}\right). \end{aligned} \tag{9}$$

The time congestion, say B_T, is defined as $B_T = P(Y_\alpha = s)$, so that an approximation for it is obtained from (9) simply by setting $k = s$.

In further support of the conditioning heuristic, we note that it is also valid for the diffusion process limits that arise in several cases of heavy traffic, e.g., for the stochastic-process version of (3) in the case of exponential service-time distributions (see Ref. 1 and 3).

We have applied the conditioning heuristic to the distributions at an arbitrary time instead of at arrival epochs. Since Poisson arrivals see time averages,[27,30,32] these two distributions are the same for M/M/s/0 systems. Also, the heavy-traffic limits for these distributions are the same for G/GI/∞ system. (See Ref. 3 for the renewal arrival process case.) However, these distributions are definitely not the same for the G/GI/s/r system or even the GI/M/s/0 special case. The conditioning heuristic seems to perform much better when applied to the distribution at an arbitrary time, as will be clear from the next two sections. This might not be surprising, but a good explanation is still needed.

IV. A CONSERVATION LAW AND THE BLOCKING APPROXIMATION

Let B_C be the probability that an arriving customer in the G/GI/s/0 system is blocked (call congestion). A basic conservation law enables us to express B_C in terms of EY_α. In particular, since the average rate of accepted arrivals equals the average departure rate, not counting lost calls (see Heyman[17]),

$$\lambda(1 - B_C) = \mu E Y_\alpha. \tag{10}$$

Hence, we can combine (9) and (10) to obtain an approximation for B_C.

It is well known and easy to show that

$$E[N(0, 1) \,|\, N(0, 1) \le \theta] = -\phi(\theta)/\Phi(\theta), \tag{11}$$

so that

$$EY_\alpha \approx \alpha - \sqrt{\alpha z}\phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right)\bigg/\Phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right) \tag{12}$$

and

$$B_C = 1 - \alpha^{-1}EY_\alpha \approx \sqrt{z/\alpha}\phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right)\bigg/\Phi\left(\frac{s-\alpha}{\sqrt{\alpha z}}\right) \approx zB_T \tag{13}$$

(e.g., see Appendix A of Delbrouck[14]).

We thus suggest (9) as an approximation for $P(Y_\alpha = k)$, (9) with $k = s$ for an approximation for B_T, and (13) as an approximation for B_C.

Let Z_α be the equilibrium number of busy servers in the G/GI/s/0 system at arrival epochs, again as a function of the offered load α. For G/M/s/0 systems (exponential service-time distributions, but still general stationary arrival process), we have the exact relationship

$$P(Z_\alpha = k - 1) = kP(Y_\alpha = k)/\alpha \tag{14}$$

for $1 \leq k \leq s$ (p. 113 of Franken et al.[27]), which is a refinement of (10), because (10) is obtained from (14) by simply summing over k. We thus propose (14) as an approximation for G/M/s/0 systems and also G/GI/s/0 systems. Improved approximations for the nonexponential service-time distribution case can perhaps be obtained from the relationships in Section 4.3.2 of Ref. 27, but this does not appear easy.

These approximations are expressed in terms of the asymptotic peakedness z in (4). However, if this parameter is not available, other expressions for the peakedness can be used instead.[6] For example, the general formula for the peakedness of a renewal arrival process with respect to an exponential service-time distribution is given here in (27).

Formulas (4) and (5) can also be used to calculate a revised peakedness if there is a change in the service-time distribution, as suggested in Section 6 of Ref. 33 and as has been done in practice. Suppose that z_1 has been previously determined based on the service-time cdf G_1, but now we are going to consider the same G/GI/s/0 system with new service-time cdf G_2. For this purpose let $\eta_i = \eta(G_i)$ in (5). Using (4), we obtain an approximation for c^2, namely,

$$c^2 = 1 + (z_1 - 1)/\eta_1\,. \tag{15}$$

We can thus approximate z_2 based on (4), (5) and

$$z_2 = 1 + (c^2 - 1)\eta_2 = 1 + (z_1 - 1)\eta_2/\eta_1. \tag{16}$$

More general transformations based on Mellin transforms have also been developed by Jagerman for the entire peakedness functionals

$z[F]$ considered in Ref. 6 to describe the effect of changing the service-time distribution.[34]

Approximation formulas (9), (13), and (14) can also be used to generate approximate multiplicative correction factors to be used with the exact M/M/s/0 formulas. For example, we can obtain our approximation by multiplying the exact blocking probability for the M/M/s/0 system by the ratio $B_C(z)/B_C(1)$, where $B_C(z)$ is B_C in (13) as a function of z in (4). This procedure is slightly more complicated, but it is exact for M/M/s/0 systems.

