Lecture Notes
in Economics and
Mathematical Systems

Managing Editors: M. Beckmann, Providence, and H. P. Künzi, Zürich

Operations Research

98

# Mathematical Methods in Queueing Theory

Proceedings of a Conference at Western Michigan University, May 10–12, 1973

Sponsored jointly by Western Michigan University and the U. S. Office of Naval Research under Grant Number NONR(G)-00017-73

Edited by A. B. Clarke

Springer-Verlag
Berlin · Heidelberg · New York 1974

# HEAVY TRAFFIC LIMIT THEOREMS FOR QUEUES: A SURVEY

Ward Whitt[1]
Yale University

CONTENTS



---

[1]Partially supported by NSF Grant GK-38149.

1. INTRODUCTION. Heavy traffic research is part of a general program to obtain simple descriptions and useful approximations for queueing models. Throughout this paper, I try to put heavy traffic research into this broader perspective. I discuss the two principal objectives of heavy traffic research, namely, (1) to describe unstable queueing systems and (2) to approximate stable queueing systems. These objectives are each related to more general themes. In an unstable queueing system the queueing processes are growing processes, so that descriptions of queueing processes in unstable queueing systems are similar to descriptions of other growing processes associated with queues. In the same way, heavy traffic approximations for stable queues are part of a large class of approximations for queueing systems. I relate heavy traffic research to these more general activities.

The central theme here is limit theorems, with the central limit theorem being truly central. To give an accurate picture, I discuss other limit theorems in addition to heavy traffic limit theorems, but I emphasize heavy traffic. I review the basic techniques and survey recent progress. Above all, I try to make this an informal discussion concentrating on essential ideas. This seems to be a good forum to wax philosophical and wane mathematical. There is a bibliography on heavy traffic which I have tried to make as complete as possible. It indicates where theorems and proofs can be found. A sample of the mathematics is available in the paper by LOULOU (1973b) in these proceedings.

To a large extent, this conference can be regarded as a sequel to the Symposium on Congestion Theory held at the University of North Carolina in 1964. From that point of view, this paper is a sequel to the papers of HEATHCOTE (1965) and KINGMAN (1965a). The name "heavy traffic" is due to Kingman who initiated the work. (The first heavy traffic limit theorem seems to have been proved by KENDALL (1957).) The term "heavy traffic" is now used in a somewhat broader way to refer to unstable systems as well as highly saturated stable systems. After KINGMAN (1961, 1962), important work was done by several Russians, chiefly PROHOROV (1963) and BOROVKOV (1964, 1965, 1967a,b,c). My own involvement began with my doctoral dissertation in the Department of Operations Research at Cornell University which was directed by Donald Iglehart, cf. WHITT (1968). IGLEHART (1965a,b, 1967) had previously investigated the asymptotic behavior of unstable queues and diffusion approximations for queueing processes. Active interest in this problem was also shared by N. U. Prabhu at Cornell.

In addition to his own research in this area, Prabhu directed theses by LALCHANDANI (1967) and WORTHINGTON (1967) which preceded my work.

Previous surveys of heavy traffic research appear in KINGMAN (1965a) and WHITT (1968). Asymptotic methods in queueing have been reviewed by COHEN (1972) and IGLEHART (1972c). Surveys of diffusion approximations and/or weak convergence in function spaces appear in BILLINGSLEY (1968, 1971), IGLEHART (1967, 1972b), NEWELL (1971), and WHITT (1968, 1970b). As far as queueing texts are concerned, the Russian books appear to be most in the spirit of the investigations reported here, cf. GNEDENKO and KOVALENKO (1968) and BOROVKOV (1972).

In addition to making a survey, I attempt to point out important open problems. Naturally, I will be happy to hear about solutions.

2. DESCRIBING QUEUEING PROCESSES. The extensive literature on queues has primarily been concerned with describing the stochastic processes of interest in various queueing models. While several papers presented at this conference reflect serious work on other important topics (e.g., control, computation, and statistical analysis), the description of queueing processes remains an active area of research. That there still remains something to say after the great outpouring of papers and books describing queueing processes is a testimonial to the fertility of queueing theory. I believe the motivation has been mathematical for the most part. Queueing processes attract attention because they constitute just the right blend of simplicity and complexity. Queueing processes are enough like basic processes such as renewal processes and random walks to suggest the possibility of successful analysis. At the same time, they are sufficiently different to present a serious challenge. In this regard, queues might be thought of as the Sirens of probability theory. I make this remark because queueing theory has not had as much to say about practical problems as either the practitioners or the theoreticians would like. It has been said that the principal value of queueing theory is as a probability training ground. I think the value of such a training ground should not be minimized, but I also think the implied criticism of queueing theory is excessive. It is evident that queueing theory is moving forward and that applicability is being enhanced. Nevertheless, periodic reevaluations from the point of view of applicability can only help the future of queueing theory.

The limitations of queueing theory are obviously due in part to the inability to obtain results in a usable form. To a large extent, queueing theory remains behind the Laplacian curtain, cf. KENDALL

(1964). At present, it is not easy to work with double, triple, or quadruple transforms. While ever-increasing computational power will probably make such results more relevant, it appears that the transforms may be bypassed altogether when serious computation is to be performed. As Marcel Neuts has emphasized during this conference, it is appropriate to formulate and analyze queueing models from the outset with the intended computation in mind. The computation should not be thought of as something you tack on after the mathematical analysis has gone as far as it can. Early work in this direction suggests that it will be more fruitful to work with basic structural relations than the transforms. Thus, the transforms may never be used at all.

The most serious problem may not be inverting the transforms or computing the complex queueing formulas. In many instances, the specific queueing phenomenon is just not well enough understood to warrant such a detailed description. Furthermore, the relevant decisions may not depend on such detail. As Gordon Newell observed during this conference, the successful analysis of an actual queueing problem can often reduce to the intelligent use of a straightedge.

Thus, it appears that the utility of queueing theory could be enhanced by developing approximation procedures and underlying common principles, that is, quantitative propositions which have sufficient generality and consequence for decision making. The idea is to avoid unnecessary complexity and strive for simple representations which capture enough of the essential features to be useful. The book by NEWELL (1971) is certainly in this spirit, and the research related here is intended to be. The approximation procedures and common principles here stem from limit theorems. While these asymptotic descriptions of queueing processes already constitute an important chapter in queueing theory, the practical value should not be overemphasized. The limit theorems suggest and support certain approximations, but the approximations can be used without any limit theorems. If faced with a practical queueing problem, I certainly would not sit down and try to prove a functional limit theorem. But of course this is not the way limit theorems are applied. The central limit theorem does not have to be re-proved every time it is applied. The limit theorems enhance our understanding and help us make useful approximations. They form a background for successful problem solving.

Open Problem. The practical utility of the approximations suggested by the limit theorems in this paper remains largely unknown. There is a great need for computational experience. Some work in this direction for heavy traffic limit theorems has been reported by GAVER

(1968), but more needs to be done.

Open Problem. Approximations need not be generated by limit theorems. More approximating procedures for queueing processes should be examined.

3. LIMIT THEOREMS FOR GROWING PROCESSES. Early work describing queueing processes focused on the steady-state or limiting behavior. If the traffic intensity $\rho$ is below its critical value, the standard queueing processes usually have limiting distributions as time gets large and these limiting distributions can often be determined by appropriate balance equations. Attention next moved to the time-dependent or transient behavior of queueing processes. It was discovered that the distribution of many queueing processes at finite time points could be described in detail, albeit a bit cumbersomely. The principal texts on queueing theory give a good account of available results on both steady-state and transient behavior, at least for the GI/G/1 system. The descriptions here fall into neither of these two familiar categories. The descriptions here involve growing processes and system approximations. Growing processes are discussed in this section and the system approximations are discussed in the next section.