V. BOROVKOV'S HEAVY-TRAFFIC LOCAL LIMIT THEOREM

A rigorous justification of (13) is provided by Theorem 15 (2), p. 226, of Borovkov.[2] For the GI/M/s/0 model, Borovkov established that

$$\lim_{\alpha\to\infty} \sqrt{\alpha} B_C = \sqrt{z}\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{17}$$

if $(s-\alpha)/\sqrt{\alpha} \to \beta$ as $\alpha \to \infty$, where $z = (c^2+1)/2$ as specified in (4) in the case of an exponential service-time distribution.

Borovkov identifies z in (17) as $(c^2+1)/2$ rather than the heavy-traffic peakedness in (4), but we believe (4) is appropriate for generalizations to nonexponential service-time distributions and nonrenewal arrival processes.

Borovkov's arguments can also be applied to yield a related local limit theorem for $P(Z_\alpha = k)$ in GI/M/s/0 systems, namely,

$$\lim_{\alpha\to\infty} \sqrt{\alpha} P(Z_\alpha = k) = \sqrt{z}\phi(\beta'/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{18}$$

if $(s-\alpha)/\sqrt{\alpha} \to \beta$ and $(k-\alpha)/\sqrt{\alpha} \to \beta' < \beta$ as $\alpha \to \infty$. We can then apply (14) to deduce that

$$\lim_{\alpha\to\infty} \sqrt{\alpha} P(Y_\alpha = k) = (1/\sqrt{\alpha z})\phi(\beta'/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{19}$$

and

$$\lim_{\alpha\to\infty} \sqrt{\alpha} B_T = (1/\sqrt{\alpha z})\phi(\beta/\sqrt{z})/\Phi(\beta/\sqrt{z}) \tag{20}$$

under the same asymptotic conditions. The limit in (19) coincides with Theorem 16, p. 232, of Borovkov,[2] which is stated there without proof.

Hence, we have theoretical justification for all our approximations in the case of GI/M/s/0 systems. We conjecture that (17) through (20) are still valid for general stationary arrival processes and nonexponential service-time distributions, i.e., under the conditions for Borovkov's G/GI/∞ theorem (3)'. Since (14) is valid for general nonrenewal arrival processes but not for general service-time distributions, the conjecture seems much more likely to be true for the generalization of the arrival process than for the generalization of the service-time distribution.

VI. CONNECTIONS TO OTHER WORK

6.1 Beginning with Erlang

For the Markovian M/M/s/0 model as well as the M/G/s/0 model, the exact blocking probability is given by the classic Erlang loss (B) formula

$$B(s, \alpha) = (\alpha^s/s!) \Big/ \sum_{k=0}^{s} (\alpha^k/k!). \tag{21}$$

Since Poisson arrivals see time averages,[27,30,32] $B_C = B_T$ and (18) and (19) coincide in this case. In this case, approximation (9) reduces to a rather standard normal approximation which was evidently known in 1924 by Erlang.[18] The M/M/s/0 case of the limit theorem (17) plus various asymptotic expansion improvements were given by Brockmeyer[7] and Vaulot.[8] Asymptotic expansion improvements also follow from Theorem 14 of Jagerman[9] and are discussed on p. 88 of Delbrouck.[14] A simple proof of the M/M/s/0 version of (17) using the central limit theorem and Stirling's formula[19] is given here in the Appendix.

6.2 The Hayward approximation

A relatively simple approximation for the blocking probability B_C in G/M/s/0 and even G/G/s/0 systems as a function of s, α and z proposed by Hayward is

$$B_C \equiv B_C(s, \alpha, z) \approx B_C(s/z, \alpha/z, 1) = B(s/z, \alpha/z) \tag{22}$$

using (21) (see Fredericks[5] and Eckberg[6]). It is significant that the approximation (9) and the limit theorem (17) are consistent with (22); for those expressions, $B(s, \alpha, z) \approx B(s/z, \alpha/z, 1)$.

There are several possible interpretations and applications. We can interpret (17) through (20) as additional evidence in support of the Hayward approximation. The limit theorems and approximations provide additional evidence that the Hayward approximation should perform well under heavy loads. The Hayward approximation is particularly appealing, given that there are convenient computer programs to calculate $B(s, \alpha)$ in (21) extended to nonintegral s, as have been developed by Jagerman.[35]

On the other hand, we can use the Hayward approximation as an additional justification for (9). We can derive (9) by applying the limit theorem for M/M/s/0 systems described in Section 6.1 together with the Hayward approximation.

The connection also suggests that improved approximations can be obtained for G/M/s/0 systems and possibly G/G/s/0 systems by applying the asymptotic expansions of Brockmeyer,[7] Vaulot,[8] or Jagerman[9] for $B(s, \alpha)$ after applying the Hayward approximation to treat the peakedness z. In particular, formula (5) in Jagerman,[9] which is based on Theorem 14 there, seems particularly promising in combination with Hayward's approximation. However, testing remains to be done. Of course, improved approximations would also be obtained from asymptotic expansions related to (17). This seems to be a promising direction of research.

6.3 Bounds for the blocking probability

Sobel[16] and Heyman[17] recently established a lower bound for the blocking probability in a G/G/s/r system when $\rho > 1$, namely,

$$B_C \geq 1 - \rho^{-1}, \tag{23}$$

and observed that the lower bound often is a good approximation when $\rho > 1$.

We partly explain why the lower bound is a good approximation by showing that it appears in limiting lower and upper bounds as $\alpha \to \infty$ in our heavy-traffic approximation (9) for G/G/s/0 systems. (This paper is a revised version of Ref. 11 in Sobel,[16] where our result is mentioned.)

We use the familiar bounds for the tail of a normal distribution

$$(x^{-1} - x^{-3})\phi(x) < 1 - \Phi(x) = \Phi(-x) < x^{-1}\phi(x), \qquad x > 0; \tag{24}$$

see p. 175 of Feller.[19] To make the connection to (9) and (17), let $x = (\alpha - s)/\sqrt{\alpha z}$. Since $x\sqrt{z/\alpha} = (1 - \rho^{-1})$, and

$$1 \leq \frac{x\phi(x)}{\Phi(-x)} \leq (1 - x^{-2})^{-1} \tag{25}$$

by (24), from (9) we obtain the approximate bounds as $x \to \infty$

$$(1 - \rho^{-1}) \leq B_C \leq (1 - \rho^{-1})(1 - x^{-2})^{-1}, \tag{26}$$

which are useful for $\rho \geq 1$. The distance between the bounds goes to zero as $x \to \infty$.

Holtzman also establishes bounds for B_C in the GI/M/s/0 system.[15] In fact, he described the range of all possible values of B_C given only the offered load α and the peakedness z' of a renewal process, which is

$$z' = [1 - \phi(\mu)]^{-1} - \alpha, \tag{27}$$

where

$$\phi(\mu) = \int_0^\infty e^{-\mu t} dF(t) \tag{28}$$

with $F(t)$ the cdf of an interarrival time. By (17), we know that the range approaches 0 as $\alpha \to \infty$ with $(s - \alpha)/\sqrt{\alpha} \to \beta$. The z' in (27) approaches z in (4) and, by (9) and (17), B_C depends only on α, s and z for large α.

6.4 *The normal-distribution method*

The approximations here for $P(Y_\alpha = k)$, B_T, and B_C in (9), (13), and (14) coincide with the normal-distribution method (NDM) of Rahko[10-12] and Hertzberg[13] and the normal approximation for the Bernoulli-Poisson-Pascal (BPP) approximation of Delbrouck,[14] but the analysis here is different.

6.5 *Light-traffic approximations and interpolations*

The approximations here have been developed by considering the service systems in heavy traffic, i.e., as the offered load increases. Improved approximations for lighter loads may be possible by considering the service systems in light traffic, i.e., as the offered load decreases. Better approximations might be obtained by making interpolations between light and heavy traffic. This seems to be another promising direction for future research. Previous work on light-traffic approximations for queues is contained in Bloomfield and Cox,[36] Newell,[37] and Burman and Smith.[38,39] Interpolations between light and heavy traffic have been considered by Burman and Smith[39] and Reiman and Simon.[40] The hybrid approximations for queues with superposition arrival processes developed by Albin[41] and used in Ref. 42 to approximate networks of queues are also in this spirit.

VII. FINITE WAITING ROOMS

Corresponding approximations can be developed for G/GI/s/r systems with r extra waiting spaces. We can apply heavy-traffic limit theorems for G/GI/s/∞ systems (see Ref. 4), together with the conditioning heuristic of Section III. The conditioning relationship (7) is also valid for M/M/s/r systems with different values of r. As Ref. 4 describes, there are several possible heavy-traffic limit theorems to apply. For GI/M/s/∞ systems, we suggest the heavy-traffic limit theorems in Ref. 4 with $(s - \alpha)/\sqrt{\alpha} \to \beta$ as $\alpha \to \infty$ or, equivalently, $(1 - \rho)\sqrt{s} \to \beta$, where $\alpha = \lambda/\mu$ and $\rho = \alpha/s$. This leads to a promising analog of Hayward's approximation (22) for the case of a finite waiting room, namely, (42) below.