It is possible to look at either the time-dependent or the asymptotic behavior of growing processes, but only the asymptotic behavior is considered here. Of course, growing processes do not have limits in the usual sense, so that they must be normalized by subtracting and dividing with appropriate functions of time before letting time get large in order to make useful statements. For example, if $\{X(t), t \geq 0\}$ is a growing stochastic process arising in some queueing model, the object is to find deterministic functions $a(t)$ and $b(t)$ and a nondegenerate random variable $L$ such that

$$\frac{X(t) - a(t)}{b(t)} \to L \quad \text{as} \quad t \to \infty, \tag{3.1}$$

where $\to$ denotes convergence in distribution (weak convergence) or some other mode of stochastic convergence. Often, corresponding to the central limit theorem, $a(t) = at$ and $b(t) = bt^{\frac{1}{2}}$.

3.1 The Classical Limit Theorems. Of course, there is nothing novel about considering such limit theorems for growing processes. These limit theorems are just variations of the classical limit theorems in probability theory. The classical limit theorems include the central limit theorems and the laws of large numbers as well as somewhat less

familiar limit theorems such as laws of the iterated logarithm and extreme value theorems. The simplest setting for all these classical theorems is a sequence of i.i.d. (independent and identically distributed) random variables $\{X_n, n \geq 1\}$ with appropriate moments finite, but the theorems also hold when the sequence $\{X_n\}$ is only almost i.i.d. Since these limit theorems are distribution-free; that is, since they do not depend on the actual distribution of $X_n$, these theorems are said to be invariance principles. (This name is also somewhat inappropriately applied to function space generalizations of the classical limit theorems. The name is inappropriate because they are not unique in this regard.) Since the limit theorems tend to be relatively insensitive to both the specific distributions and the i.i.d. assumptions, the limit theorems and the resulting approximations are called robust.

The statements of the classical limit theorems involve the asymptotic behavior as $n \to \infty$ of the associated sequence of partial sums $\{S_n, n \geq 0\}$, where $S_n = X_1 + \ldots + X_n$ and $S_0 = 0$, or the associated sequence of extreme values $\{E_n, n \geq 1\}$, where $E_n = \max_{1 \leq k \leq n} X_k$. These classical limit theorems play a vital role in probability theory because they reveal the statistical regularity associated with a macroscopic view of uncertainty. Just as in statistical mechanics or macroeconomics, there is order associated with a macroscopic view which is not apparent from the microscopic view. For example, in the setting above the exact distribution of $S_n$ is an $n$-fold convolution of the distribution of $X_1$, which is usually quite complicated to compute and express, but as $n$ grows, the distribution of $S_n$ can very rapidly be described quite accurately by the limiting normal distribution in the classical central limit theorem. What has just been said suggests that the limit theorem is useful primarily because it might be difficult to compute the distribution of $S_n$. In fact, the limit theorem is of value whether or not the distribution of $S_n$ can be computed. Obviously, a computer program could be devised to approximate the distribution of $S_n$ arbitrarily closely in any specific situation, but this cannot replace the understanding provided by the limit theorem. Being able to compute the distribution of $S_n$ or even actually doing so in any instance does not capture the general tendency to the bell-shaped normal curve. It is the general principle provided by the limit theorem which is most important. The limit

theorem acts like a physical law explaining common tendencies in a wide range of phenomena. This is the reason that the classical limit theorems form the core of probability theory.

The ideas behind classical limit theorems for queues are quite simple. In order to focus on the main ideas and keep everything as simple as possible, for the most part I only talk about the sequence $\{W_n, n \geq 1\}$ of waiting times (until beginning service) of successive customers in a standard GI/G/1 queue. While attention is focused on this particular process in this particular queueing system, it is significant that similar results are usually available for other queueing processes and much more complicated systems. This is especially true for the heavy traffic limit theorems.

For the classical limit theorems for queues, it is important to identify two cases: stable queues and unstable queues. Classical limit theorems for stable queues follow by relating the sequence $\{W_n, n \geq 1\}$ to the underlying sequence $\{X_n, n \geq 1\}$ in the classical setting. While the sequence $\{W_n, n \geq 1\}$ is certainly not i.i.d., it is close enough for the limit theorems. Classical limit theorems for unstable queues follow by relating the sequence $\{W_n, n \geq 1\}$ to the sequence of partial sums $\{S_n, n \geq 1\}$ in the classical setting. When $\rho > 1$, $\{W_n\}$ usually has the same limit behavior as an appropriate sequence of partial sums $\{S_n\}$, cf. Section 3.3.1. When $\rho = 1$, the limit behavior of $\{W_n\}$ is not the same as $\{S_n\}$ but it is easily obtained as a consequence, cf. Section 3.3.2. These are the essential ideas; the rest of this section only develops them in slightly more detail.

It is also important that the various limit theorems for queueing processes described below can usually be proved by applying existing limit theorems for basic processes instead of imitating the proofs of these earlier theorems. This is important, not only for reducing the length of the argument, but also for understanding the limiting phenomenon.

3.2 Stable Queues ($\rho < 1$). Consider a standard GI/G/1 queue determined by two independent sequences of i.i.d. random variables: $\{u_n, n \geq 1\}$ and $\{v_n, n \geq 0\}$. Assume a $0^{th}$ customer arrives at time $t = 0$ to find a free server. Let $v_n$ represent the service time of the $n^{th}$ customer and let $u_n$ represent the inter-arrival time between the $(n-1)^{st}$ and $n^{th}$ customers. Make the following definitions:

$$\rho = Ev/Eu \ ,$$

$$Y_n = v_{n-1} - u_n \ , \quad n \geq 1 \tag{3.2}$$

$$W_{n+1} = [W_n + Y_{n+1}]^+ \ , \quad n \geq 0 \ ,$$

with $W_0 = 0$, where $[x]^+ = \max\{0,x\}$. If $\rho < 1$, then the GI/G/1 queue is stable. In particular, there is a nondegenerate random variable $W$ such that $W_n \Rightarrow W$ as $n \to \infty$. (I use $\Rightarrow$ to denote weak convergence.)

3.2.1 Cumulative Processes. The main classical limit theorems for $\{W_n\}$ in this setting involve time averages. In particular, under appropriate moment conditions, the central limit theorem, strong law of large numbers, and law of iterated logarithm state, respectively:

$$(\sigma^2 n)^{-\frac{1}{2}} \left[\sum_{k=1}^{n} W_k - nEW\right] \Rightarrow N(0,1),$$

$$n^{-1} \sum_{k=1}^{n} W_k \to EW \quad \text{a.s.}, \tag{3.3}$$

and

$$\limsup_{n\to\infty} \frac{\sum_{k=1}^{n} W_k - nEW}{(2\sigma^2 n \log\log n)^{\frac{1}{2}}} = 1 \quad \text{a.s.},$$

where $\sigma^2$ is a positive constant and $N(0,1)$ represents a random variable with the normal distribution, mean 0, and variance 1; cf. IGLEHART (1971b). One explanation for the limits in (3.3) is that the events $\{W_k = 0\}$ are regeneration points for the process $\{W_n, n \geq 1\}$. Furthermore, $W_k = 0$ for infinitely many $k$ with probability one. Thus there is an embedded persistent renewal process in $\{W_n, n \geq 1\}$. In other words, $\{W_n, n \geq 1\}$ is a regenerative process and $\{\sum_{k=1}^{n} W_k, n \geq 1\}$ is a cumulative process, cf. SMITH (1955, 1958). This means that the sequence $\{W_n\}$ can be broken up into i.i.d. blocks. Consequently, even though $\{W_n\}$ is not itself i.i.d., the theory for a sequence of i.i.d. random variable can be applied.

In particular, a partial sum $\sum_{k=1}^{n} W_k$ can be expressed as a random sum of i.i.d. blocks plus one additional term which is asymptotically negligible with the normalization.

The results in (3.3) also hold for the GI/G/s queue with $s > 1$ if $\rho = Ev/sEu < 1$ and $P\{u_n > v_n\} > 0$, cf. WHITT (1972b). The extra condition requiring that $P\{u_n > v_n\} > 0$ is needed to assure that the embedded renewal process be persistent. When an embedded persistent renewal process is not available, it seems that the classical limit theory for stationary processes with appropriate mixing conditions could be applied, but this remains to be examined carefully.