From Section 4 of Ref. 4, it is evident that the extension to nonexponential service times is more difficult when $r > 0$, but we conjecture that the GI/M/s/∞ results in Ref. 4 extend to G/M/s/∞ systems (nonrenewal arrival processes) and that the conditioning heuristic is valid in heavy traffic for G/M/s/r systems. In support of this, the conservation relationships (10) and (14) extend to G/M/s/r systems. (The factor k on the right side of (14) is replaced by $\min\{k, s\}$.) The corresponding heavy-traffic local limit theorem for M/M/s/r systems is easy to prove using the methods of Ref. 4 or the Appendix. We conjecture that heavy-traffic local limit theorems corresponding to (17) through (20) are also valid for GI/M/s/r systems as well as for the more general G/M/s/r systems. Moreover, we conjecture that the form of the limits will coincide with what we get by applying the conditioning heuristic to the GI/M/s/∞ limits in Ref. 4.

In this section, let X_α and Y_α be the equilibrium number of customers in G/M/s/∞ and G/M/s/r systems, respectively, at an arbitrary time. For α large with $(1-\rho)\sqrt{s} = \beta$, the approximations derived from Ref. 4, where the limit theorem is proved only for renewal arrival processes (see Propositions 1 and 2 and Theorems 1 and 4), are

$$P(X_\alpha \geq s) \equiv \gamma \approx [1 + \beta'\Phi(\beta')/\phi(\beta')]^{-1}, \tag{29}$$

$$P(X_\alpha > s + r \mid X_\alpha \geq s) \approx \eta \equiv e^{-\beta' r/\sqrt{s}}, \tag{30}$$

$$P(Y_\alpha \geq s) \equiv \xi \approx \gamma(1 - e^{-\beta' r/\sqrt{s}})/(1 - e^{-\beta' r/\sqrt{s}}), \tag{31}$$

$$\sqrt{s}P(Y_\alpha = k \mid Y_\alpha \leq s) \approx \phi(\beta' + \delta)/\Phi(\beta'), \tag{32}$$

$$\sqrt{s}P(Y_\alpha = k \mid Y_\alpha \geq s) \approx (\gamma\beta'/\xi)e^{-\beta'\delta}, \tag{33}$$

and

$$\sqrt{s}B_T \approx \beta' e^{-\beta' r/\sqrt{s}} \tag{34}$$

for $(k - s)/\sqrt{s} = \delta$ and $\beta' = \beta/z$ with z being the peakedness in (4).

Since

$$E(\min\{X_\alpha, s\}) = \alpha = s\rho \tag{35}$$

for all G/G/s/∞ systems, e.g., by (4.2.3) of Ref. 27,

$$\begin{aligned} E(\min\{Y_\alpha, s\}) &= \frac{sP(s \leq X_\alpha \leq s + r) + s\rho - sP(X_\alpha \geq s)}{P(X_\alpha \leq s + r)} \\ &\approx [s(\gamma - \gamma\eta) + s(\rho - \gamma)]/(1 - \gamma\eta) \\ &\approx s(\rho - \gamma\eta)/(1 - \gamma\eta), \end{aligned} \tag{36}$$

for γ and η in (29) and (30), so that by (10)

$$B_C = 1 - \alpha^{-1}E(\min\{Y_\alpha, s\}) \approx 1 - \rho^{-1}(\rho - \gamma\eta)/(1 - \gamma\eta)$$
$$\approx (1 - \rho)\gamma\eta/\rho(1 - \gamma\eta) \tag{37}$$

and

$$\sqrt{s}B_C \approx \beta\gamma\eta/\rho(1 / \gamma\eta) \approx (z/\rho)\sqrt{s}B_T \tag{38}$$

so that $B_C \approx zB_T$ as in (13), although the correction with ρ in (38) may be useful.

Also note that

$$\eta \equiv \eta(\rho, s, r, z) \equiv \eta(\rho, r, z) \approx \eta(\rho, r/z, 1) \tag{39}$$

and

$$\gamma \equiv \gamma(\rho, s, r, z) \equiv \gamma(\rho, s, z) \approx \gamma(\rho, s/z^2, 1), \tag{40}$$

so that

$$B_C \equiv B_C(\rho, s, r, z) \approx B_C(\rho, s/z^2, r/z, 1) \tag{41}$$

or, equivalently,

$$B_C \equiv B_C(\alpha, s, r, z) \approx B_C(\alpha/z^2, s/z^2, r/z, 1) \tag{42}$$

in the manner of Hayward's approximation in Section 6.2. If we express B_C in terms of $1 - \rho$, then we have the alternate expression

$$B_C \equiv B_C(1 - \rho, s, r, z) \approx B_C[(1 - \rho)/z, s, r, 1]. \tag{43}$$

We can achieve (42) by fixing s, r and μ and then changing λ so that $(1 - \rho)/z$ is unchanged while z is replaced by 1. As with Hayward's approximation, we can use M/M/s/r formulas when $z = 1$.

Note that the normalizations in (41) through (43) are not the same as in Hayward's approximation in Section 6.2. Comparing (42) with (22), we see that now α and s are divided by z^2 in (42) instead of z, so that evidently the peakedness has a much greater impact when there is a finite waiting room. This might be expected because now the boundary where losses occur is further away from the center of mass, with ρ required to be less than one. The different normalization of r and s might be expected because this approximation is based on r being of order $\sqrt{s}$.

Of course, the approximations developed in this section need to be tested, which has not yet been done. From the theory, we know that the modification of Hayward's approximation for finite waiting rooms in (42) should work well when ρ is high but less than one, s is large, and r is of order $\sqrt{s}$, but it remains to determine the actual range over which the approximation is good.

VIII. TESTING THE BLOCKING APPROXIMATION

In this section we present numerical comparisons to test the approximation for the approximate blocking probability B_C in (13). Our testing here is confined to G/M/s/0 systems. Since formula (13) coincides with the normal approximations of Rahko[10-13] and Delbrouck,[14] their comparisons are relevant too.

Tables I through IV compare the heavy-traffic approximation (13) with exact M/M/s/0 results and the equivalent random method. We selected seven different blocking probabilities: 0.001, 0.01, 0.05, 0.10, 0.20, 0.40, and 0.60. The higher numbers were selected so that we could test the lower bound in (23), which is only applicable when $\alpha > s$. We also selected four different (z, s)-pairs: (1, 5), (1, 50), (2, 50), and (10, 400). In each case, we used the charts on pages 23–32 of Wilkinson[43] to determine the corresponding load in Erlangs dictated by the equivalent random method for the given blocking probabilities and parameters s and z. When the peakedness was not 1 (Tables III and IV), we also calculated Hayward's approximation (22). For Hayward's approximation we often used the more detailed graphs in Appendix A of Cooper.[44] (The results were also checked using Jagerman's computer programs.[35]) In parentheses next to the lower bound is the upper bound obtained from (26). It should be noted that the upper bound in (26) is an upper bound for our normal approximation, not necessarily on the true blocking probability.

In interpreting the tables, remember that the equivalent random method is only an approximation too when the arrival process is not Poisson. Moreover, from Holtzman[15] we know that the range of possible blocking probabilities consistent with the partial information provided by the peakedness can be quite wide. (As we indicated in Section 6.3, this is not true in heavy traffic.) Hence, there often is little reason to prefer the numerical accuracy of exact calculations according to the Erlang loss formula (21) over the normal approximation (13).

Table I—The blocking probability $B_C(s, \alpha, z)$ for $z = 1$ and $s = 5$

Load in Erlangs α	Blocking Probability	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
0.77	0.001	0.000	
1.37	0.01	0.003	
2.22	0.05	0.05	
2.86	0.10	0.11	
4.0	0.20	0.23	
6.6	0.40	0.39	0.24 (0.33)
11.5	0.60	0.49	0.57 (0.77)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

Table II—The blocking probability $B_C(s, \alpha, z)$ for $z = 1$ and $s = 50$

Load in Erlangs α	Blocking Probability	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
32.5	0.001	0.001	
38.0	0.01	0.01	
44.5	0.05	0.05	
49.7	0.10	0.10	
59.0	0.20	0.21	0.15 (0.16)
82.0	0.40	0.42	0.39 (0.40)
122.0	0.60	0.59	0.59 (0.61)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

As we expected, the quality of the approximation, as measured against the exact formula when $z = 1$ or the equivalent random method when $z \neq 1$, improves as α or s increases and z decreases. The parameter α/z gives a good indication of the quality to be expected; i.e., the quality depends approximately on α/z and improves as α/z increases. The heavy-traffic approximation tends to degrade as the number of servers gets beyond two or three standard deviations ($\sqrt{\alpha z}$) away from the G/GI/∞ mean (the load α).

Table V presents some of Kuczura's[45] results (as displayed in his Figures 1–3) for GI+M/M/s/0 systems (having an arrival process that is a superposition of a renewal process and a Poisson process) together with our heavy-traffic approximation. A significant feature of these systems is that the blocking experienced by the customers in the different streams is not the same. Since the arrival rates in the two streams are identical in each case, the blocking experienced by an arbitrary customer, which is what our approximations are for, is the average of the blocking probabilities associated with the separate streams.