Open Problem. I conjecture that the limit theorems in (3.3) hold for $\{W_n, n \geq 1\}$ and other queueing processes such as the queue length process in the GI/G/s queue with $s > 1$ if $\rho < 1$ and appropriate moments are finite. In other words, I conjecture that the condition $P\{u_n > v_n\} > 0$ can be dropped.

Open Problem. Iglehart has derived expressions for the constant $\sigma^2$ in (3.3) which will appear. Descriptions of other norming constants are still needed.

3.2.2 Estimation. The limits for cumulative processes obviously have implications for statistical estimation. For example, (3.3) says that $n^{-1} \sum_{k=1}^{n} W_k$ is a consistent and asymptotically normal estimator of $EW$. The regenerative structure also suggests using the i.i.d. blocks as the basis of the statistical analysis. In simulation this solves the problem of finding an appropriate initial state. It is not necessary to let the system run until steady-state is approximated; the simulation can be begun at a regeneration point. The i.i.d. structure also permits the application of standard statistical techniques. For further discussion, see CRANE and IGLEHART (1973) and FISHMAN (1972a,b, 1973).

3.2.3 Occupation Times. A growing process can also be obtained by looking at the total time in the interval $[0,t]$ that a queueing process spends in some set $A$. This is called an occupation time process. For example, the process $\{I(t), t \geq 0\}$ where $I(t)$ represents the accumulated idle time up to time $t$ is such a process. Corresponding to (3.3), for a GI/G/1 queue with $\rho < 1$ under appropriate

moment conditions,

$$(\alpha^2 t)^{-\frac{1}{2}} [I(t) - (1-\rho)t] \Rightarrow N(0,1),$$

$$t^{-1} I(t) \to (1-\rho) \quad \text{a.s.}, \tag{3.4}$$

and

$$\limsup_{t\to\infty} \frac{I(t) - (1-\rho)t}{(2\alpha^2 t \log\log t)^{\frac{1}{2}}} = 1 \quad \text{a.s.},$$

where $\alpha^2$ is a positive constant, cf. IGLEHART (1971b) and WHITT (1971a). This particular occupation time process as well as more general occupation time processes in the GI/G/1 and GI/G/s queues are easy to treat, with the possible exception of the norming constants, because these occupation time processes are also cumulative processes in the sense of Section 3.2.1 above, cf. Remark (iii) on p. 280 of IGLEHART (1971b).

Occupation time processes are also investigated by different methods in the paper by TAKACS (1973) in these proceedings. The specific results in (3.4) were obtained by Takacs about fifteen years ago using similar methods. However, it appears that neither of the two kinds of independence assumptions in TAKACS (1973) covers as general occupation time processes as the cumulative process results which are applicable to most GI/G/s queues. It is also perhaps of interest to note that some of the methods used by TAKACS (1973) such as Dobrushin's composition results can be extended to more general modes of convergence in the function space setting, cf. WHITT (1972c, 1973a).

3.2.4 Extremal Processes. If $W_n^{\uparrow} = \max\{W_k,\ 0 \le k \le n\}$, then $\{W_n^{\uparrow},\ n \ge 0\}$ is a growing process. Since $\{W_n\}$ is something like an i.i.d. sequence when $\rho < 1$, it is reasonable to expect that the classical extreme value theorems for i.i.d. random variables should apply to $\{W_n\}$. This appears to be true but it has not yet been verified completely.

One approach is via stationary mixing structure. Let $\mathcal{F}_a^b$ denote the $\sigma$-field generated by events of the form $\{(X_{i_1}, \ldots, X_{i_m}) \in E\}$ where $0 < a \le i_1 \le \ldots \le i_m \le b$ and $E$ is a measurable subset of $R^m$. A sequence $\{X_n,\ n \ge 0\}$ is strong-mixing if

$$\sup\{|P(AB) - P(A)P(B)| : A \in \mathcal{F}_0^m,\ B \in \mathcal{F}_{m+k}^{\infty}\} \le \alpha(k),$$

where $\alpha(k) \to 0$ as $k \to \infty$. LOYNES (1965) showed that if there exists a nondegenerate c.d.f. $\Phi$ and constants $a_n$ and $b_n$ such that

$$\lim_{n\to\infty} P\{a_n^{-1}[X_n^{\uparrow} - b_n] \le x\} = \Phi(x)$$

for a stationary strong-mixing sequence $\{X_n, n \ge 0\}$, then $\Phi$ must belong to one of three classical extreme value distribution types, just as in the i.i.d. case. These three types are

$$\Phi_\alpha(x) = \begin{cases} 0 & , \ x \le 0 \\ \exp\{-x^{-\alpha}\} & , \ x > 0 , \ \alpha > 0 , \end{cases}$$

(3.5)
$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & , \ x < 0 , \ \alpha > 0 \\ 1 & , \ x \ge 0 \end{cases}$$

$$\Lambda(x) = \exp\{-e^{-x}\} , \quad -\infty < x < \infty.$$

Earlier, LOYNES (1962) showed that the single-server queue is stable ($W_n \Rightarrow W$) if the sequence $\{Y_n\}$ in (3.2) is stationary with $EY_1 < 0$ ($\rho < 1$). Furthermore, if the limiting distribution $W$ is used as the initial distribution for $\{W_n\}$, then the sequence $\{W_n\}$ becomes stationary too, cf. Lemma 1 and Theorem 3 of LOYNES (1962). It is also not difficult to show that the sequence $\{W_n\}$ is strong-mixing whenever the sequence $\{Y_n\}$ is strong-mixing, cf. Theorem 3.1 of WHITT (1971b). This means that the only possible nondegenerate limits for $\{a_n^{-1}[W_n^{\uparrow} - b_n]\}$ when $\{Y_n\}$ is stationary strong-mixing are the three classical types in (3.5). This obviously applies to the GI/G/1 queue as a special case. The hitch here is that, unlike the i.i.d. case, there are not necessary and sufficient conditions for convergence to each of the three types. There are sufficient conditions, but these have not yet been shown to be applicable.

A second approach is via regenerative structure. The case of a GI/G/1 queue in which $Y_1$ has an exponential tail has been treated by IGLEHART (1972a) using the regenerative structure. Assuming that $Ee^{aY_1} = 1$ and $EY_1e^{aY_1} < \infty$ for some constant $a$, Iglehart showed that

(3.6)
$$\lim_{n\to\infty} P\{aW_n^{\uparrow} - \log bn \le x\} = \exp\{-m^{-1}e^{-x}\},$$

for constants a, b, and m. The constant a is the root of $Ee^{aY_1} = 1$; it appears to be an important quantity. As a consequence of (3.6),

$$\frac{W_n^{\uparrow}}{\log n^{1/a}} \Rightarrow 1 \quad \text{as} \quad n \to \infty. \tag{3.7}$$

So far, other results seem to be restricted to the M/G/1 and GI/M/1 queues. See Section 8 of COHEN (1972) for a survey.

Open Problem. Extreme value theorems for all cases in the GI/G/1 queue have not yet been proved. Markovian assumptions for existing results should be unnecessary. Nothing at all has yet been said about the GI/G/s queue.

3.3 Unstable Queues ($\rho \geq 1$). If the input rate is greater than or equal to the output rate in a queueing system, then that system is unstable and the queueing processes tend to "blow up." For example, if $\rho \geq 1$ in a GI/G/1 queue, then for any $K < \infty$

$$\lim_{n\to\infty} P\{W_n \leq K\} = 0.$$

Heavy traffic limit theorems describe the growth in more detail. These limit theorems are useful because the input rate does often temporarily exceed the output rate in many queueing systems. The limit theorems yield approximations for systems which are unstable for a period of time.