Table III—The blocking probability $B_C(s, \alpha, z)$ for $z = 2$ and $s = 50$

Load in Erlangs α	Approximate Blocking Probability (Equiv. Rand.)	Hayward's Approximation B(25, α/2)	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho >$ 1
26.5	0.001	0.001	0.001	
32.6	0.01	0.01	0.01	
40.2	0.05	0.05	0.06	
45.8	0.10	0.10	0.11	
55.5	0.20	0.20	0.21	0.10 (0.10)
78.9	0.40	0.40	0.38	0.37 (0.38)
120.0	0.60	0.59	0.58	0.58 (0.61)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

Table IV—The blocking probability $B_C(s, \alpha, z)$ for $z = 10$ and $s = 400$

Load in Erlangs α	Approximate Blocking Probability (Equiv. Rand.)	Hayward's Approximation B(40, α/10)	Heavy-Traffic Approximation (13)	Bound (23) $1 - \rho^{-1}$ for $\rho > 1$
256	0.001	0.002	0.001	
290	0.01	0.01	0.01	
340	0.05	0.05	0.05	
385	0.10	0.10	0.10	
460	0.20	0.20	0.22	0.13 (0.14)
640	0.40	0.40	0.41	0.38 (0.39)
900	0.60	0.56	0.56	0.56 (0.57)

Note: In parentheses to the right of the lower bound (23) is the approximate upper bound in (26).

To obtain the heavy-traffic approximation, it is necessary to specify the peakedness of the arrival process. When the arrival process is the superposition of independent renewal processes, at least one of which is not Poisson, the superposition process is not a renewal process (see Ref. 26 and its references). However, the peakedness of the superposition process is clearly the convex combination of the individual peakedness values. Suppose there are n independent streams with λ_i the arrival rate and z_i the peakedness of stream i. Then clearly

Table V—The blocking probability for a GI+M/M/s/0 model with $s = 25$: comparison with Kuczura[45]

System	Load in Erlangs, α		
	20	25	30
M/M/s/0			
Arbitrary arrival	0.050	0.144	0.245
Heavy traffic	0.054	0.148	0.229
D + M/M/s/0			
Poisson arrival	0.045	0.15	0.28
Renewal arrival	0.027	0.11	0.20
Arbitrary arrival	0.036	0.13	0.24
Heavy traffic ($z = 0.75$)	0.037	0.124	0.214
Hayward ($z = 0.75$)	0.035	0.126	0.232
GI + M/M/s/0 [$z(G) = 2$]			
Poisson arrival	0.058	0.14	0.21
Renewal arrival	0.094	0.20	0.31
Arbitrary arrival	0.076	0.17	0.26
Heavy traffic ($z = 1.5$)	0.084	0.186	0.274
Hayward ($z = 1.5$)	0.077	0.172	0.268
GI + M/M/s/0 [$z(G) = 3$]			
Poisson arrival	0.06	0.14	0.20
Renewal arrival	0.15	0.26	0.38
Arbitrary arrival	0.115	0.20	0.29
Heavy traffic ($z = 2$)	0.12	0.21	0.30
Hayward ($z = 2$)	0.101	0.195	0.286
Equivalent random method ($z = 2$)	0.105	0.200	0.285

Note: All entries except the heavy-traffic, equivalent-random method, and Hayward values are from Kuczura.[45]

$$\lambda = \sum_{i=1}^{n} \lambda_i \qquad \text{and} \qquad \sum_{i=1}^{n} \lambda_i z_i/\lambda. \tag{44}$$

We use (44) to obtain the peakedness values for the heavy-traffic approximation given in Table V. For the case in which the peakedness of the superposition process is $z = 2$, we also compared the blocking probabilities with those obtained using the equivalent random method, Chart 2 on page 24 of Wilkinson,[43] and Hayward's approximation using Jagerman's program.[35] The results seem to be good, about the same as those in Tables I through IV.

IX. ACKNOWLEDGMENTS

I am grateful to K. J. Biewenga, O. J. Boxma, and J. W. Cohen of the University of Utrecht in the Netherlands for first pointing out the early work described in Section 6.1, and to D. L. Jagerman and M. Segal for their interest and assistance.

REFERENCES

1. A. A. Borovkov, "On Limit Laws for Service Processes in Multi-Channel Systems," Siberian Math. J., *8* (1967), pp. 746–763 (English translation).
2. A. A. Borovkov, *Stochastic Processes in Queueing Theory,* New York: Springer-Verlag, 1976.
3. W. Whitt, "On the Heavy-Traffic Limit Theorem for GI/G/∞ Queues," Adv. Appl. Prob., *14,* No. 1 (March 1982), pp. 171–190.
4. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues With Many Exponential Servers," Oper. Res., *29,* No. 3 (May–June 1981), pp. 567–588.
5. A. A. Fredericks, "Congestion in Blocking Systems—A Simple Approximation Technique," B.S.T.J., *59,* No. 6 (July–August 1980), pp. 805–827.
6. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.3.
7. E. Brockmeyer, "Approximative Formulae for Loss and Improvement," Appendix 2 in *The Life and Work of A. K. Erlang,* E. Brockmeyer, H. L. Halstrom, and A. Jensen, eds., Copenhagen: Danish Academy of Sciences, 1948, pp. 120–126.
8. E. Vaulot, "Les Formules d' Erlang et leur Calcul Pratique," Ann. Telecom., *6* (1951), pp. 279–286.
9. D. L. Jagerman, "Some Properties of the Erlang Loss Functions," B.S.T.J., *53,* No. 3 (March 1974), pp. 525–551.
10. K. Rahko, "The Dimensioning of Local Telephone Traffic Routes Based on the Distribution of the Traffic Carried," Acta Polytechnica Scandinavica, 1967, El 14, 95.
11. K. Rahko, "Dimensioning of Traffic Routes According to the EERT-Method and Corresponding Methods," Proc. Eighth Inter. Teletraffic Congress, Melbourne, November 1976, p. 143.
12. K. Rahko, "Some Methods for Approximation of Congestion," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.6.
13. S. Hertzberg, "Approximative Dimensioning Formulas with the Normal Distribution Method," Reports 3/79 and 5/82, Telecommunication Switching Laboratory, Helsinki University of Technology, 1979 and 1982 (in Swedish).
14. L. E. N. Delbrouck, "A Unified Approximate Evaluation of Congestion Functions for Smooth and Peaky Traffics," IEEE Trans. Comm., *COM-29,* No. 2 (February 1981), pp. 85–91.
15. J. M. Holtzman, "The Accuracy of the Equivalent Random Method With Renewal Inputs," B.S.T.J., *52,* No. 2 (May 1973), pp. 1673–1679.
16. M. Sobel, "Simple Inequalities for Multiserver Queues," Management Sci., *26,* No. 9 (September 1980), pp. 951–956.

17. D. P. Heyman, "Comments on a Queueing Inequality," Management Sci., *26,* No. 9 (September 1980), pp. 956–959.
18. A. K. Erlang, "On the Rational Determination of the Number of Circuits," *The Life and Works of A. K. Erlang,* E. Brockmeyer, H. L. Halstrom, and A. Jensen, eds., Copenhagen: Danish Academy of Technical Sciences, 1948, pp. 216–221.
19. W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol. I, 2nd Ed., New York: John Wiley and Sons, 1968.
20. J. DeBoer, "Limits and Asymptotes of Overflow Curves," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.4.
21. A. A. Fredericks, "Approximating Parcel Blocking Via State Dependent Birth Rates," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.2.
22. K. Lindberger, "Simple Approximations of Overflow System Quantities for Additional Demands in the Optimization," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 5.3.3.
23. B. Sanders, W. H. Haemers, and R. Wilcke, "Simple Approximation Techniques for Congestion Functions for Smooth and Peaked Traffic," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.1.
24. E. A. van Doorn, "Some Analytical Aspects of the Peakedness Concept," Proc. Tenth Inter. Teletraffic Congress, Montreal, June 1983, p. 4.4b.5.
25. G. F. Newell, "The $M/M/\infty$ Service System With Ranked Servers in Heavy Traffic," Institute of Transportation Studies, University of California at Berkeley, 1983.
26. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30,* No. 1 (January–February 1982), pp. 125–147.
27. P. Franken, D. König, U. Arndt and V. Schmidt, *Queues and Point Processes,* Berlin: Akademic-Verlag, 1981.
28. R. W. Wolff, "The Effect of Service Time Regularity on System Performance," in *Computer Performance,* K. M. Chandy and M. Reiser, eds., Amsterdam: North-Holland, 1977, pp. 297–304.
29. W. Whitt, "Minimizing Delays in the GI/G/1 Queue," Oper. Res., *32,* No. 2 (March–April 1984), to be published.
30. D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research,* Vol. 1, New York: McGraw-Hill, 1982.
31. D. L. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54,* No. 3 (March 1975), pp. 625–661.
32. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30,* No. 2 (March–April 1982), pp. 223–231.
33. W. Whitt, "Approximating a Point Process by a Renewal Process: The View Through a Queue, An Indirect Approach," Management Sci., *27,* No. 6 (June 1981), pp. 619–636.
34. D. L. Jagerman, unpublished work.
35. D. L. Jagerman, unpublished work.
36. P. Bloomfield and D. R. Cox, "A Low Traffic Approximation for Queues," J. Appl. Prob., *9,* No. 4 (December 1972), pp. 832–840.
37. G. F. Newell, *Applications of Queueing Theory,* Second Edition, London: Chapman Hall, 1982.
38. D. Y. Burman and D. R. Smith, "A Light Traffic Theorem for Multiserver Queues," Math. Oper. Res., *8,* No. 1 (February 1983), pp. 15–25.
39. D. Y. Burman and D. R. Smith, "Asymptotic Analysis of a Queueing Model With Bursty Traffic," B.S.T.J., *62,* No. 6 (July–August 1983), pp. 1433–1453.
40. M. I. Reiman and B. Simon, unpublished work.
41. S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes of Queues," Oper. Res., *32* (1984).
42. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *63,* No. 9 (November 1983), pp. 2779–2815.
43. R. I. Wilkinson, *Nonrandom Traffic Curves and Tables,* Traffic Studies Center, Bell Laboratories, 1970.
44. R. B. Cooper, *Introduction to Queueing Theory,* New York: The Macmillan Company, 1972.
45. A. Kuczura, "Loss Systems with Mixed Renewal and Poisson Inputs," Oper. Res., *21,* No. 3 (May–June 1973), pp. 787–795.