There are two heavy traffic cases for a single queueing system: $\rho > 1$ and $\rho = 1$. In the context of Markov chains, these cases correspond to transience and null recurrence respectively. For example, if $\rho > 1$, and $K < \infty$, then $W_n < K$ for only finitely many $n$ with probability one. If $\rho = 1$, then $W_n < K$ for infinitely many $n$ with probability one. However, when $\rho = 1$, the expected time between epochs when customers arrive to find a free server is infinite. These facts are easy to verify given the relationship between $\{W_n\}$ and the random walk $\{S_n\}$. It follows from (3.2) by induction that

$$W_n = S_n - \min\{S_k,\ 0 \leq k \leq n\}, \quad n \geq 0, \tag{3.8}$$

where $S_n = Y_1 + \ldots + Y_n$ and $S_0 = 0$. If the system is a GI/G/1 queue, then

$$W_n \sim M_n = \max\{S_k,\ 0 \leq k \leq n\}, \tag{3.9}$$

where $\sim$ means equality in distribution. Consequently, the asymptotic behavior of $W_n$ coincides with the known asymptotic behavior of $M_n$, cf. Chapter 8 of CHUNG (1968). However, note that (3.8) holds everywhere for all $n$ while (3.9) holds only in distribution for a single $n$.

As before, I discuss only $\{W_n\}$ in the GI/G/1 queue, but results are available for other processes and other systems. The heavy traffic theory for all standard processes in a single unstable GI/G/1 queueing system appears in WHITT (1968, 1971a). Extensions to single unstable multi-channel systems, including the GI/G/s queue, appear in IGLEHART and WHITT (1970a) and LOULOU (1971, 1973a,b). The law of the iterated logarithm is discussed by IGLEHART (1971a). See the bibliography for a full list of references.

3.3.1 $\rho > 1$. It is well known from the theory of random walks that there is a nondegenerate random variable $L$ such that $\min\{S_k, 0 \le k \le n\} \Rightarrow L$ as $n \to \infty$ if and only if $EY_1 > 0$, which is equivalent to $\rho > 1$. This means that with normalization $\{W_n\}$ and $\{S_n\}$ will have the same limit behavior if $\rho > 1$. This is so because

$$\left|\frac{W_n - a_n}{b_n} - \frac{S_n - a_n}{b_n}\right| = \frac{|W_n - S_n|}{b_n} = \frac{-\min\{S_k, 0 \le k \le n\}}{b_n}, \tag{3.10}$$

which converges to zero if $b_n \to \infty$. Thus, the limit behavior of $\{W_n\}$ is obtained from known results for the basic process $\{S_n\}$. The connection is the convergence together theorem which says that if the distance between random elements $U_n$ and $V_n$ converges to 0 as $n \to \infty$, then $V_n$ converges to $V$ if and only if $U_n$ converges to $V$, cf. Theorem 4.1 of BILLINGSLEY (1968). As a consequence, if $\rho > 1$, then

$$(\sigma^2 n)^{-\frac{1}{2}} [W_n - n\mu] \Rightarrow N(0,1),$$

$$n^{-1} W_n \to \mu \quad \text{a.s.}, \tag{3.11}$$

and

$$\limsup \frac{W_n - n\mu}{(2\sigma^2 n \log\log n)^{\frac{1}{2}}} = 1 \quad \text{a.s.},$$

where $\mu = EY_1$ and $\sigma^2 = \text{Variance}\ (Y_1)$.

3.3.2 $\rho = 1$. When the traffic intensity is right at its critical value, the situation is more delicate. From (3.8), it is apparent that the limit behavior of $\{W_n\}$ is closely related to the limit behavior of $\{S_n\}$, but the limit behavior is not the same when $\rho = 1$. From (3.8), it is apparent that $W_n$ is a function of the initial segment $\{S_k, 0 \le k \le n\}$ and not just the single variable $S_n$. The desired connection between $\{W_n\}$ and $\{S_n\}$ relates $W_n$ or $\{W_k, 0 \le k \le n\}$ to $\{S_k, 0 \le k \le n\}$. This can be done by inducing for each $n$ appropriate stochastic processes in the space $D[0,1]$ of all right-continuous real-valued functions on $[0,1]$ with limits from the left everywhere. Including normalization, let

$$\underset{\sim}{W}_n \equiv \underset{\sim}{W}_n(t) = \frac{W_{\lceil nt \rceil}}{b_n}, \quad 0 \le t \le 1, \tag{3.12}$$

and

$$\underset{\sim}{S}_n \equiv \underset{\sim}{S}_n(t) = \frac{S_{\lceil nt \rceil}}{b_n}, \quad 0 \le t \le 1,$$

where $\lceil x \rceil$ is the greatest integer less than or equal to $x$. Note that the stochastic processes $\underset{\sim}{W}_n$ and $\underset{\sim}{S}_n$ in (3.12) are continuous-time processes with sample paths in the space $D[0,1]$. These processes conveniently represent normalizations of the initial segments $\{W_k, 0 \le k \le n\}$ and $\{S_k, 0 \le k \le n\}$ from the sequences $\{W_n\}$ and $\{S_n\}$. From (3.8), it is apparent that $\underset{\sim}{W}_n = f(\underset{\sim}{S}_n)$, where $f : D[0,1] \to D[0,1]$ is defined by

$$f(x)(t) = x(t) - \inf\{x(s), 0 \le s \le t\}, \quad 0 \le t \le 1. \tag{3.13}$$

The desired limit theorems follow from known <u>functional limit theorems</u> for $\underset{\sim}{S}_n$ such as Donsker's theorem (Theorem 16.1 of BILLINGSLEY (1968)) and <u>continuous mapping theorems</u> (Section 5 of BILLINGSLEY (1968)). Donsker's theorem asserts that $\underset{\sim}{S}_n$ converges weakly in $D[0,1]$ to Brownian motion, say $B$. Continuous mapping theorems assert that $f(U_n)$ converges to $f(U)$ if $U_n$ converges to $U$ and $f$ is continuous. (Further discussion appears in Section 3.4.) With an appropriate topology such as Skorohod's $J_1$ topology (induced by the metrics $d$ and $d_0$ in Chapter 3 of BILLINGSLEY (1968)), the function $f$ in (3.13) is continuous as is the projection

$\pi_1 : D[0,1] \to R$, defined by $\pi_1(x) = x(1)$. Consequently,

$$(\sigma^2 n)^{-\frac{1}{2}} W_n \Rightarrow f(B)(1) = PN(0,1),$$

$$n^{-1} W_n \to 0 \quad \text{a.s.},$$

(3.14)

$$\limsup_{n\to\infty} \frac{W_n}{(2\sigma^2 n \log\log n)^{\frac{1}{2}}} = 1 \quad \text{a.s.},$$

where $\sigma^2$ = Variance $(Y_1)$, $B$ is Brownian motion, and $PN(0,1)$ denotes the positive normal distribution, which is the normal distribution conditioned to be nonnegative. To be slightly more specific, the first limit in (3.14) holds because $\underset{\sim}{S}_n \Rightarrow B$, $(\sigma^2 n)^{-\frac{1}{2}} W_n = \pi_1 \circ f(\underset{\sim}{S}_n)$, and $\pi_1 \circ f : D[0,1] \to R$ is continuous.

It is interesting that the general theory of convergence for probability measures on more abstract spaces actually began with the limits in (3.14), cf. ERDOS and KAC (1946) or p. 199 of CHUNG (1968). The original object was to determine the asymptotic behavior of $(\sigma^2 n)^{-\frac{1}{2}} M_n$ when $EY_1 = 0$, but by (3.9) its distribution coincides with that of $(\sigma^2 n)^{-\frac{1}{2}} W_n$ for each $n$.

3.3.3 Techniques. I have already mentioned several important techniques for proving heavy traffic limit theorems. The heavy traffic limit theorems are proved by applying known limit theorems for related basic processes such as the random walk $\{S_n\}$. When $\rho > 1$, convergence together theorems are used, and when $\rho = 1$, continuous mapping theorems are used. For the continuous mapping theorems, it is essential to employ convergence in the function space context because the queueing processes are continuous images of more basic processes only when both processes are regarded as random elements of an appropriate function space. The necessary function space arguments are discussed by BILLINGSLEY (1968), VERVAAT (1972), and WHITT (1972c).

In the GI/G/1 queue with $\rho \geq 1$ it is natural to start proving heavy traffic limit theorems by focusing on the sequence of waiting times $\{W_n\}$ which is so closely related to the random walk $\{S_n\}$. This approach is described above and was used by PROHOROV (1963), VISKOV (1964), and WHITT (1968). The results for $\{W_n\}$ can then be used to obtain corresponding results for other processes. For example, the continuous-time virtual waiting time process $\{W(t), t \geq 0\}$

can be treated by means of a random time change. The queue length process $\{Q(t), t \geq 0\}$ can then be treated by showing that $W(t)/Q(t) \Rightarrow Ev$ as $t \to \infty$ if $\rho \geq 1$, cf. WHITT (1968).

Alternatively, it is possible to start with the total workload to enter the system up to time $t$, which is a random sum, and then use it to treat the virtual waiting time process, cf. HOOKE (1970), and WHITT (1971a). In this approach, the sequence $\{W_n\}$ can then be treated as an embedded sequence.

An entirely different approach based on the arrival and service counting processes was used by BOROVKOV (1965) and IGLEHART and WHITT (1970a) to treat multi-channel systems. The first heavy traffic limit theorems in this approach are for the queue length process. In each approach, limits for one queueing process are then used to obtain limits for the other queueing processes.

In order to obtain heavy traffic limit theorems for relatively complex queueing systems such as the GI/G/s queue or more general multi-channel systems, it is useful to construct modified systems which are easier to analyze directly than the original system but which behave the same in heavy traffic. The first modified system for studying multi-channel system was introduced by BOROVKOV (1965) and applied by IGLEHART and WHITT (1970a). It has each server remain in operation even when idle and assigns customers to the servers that can complete their service first, using residual service times. Another modified system designed for studying the virtual waiting time process in an s-server queue was introduced by LOULOU (1971, 1973a,b). If there are $s$ servers with only $k < s$ busy, it has all $s$ servers sharing the work of the $k$ customers. With this device, LOULOU (1971, 1973a,b) obtained heavy traffic limit theorems for the virtual waiting time process and the sequence of waiting times when $\rho > 1$ in an s-server queue, thus filling a gap in Section 6 of IGLEHART and WHITT (1970a). However, Loulou's modified system apparently cannot be used to treat as general multi-channel systems as Borovkov's modified system. For example, difficulty is encountered with several arrival channels.

Open Problem. LOULOU (1971, 1973a,b) filled a gap in Section 6 of IGLEHART and WHITT (1970a) with his modified system. However, it still remains to treat the queue length at the $i^{th}$ service channel when $\rho > 1$ (Section 5) as well as the load and the virtual waiting time process in general multiple channel systems (Section 6).

3.4 Functional Limit Theorems. All the limit theorems in this section can be extended to functional limit theorems. I have mentioned functional limit theorems in Section 3.3.2 as a useful tool for proofs. It is also desirable to have the final limit theorems be functional limit theorems whenever possible because such results imply many limit theorems for related functionals and processes as well as the ordinary limit theorem itself.

The idea is to replace the convergence of real-valued random variables with convergence of associated stochastic processes. For example, instead of (3.1) or the convergence

$$\frac{X(t) - at}{bt^c} \to L \quad \text{as} \quad t \to \infty , \tag{3.15}$$

the functional limit theorem typically states

$$\left\{ \frac{X(st) - ast}{bt^c} , \quad s \geq 0 \right\} \to \{L(s), \ s \geq 0\} \quad \text{as} \quad t \to \infty. \tag{3.16}$$

Obviously the modification in (3.16) is quite artificial, but it is useful. Since $st \to \infty$ as $t \to \infty$ for all $s > 0$, the convergence in (3.16) is not so different from (3.15), and yet it captures more of the asymptotic behavior of the entire process. The mode of convergence has not been specified in (3.15) or (3.16) because there are functional versions of all the classical limit theorems. For example, there are functional central limit theorems, functional laws of large numbers, and functional laws of the iterated logarithm. In particular, there are functional generalizations of all the results previously quoted in this section. Most of the papers I have mentioned state their conclusions in this way.

If the mode of convergence is convergence in distribution, then (3.16) involves convergence in distribution of a sequence of stochastic processes. This could mean convergence of all sequences of random variables obtained by evaluating the stochastic processes at single time points. More generally, this could mean convergence of all finite-dimensional distributions. The appropriate mode of convergence, weak convergence in the function space, turns out to be even more stringent. Weak convergence in the function space requires convergence of the stochastic processes regarded as measures on the space of all sample paths. This typically means convergence of all finite-dimensional distributions plus an additional condition on the fluctuations of the sample paths (tightness). This more general mode of convergence is important because it permits the continuous mapping

theorem to be applied with force. For example, limit theorems (and functional generalizations) for $W_n^\uparrow = \max\{W_k,\ 0 \le k \le n\}$ in heavy traffic are an easy consequence of functional limit theorems for $\{W_n\}$ and the continuous mapping theorem. It is only necessary to apply the supremum function, cf. VERVAAT (1972) and WHITT (1972c). If $\rho > 1$, $\{W_n^\uparrow\}$ has the same limit behavior as $\{W_n\}$, which was exhibited in (3.11). If $\rho = 1$, then

$$(\sigma^2 n)^{-1/2} W_n^\uparrow \Rightarrow \sup\{f(B)(t),\ 0 \le t \le 1\},$$

$$n^{-1} W_n^\uparrow \to 0, \text{a.s.}, \tag{3.17}$$

$$\limsup_{n \to \infty} \frac{W_n^\uparrow}{(2\sigma^2 n \log\log n)^{1/2}} = 1 \text{ a.s.},$$

where

$$P\{\sup\{f(B)(t)\} \le x\} = 1 - (4/\pi) \sum_{k=1}^{\infty} [(-1)^k/(2k+1)] \exp\{-[\pi^2 (2k+1)^2/8x^2]\}.$$

Functional generalizations of the extreme value theorems are also possible when $\rho < 1$. Even for the basic process, this is a fairly active area of research at present, cf. VERVAAT (1973) and WICHURA (1973).

Since functional central limit theorems often have Brownian motion as a limit process, it is important to be able to describe functionals of Brownian motion. Several of the useful functionals of Brownian motion are listed in Section 7 of IGLEHART (1972b).

I do not intend to dwell at great length on functional generalizations here. For further discussion, see BILLINGSLEY (1968), especially the Introduction; the survey papers by IGLEHART (1972b,c); and the paper by LOULOU (1973b) in these proceedings.

4. SYSTEM APPROXIMATIONS. The limit theorems for growing processes concerned a single stochastic process in a single queueing system. This single stochastic process was described by letting time get large. Sequences of stochastic processes were considered too, but only in the functional limit theorems and only as artificial constructions to obtain more general results for the single stochastic process under

consideration.

In this section, the limit theorems are for a sequence of queueing systems. Instead of describing the behavior of a single queueing process as time gets large, the object is to describe the behavior of a sequence of queueing processes associated with a sequence of queueing systems changing systematically. For example, the size of a finite waiting room may increase; the number of servers may increase, the traffic intensity may approach its critical value (heavy traffic), or the service time distributions may approach the exponential distribution. The corresponding limit theorems lead to approximations for queueing systems. Many of these approximations are approximations of entire queueing processes. This section just covers a few of the limit theorems of this kind which are possible. The wide range of possibilities is well illustrated for stochastic models in genetics by KARLIN and McGREGOR (1964).

4.1 The Poisson Tendency. The arrival process of a queueing system frequently can be represented quite well by a Poisson process. This phenomenon can be explained by a limit theorem. In many queueing systems the arrival process can be thought of as the sum of a large number of independent point processes of relatively small intensity. For example, each subscriber's use of a telephone might be described by a stationary renewal process or a more general point process. Such processes for different subscribers would typically be independent. Then the total demand on the telephone exchange is the sum of a large number of these processes. The appropriate limit theorem asserts that the superposition of independent uniformly sparse point processes converges to a Poisson process as the number of processes added becomes large with the intensity of each component process becoming asymptotically negligible. Such a theorem was proved by KHINTCHINE (1960). It is interesting that this theorem was proved with this queueing application in mind. Others had previously noted that such a limit theorem should hold, but Khintchine was the first to verify convergence of the finite-dimensional distributions. For further discussion, see Chapter 2 of GNEDENKO and KOVALENKO (1966). The mode of convergence can also be extended to weak convergence in the function space setting, cf. STRAF (1971) and WHITT (1971d).