APPENDIX

A Heavy-Traffic Local Limit Theorem for the Erlang Loss Formula

For the elementary M/G/s/0 system, where the blocking probability is given by the Erlang loss formula (21), approximations (9) and (13) follow easily from well-known limit theorems for the Poisson distribution. We present one possible argument here.

Let $p(k, \lambda)$ denote the probability mass function of the Poisson distribution with mean λ, i.e.,

$$p(k, \lambda) = e^{-\lambda}\lambda^k/k!. \tag{45}$$

The Erlang loss formula then can be expressed as

$$B(s, \alpha) = p(s, \alpha) \Big/ \sum_{k=0}^{s} p(k, \alpha). \tag{46}$$

As before, let $\phi(x)$ and $\Phi(x)$ be the density and cdf, respectively, of a standard normal random variable, $N(0, 1)$.

Theorem: In an M/G/s/0 system,

$$\lim_{\alpha\to\infty} \sqrt{\alpha}B([\alpha + s\sqrt{\alpha}], \alpha) = \phi(s)/\Phi(s),$$

where $[x]$ *is the greatest integer less than or equal to x.*

Proof: Since a Poisson random variable with mean λ has the same distribution as the sum of n i.i.d Poisson random variables with mean λ/n, the central limit theorem can be applied to obtain

$$\lim_{\alpha\to\infty} \sum_{k=0}^{[\alpha+s\sqrt{\alpha}]} p(k, \alpha) = \Phi(s)$$

(see Problem 9, p. 194, and Example $X(c)$, p. 245, of Feller[19]). Hence, it remains to show that

$$\lim_{\alpha\to\infty} \sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] = \phi(s) = (2\pi)^{-1/2}e^{-s^2/2}.$$

We can establish this result using Stirling's formula (p. 52 of Feller[19]). As in Stirling's formula, let the symbol ~ below mean that the ratio of the two sides tends to 1 as $\alpha \to \infty$. We have

$$\begin{aligned}\sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] &= \frac{\sqrt{\alpha}e^{-\lambda}(\alpha^{(\alpha+s\sqrt{\alpha})})}{(\alpha + s\sqrt{\alpha})!}\\ &\sim \frac{\sqrt{\alpha}e^{-\alpha}(\alpha^{(\alpha+s\sqrt{\alpha})})}{\sqrt{2\pi}(\alpha + s\sqrt{\alpha})^{(\alpha+s\sqrt{\alpha})}(\alpha + s\sqrt{\alpha})^{1/2}e^{-(\alpha+s\sqrt{\alpha})}}\\ &\sim \frac{e^{s\sqrt{\alpha}}}{\sqrt{2\pi}(1 + s/\sqrt{\alpha})^{\alpha}(1 + s/\sqrt{\alpha})^{s\sqrt{\alpha}}}\\ &\sim \frac{e^{-s^2}}{\sqrt{2\pi}}\frac{e^{s\sqrt{\alpha}}}{(1 + s/\sqrt{\alpha})^{\alpha}},\end{aligned}$$

where

$$\log(e^{s\sqrt{\alpha}}(1 + s/\sqrt{\alpha})^{-\alpha}) = s\sqrt{\alpha} - \alpha \log(1 + s/\sqrt{\alpha})$$

$$= s\sqrt{\alpha} - \alpha\left(\frac{s}{\sqrt{\alpha}} - \frac{1}{2}\left(\frac{s}{\sqrt{\alpha}}\right)^2 + o(\alpha)\right) = \frac{s^2}{2} + 0(1).$$

Hence,

$$e^{s\sqrt{\alpha}}(1 + s/\sqrt{\alpha})^{\alpha} \sim e^{s^2/2}$$

and

$$\sqrt{\alpha}p[(\alpha + s\sqrt{\alpha}), \alpha] \sim (2\pi)^{-1/2}e^{-s^2/2}$$

as claimed.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department in the Systems Analysis Center.