An extensive account of research on point processes appears in LEWIS (1972). Research on superposition has been surveyed by ÇINLAR (1972). I have been concerned with determining how close the limiting Poisson process is to the superposition process. Bounds on the distance can be computed, where "distance" here means distance between

the stochastic processes regarded as measures on the function space, cf. WHITT (1972d). The specific distance used was the Prohorov distance for measures with the Skorohod distance on the function space D[0,1], cf. Section 5.1. While this distance has no unique claim to relevance, it does suggest that the rate of convergence is quite fast (of order $n^{-1}$ as compared with $n^{-1/2}$ in the central limit theorem). It is also possible to use these bounds to obtain bounds on the distance between various functionals of these processes. This would involve familiar distances for measures on the real line such as the Levy distance. Such related results can be obtained because rates of convergence as well as ordinary convergence are preserved by a large class of mappings, cf. WHITT (1972e).

4.2 Continuity. The standard queueing models are specified by the arrival process and the service times. For example, the standard multi-server system is specified by the sequence of interarrival times $\{u_n\}$ and the sequence of service times $\{v_n\}$. It is reasonable to expect that if these basic processes approximately satisfy certain assumptions, then the associated queueing processes will be close to the corresponding processes when the assumptions actually hold. For example, if $\{u_n\}$ and $\{v_n\}$ are approximately independent sequences of i.i.d. random variables with exponential distributions, the queue length process and the virtual waiting time process should be close to corresponding processes in the M/M/s queue. These hypotheses can be verified by showing that the queueing system is continuous. This means that the various sequences of queueing processes associated with the sequence of queueing systems converge if the sequences of basic processes converge. Such results were obtained for the single-server queue by KENNEDY (1972a) and extended to multi-server queues by WHITT (1971c, 1973b). The key to proving these theorems is to operate in the function space context. Then the desired convergence can be obtained by repeated application of the continuous mapping theorem.

From the point of view of providing a broad structural view of queueing processes, the continuity results are closely related to corresponding monotonicity results reported in JACOBS and SCHACH (1972), STOYAN (1972), and references there.

The continuity results can be used, not only to justify the use of models whose assumptions are only approximately satisfied, but also to generate new approximations. This is illustrated by the results of SCHASSBERGER (1970, 1972). He approximated the Laplace transform of the virtual waiting time process in the GI/G/1 queue and in the

preemptive-resume priority queue by considering a sequence of queueing systems in which the interarrival times and service times are mixtures of Erlang distributions.

Open Problem. The continuity results so far concern convergence of stochastic processes. As a next step, convergence of the limiting distributions and convergence of moments should be established. Such results will require extra conditions.

4.3 Heavy Traffic Approximations. Heavy traffic approximations can be obtained by considering a sequence of stable queueing systems which become unstable or heavily loaded in the limit. This usually means that the sequence of traffic intensities associated with the sequence of queueing systems is allowed to converge to the critical value from below. Thus, each queueing system in the sequence is stable, but the average number of customers waiting tends to grow as the systems change. After appropriate normalization, various associated sequences of queueing processes and random variables often converge to nondegenerate limits. Furthermore, these limits usually take a very simple form which is independent of the specific interarrival time and service time distributions. This is to be expected because when there are many customers in the system it is reasonable to disregard the detailed effect of each individual customer. The macroscopic view associated with statistical mechanics and the classical limit theorems becomes justified.

4.3.1 The Exponential Approximation. The first heavy traffic approximation was the exponential approximation for the stationary waiting time distribution in a single-server queue obtained by KINGMAN (1961, 1962, 1965). (An earlier result for dams was obtained by KENDALL (1957).) In the GI/G/1 setting of Section 3.2, let

$$\alpha = \frac{-2EY}{Var(Y)} = \frac{2(Eu - Ev)}{Var(u) + Var(v)}, \tag{4.1}$$

so that $\alpha > 0$ iff $\rho < 1$. Consider a one-parameter family of queueing systems indexed by $\alpha$. Kingman looked at the iterated limit, first letting $n \to \infty$ to obtain the steady-state waiting time $W(\alpha)$ for each $\alpha > 0$, and then letting $\alpha \downarrow 0$ after normalizing to obtain $\alpha W(\alpha) \Rightarrow E$ in R (E denotes an exponential random variable) or

$$\lim_{\alpha \downarrow 0} P\{\alpha W(\alpha) \le x\} = \begin{cases} 1 - e^{-x}, & x \ge 0 \\ 0, & x < 0, \end{cases} \tag{4.2}$$

under appropriate regularity conditions on the family of queueing

systems. There are several possible sets of regularity conditions, one of which requires

$$\text{Var}[Y_n(\alpha)] \to \sigma^2 \ , \ 0 < \sigma^2 < \infty \ , \quad \text{as} \quad \alpha \to 0 \ ,$$

(4.3)

$$\text{and} \qquad \sup_{\alpha}\{EY_n^{2+\varepsilon}(\alpha)\} < \infty \quad \text{for some} \quad \varepsilon > 0 \ .$$

The exponential approximation in (4.2) is perhaps best understood in the context of the closely-related diffusion approximation to be discussed in the next section.

Open Problem. The exponential approximations have not yet been verified in as great generality as the diffusion approximations. The same exponential approximation also holds for the continuous-time virtual waiting time process, cf. HOOKE (1969), KYPRIANOU (1971b), and WHITT (1974), but the exponential approximation has not yet been verified for the queue length process in a GI/G/1 queue. Such a result for an M/G/1 queue appears on p. 168 of GNEDENKO and KOVALENKO (1968).

4.3.2 Diffusion Approximations. Diffusion approximations can be obtained by considering a sequence of appropriately normalized queueing processes instead of a sequence of appropriately normalized steady-state distributions. In particular, if in addition to (4.1) and (4.3) there is a sequence of GI/G/1 queueing systems indexed by $n$ with $2^{-1}\sigma\alpha_n n^{1/2} \to c \ , \ 0 < c < \infty \ , \ \text{as} \ n \to \infty \ ,$ then

$$\underset{\sim}{W}_n(t) = (n\sigma^2)^{-1/2} W^n_{[nt]} \Rightarrow f(B - ce)(t) \quad \text{in} \quad R \tag{4.4}$$

for each $t > 0$, where $B$ is Brownian motion, $e$ is the identity map: $e(t) = t$, $f$ is the function in (3.13) corresponding to the barrier at 0, and

$$P\{f(B - ce)(t) \le x\} = \Phi\left(\frac{x+ct}{t^{1/2}}\right) - e^{-2cx}\Phi\left(\frac{-x+ct}{t^{1/2}}\right), \tag{4.5}$$

where $\Phi$ is the standard normal c.d.f. (Note that the normalization in (4.4) involves a time contraction as well as a division by $n^{1/2}$.) Moreover, $\underset{\sim}{W}_n$ converges weakly to $f(B - ce)$ in the function space $D[0,1]$ or $D[0,\infty)$. In fact, (4.4) is most easily proved via this more general weak convergence. The argument differs only slightly from the $\rho = 1$ case discussed in Section 3.3.2. If

$$\underset{\sim}{S}_n = \underset{\sim}{S}_n(t) = (n\sigma^2)^{-1/2} S^n_{[nt]} \ , \quad t \ge 0 \ , \tag{4.6}$$

with $S_k^n = Y_1^n + \ldots + Y_k^n$ being the $k^{th}$ partial sum in the $n^{th}$ queueing system, then $\underset{\sim}{S}_n$ converges weakly to B - ce by virtue of PROHOROV's (1956) double-sequence extension of Donsker's theorem, cf. p. 220 of PARTHASARATHY (1967). (The translation in the limit comes from the condition $2^{-1}\sigma\alpha_n n^{1/2} \to c$.) The weak convergence for $\underset{\sim}{W}_n$ in D and (4.4) are then obtained by applications of the continuous mapping theorem, first with f in (3.13) and then with the projection at time t. The limit in (4.4) was first obtained by PROHOROV (1963). Such diffusion approximations have now been obtained for all the standard processes in quite general queueing systems, cf. BOROVKOV (1965), IGLEHART and WHITT (1970b), and WHITT (1968, 1974).

From the discussion above, it should be clear that the general weak convergence method of proof applies equally well to a sequence of queueing systems with $\rho_n \to \rho \geq 1$ or a single queueing system with $\rho \geq 1$. The limits for a single queueing system are obviously a special case obtained by letting each queueing system in the sequence of queueing systems be the same. However, it usually works the other way too. When results can be obtained for a single unstable system with $\rho \geq 1$, corresponding results can usually be obtained for a sequence of queueing systems with $\rho_n \to \rho \geq 1$. Hence, even though the objectives may be different, it is appropriate to consider both the descriptions of unstable queues and the diffusion approximations for stable queues under a common name.

The exponential approximation is related to the diffusion approximation in an obvious way. As $t \to \infty$, the left side of (4.4) approaches a normalization of the steady-state waiting time while the right side approaches the exponential distribution. This informal explanation of the exponential approximation was given by KINGMAN (1962, 1965).

Open Problem. While the exponential approximations are easily related to the diffusion approximations, it still remains to use the diffusion approximations to prove exponential approximations. For practical purposes, the exponential approximations have been justified in great generality because (4.5) approaches the exponential distribution as t gets large, but a proof justifying the interchange of limits ($n \to \infty$ and $t \to \infty$) is still needed. For the special case of $\{W_n\}$

in a GI/G/1 queue, this interchange has been justified by PROHOROV (1963) using Kolmogorov's inequality and by WHITT (1974) using a new topology on a subset of $D[0,\infty)$ for the purpose of obtaining convergence of limiting distributions from weak convergence. More work needs to be done.

4.3.3 Many Servers. When there are infinitely many servers, heavy traffic approximations can be obtained by considering a sequence of queueing systems in which the arrival rates increase. Again, the average number of customers tends to get large and limit theorems are possible after appropriate normalization. For example, consider a sequence of GI/G/$\infty$ queues. Suppose $\{A(t), t \geq 0\}$ is a renewal process and the arrival counting process for the $n^{th}$ system is defined by $A_n(t) = A(nt)$, $t \geq 0$. If $Q_n(t)$ is the queue length in the $n^{th}$ GI/G/$\infty$ system and $\underset{\sim}{Q}_n$ is the associated normalized random function:

$$\underset{\sim}{Q}_n \equiv \underset{\sim}{Q}_n(t) = \frac{Q_n(t) - nh(t)}{n^{1/2}}, \quad t \geq 0, \tag{4.7}$$

then $\underset{\sim}{Q}_n$ converges weakly to a centered stationary Gaussian process, cf. BOROVKOV (1967a) and IGLEHART (1972c, 1973a). Of course, this means that $Q_n(t)$ is approximately normally distributed.

This infinite-server result also encompasses a limit theorem for s-server queues as $s$ gets large if the arrival rate is allowed to grow with $s$ so that the traffic intensity $\rho = Ev/sEu$ remains fixed. Such a theorem was first proved by IGLEHART (1965a) for M/M/s queues. It follows from the more general infinite server results because as $s$ grows with $\rho$ fixed the probability that all $s$ servers will be busy becomes negligible.

## 5. FURTHER HEAVY TRAFFIC RESEARCH.

5.1 Rates of Convergence. After proving limit theorems, it is of interest to obtain refinements such as asymptotic expansions and bounds on the rate of convergence. Such extensions of heavy traffic limit theorems were discussed to some extent by PROHOROV (1963), BOROVKOV (1964), and WHITT (1968, 1972a), but recently an extensive treatment has been given by KENNEDY (1972b). Kennedy has shown how bounds for rates of convergence can be obtained for virtually all the functional central limit theorems for queues in heavy traffic. These rates of convergence go beyond the usual results because, like the

rates discussed in Section 4.1, they are for a sequence of stochastic processes instead of a sequence of random variables. The rates can be expressed in terms of the Prohorov metric on the space $\mathcal{P}(S)$ of all probability measures on a separable metric space $(S,m)$. For any measurable subset $A$ in $S$, let

$$A^{\varepsilon} = \{x \in S : m(x,y) < \varepsilon \text{ for some } y \in A\} . \tag{5.1}$$

The Prohorov metric $d$ can then be defined for any $P, Q \in \mathcal{P}(S)$ by

$$d(P,Q) = \inf\{\varepsilon > 0 : P(F) \leq \varepsilon + Q(F^{\varepsilon}), \quad F \text{ closed}\} . \tag{5.2}$$

For example, consider a single GI/G/1 queue with $\rho = 1$ as in Section 3.3.2. Suppose

$$\underset{\sim}{W}_n \equiv \underset{\sim}{W}_n(t) = (n\sigma^2)^{-1/2} W_{[nt]}, \quad 0 \leq t \leq 1 . \tag{5.3}$$

Let $\mathcal{L}(\underset{\sim}{W}_n)$ denote the probability measure in $\mathcal{P}(D[0,1])$ induced by $\underset{\sim}{W}_n$. Then, under the assumption that $EY_n^5 < \infty$, there exists a constant $C$ such that

$$d(\mathcal{L}(\underset{\sim}{W}_n), \mathcal{L}(f(B))) \leq \frac{C \log n}{n^{1/4}} , \tag{5.4}$$

where the uniform metric is used on $D[0,1]$. (The processes $\underset{\sim}{W}_n$ and $B$ have paths in a separable subset.) The particular result in (5.4) follows directly from a rate of convergence theorem for Donsker's theorem, Theorem 5.2 of DUDLEY (1972), and the Lipschitz mapping theorem, Corollary 3.3 of WHITT (1972e). The result in (5.4) may seem a bit artificial because it is hard to attach intuitive meaning to the Prohorov metric. However, as a consequence of (5.4),

$$|P\{F(\underset{\sim}{W}_n) \leq x\} - P\{F(f(B)) \leq x\}| \leq \frac{C_F \log n}{n^{1/4}} \tag{5.5}$$

for every functional $F : D[0,1] \to R$ with

$$|F(x) - F(y)| \leq K \sup_{0 \leq t \leq 1} |x(t) - y(t)| \tag{5.6}$$

for some constant $K$, and

$$|P\{F(f(B)) \leq x+h\} - P\{F(f(B)) \leq x\}| \leq L|h| \tag{5.7}$$

for some constant $L$ and all $x$ and $h$ ; see Corollary 3.3 and Theorem 3.5 of WHITT (1972e). Kennedy's limit theorems have all been stated in the form (5.5), but they can also be expressed in the form

(5.4). Of course, bounds for the other processes treated by Kennedy involve much more work. For a survey of rates of convergence and metric representations of stochastic convergence, see DUDLEY (1972).

The bounds on the rate of convergence in KENNEDY's (1972b) theorems are typically of order

$$g(n,p) = \{(\log n)^p/n^{\min(p-1,\ p/2)}\}^{(2p+1)^{-1}}, \tag{5.8}$$

where $p = \min\{p_1, p_2\}$, based on assuming $Eu^{p_1} < \infty$ and $Ev^{p_2} < \infty$. The bound in (5.8) is always worse than the $n^{-1/4}(\log n)^{1/2}$ in (5.4). The slight improvement in (5.4) is due to the application of Theorem 5.2 of DUDLEY (1972) instead of the previous bound on the rate of convergence for Donsker's theorem established by ROSENKRANTZ (1967) and extended by HEYDE (1969). These bounds are not as good as $0(n^{-1/2})$ which is the classical Berry-Esseen bound associated with the ordinary central limit theorem. This seems to be due in part to the method of proof which is based on the Skorohod embedding method. SAWYER (1972) has shown that the Skorohod embedding method cannot in general yield rates of convergence faster than $0(n^{-1/4})$. By an entirely different method, NAGAEV (1970) has proved a theorem which implies that if $\rho = 1$ and $EY^3 < \infty$ in a GI/G/1 queue, then there exists a constant $C$ such that

$$\left|P\left\{\frac{W_n}{\sigma n^{1/2}} \le x\right\} - (2/\pi)^{1/2}\int_0^x e^{-y^2/y}dy\right| \le \frac{C}{n^{1/2}}. \tag{5.9}$$

Open Problem. The best bounds for rates of convergence in the heavy traffic limit theorems should be $0(n^{-1/2})$, but this remains to be determined. Weaker bounds may be necessary for functional central limit theorems, but this also remains to be determined. It certainly should be possible to establish bounds of order $0(n^{-1/2})$ in many ordinary heavy traffic limit theorems.

Open Problem. No bounds at all have been established for rates of convergence in many of the other limit theorems discussed in this paper. For example, bounds on the rates of convergence for the exponential approximation are still needed.

The Skorohod imbedding method, which plays a large role in the proofs of the rate of convergence theorems above, deserves some additional discussion. The idea is to represent the processes in a converging sequence of processes as random time transformations of the

limit process. Naturally, these random time transformations must approach the identity map in the limit. Such representations are useful because it often makes bounds on rates of convergence easier to estimate. The specific representation due to SKOROHOD (1965) is for an arbitrary random variable $X$ with mean zero and finite variance. Skorohod showed that it is always possible to find a stopping time $T$ for Brownian motion $B$ such that $B(T)$ has the same distribution as $X$. Then a random walk $X_1, X_1 + X_2, \ldots$ could be represented by $B(T_1), B(T_1 + T_2), \ldots$. The Skorohod imbedding method is discussed for example in FREEDMAN (1971). A recent survey has been prepared by SAWYER (1973).

5.2 Exponential Approximations for Multi-Server Queues. Until recently, exponential approximations had not been established for multi-server queues. KINGMAN (1965a) conjectured that the limit in (4.2) also holds for GI/G/s queues if, instead of (4.1),

$$\alpha = \frac{2[Eu - Ev/s]}{Var(u) + Var(v/s)} . \tag{5.10}$$

This conjecture has now been partially verified.

First, the modified system introduced by LOULOU (1971, 1973a,b) provides bounds above and below the process $\{L_n, n \geq 0\}$, where $L_n$ depicts the total workload facing all $s$ servers in a standard multi-server queue at the epoch just prior to the $n^{th}$ arrival. In particular, it is easy to show that

$$0 \leq L_n' \leq L_n \leq L_n' + s \max_{0 \leq k \leq n-1} \{v_k\} , \tag{5.11}$$

where $L_n'$ is the workload facing all $s$ servers in the modified system. In the modified system idle servers help busy servers so that the rate of service is always $s$ per time whenever any customers are in the system. The sequence $\{L_n'\}$ thus corresponds to the sequence of waiting times in a single-server queue determined by the sequences $\{su_n\}$ and $\{v_n\}$. Consequently, if the service times are bounded, there are bounds above and below the limiting distribution of $L_n$ as $n \to \infty$. Furthermore, the distance between these bounds becomes negligible under the usual heavy traffic normalizations. This means that known exponential approximations for the limiting distribution of $\{L_n'\}$ also apply to the limiting distribution of $\{L_n\}$. But note that in order to have a finite upper bound as $n \to \infty$ in (5.11), it is necessary to assume that the service times are bounded. For further

discussion, see WHITT (1974) and the paper by LOULOU (1973b) in these proceedings. (For extensions, see KÖLLERSTRÖM (1973)-added in proof.)

Exponential approximations for multi-server queues have also recently been obtained by YU (1972, 1973). Yu establishes two-sided stochastic bounds for the waiting times and other variables in a $GI/E_k/s$ queue. Different servers are even allowed to have different Erlang service distributions, but the shape parameter $k$ must be the same for all servers. The bounds involve corresponding variables from a $GI/E_k/1$ system with the same traffic intensity. For example, Yu shows that

$$W_n(1) - A \leq W_n(s) \leq W_n(1) + B , \tag{5.12}$$

where $W_n(r)$ is the waiting time of the $n^{th}$ customer in the $GI/E_k/r$ system, $A$ and $B$ are specific finite random variables independent of $n$, and $\leq$ denotes stochastic order, i.e., $X \leq Y$ means $P\{Y \geq t\} \geq P\{X \geq t\}$ for all $t$. It is easy to show that $A$ and $B$ do not grow in heavy traffic, so $W_n(s)$ behaves the same in heavy traffic as $W_n(1)$. Hence, the known exponential approximations for $W_n(1)$ can be applied to $W_n(s)$ in the $GI/E_k/s$ queue. The proofs of YU's (1972, 1973) stochastic order relations follow STIDHAM (1970). Naturally, the bounds are of interest in their own right without reference to heavy traffic. Recent references in addition to LOULOU (1971, 1973a,b) and YU (1972, 1973) from which a relatively complete picture of the literature on bounds and inequalities for queues can be obtained are BRUMELLE (1971, 1972), JACOBS and SCHACH (1972), KINGMAN (1970), ROSS (1973), STOYAN (1972), and SUZUKI and YOSHIDA (1970).

Open Problem. It still remains to verify that the exponential approximation is valid for all GI/G/s queues. Nothing at all has been said about exponential approximations for the more general multiple-channel systems of BOROVKOV (1965) and IGLEHART and WHITT (1970a,b).

5.3 Conditioning. It is frequently of interest to consider limit theorems in the presence of conditioning. For example, heavy traffic limit theorems have been proved for $\{W_n\}$ (and other processes) under the condition that the first busy period has not yet ended. Since visits to the origin are excluded, it is natural to expect larger limits with such conditioning than without, and this turns out to be the case.

When $\rho < 1$, convergence without normalization as $t \to \infty$ has been verified for various queueing processes in the M/G/1 and GI/M/1 queues by KYPRIANOU (1971a, 1972a) and the limits are called quasi-stationary distributions. KYPRIANOU (1972b) also proved heavy traffic limit theorems for these quasi-stationary distributions.

Just as in Section 4.3.1, consider a one-parameter family of queueing systems indexed by $\alpha$, where $\alpha$ is defined in (4.1). Let $Z(\alpha)$ be a random variable with the quasi-stationary distribution associated with the sequence of waiting times in the $\alpha$-system. In particular, let

$$P\{Z(\alpha) \le x\} = \lim_{n\to\infty} P\{W_n(\alpha) \le x | W_k(\alpha) > 0,\ 0 \le k \le n\} . \tag{5.13}$$

Then, for M/G/1 queues under appropriate regularity conditions,

$$\alpha Z(\alpha) \Rightarrow G \quad \text{as} \quad \alpha \downarrow 0 , \tag{5.14}$$

where $G$ has a Gamma distribution with density $\lambda^2 x e^{-\lambda x}$ and $\lambda = 2^{-1}$. The mean of the limit in (5.14) is exactly four times the mean of the limit without conditioning in (4.2).

Of course, no quasi-stationary limit exists when $\rho \ge 1$, so the object, just as in Section 3.3, must be to obtain limits for $\{W_n | W_k > 0,\ 0 \le k \le n\}$ as $n \to \infty$ after appropriate normalization. Such theorems have recently been proved by IGLEHART (1973). If $\rho > 1$, the conditioning does not alter the limit. If $\rho = 1$, then

$$\lim_{n\to\infty} P\left\{ \frac{W_n}{\sigma n^{1/2}} \le x | W_k > 0,\ 0 \le k \le n \right\} = \begin{cases} 1 - e^{-x^2/2}, & x \ge 0 \\ 0, & \end{cases} \tag{5.15}$$

The limit in (5.15) is the Rayleigh distribution which has mean $(\pi/2)^{1/2}$. Without conditioning, the limit is the positive normal distribution with mean $(2/\pi)^{1/2}$. The conditioned result is thus $\pi/2$ times bigger, which is a smaller factor than the previous 4 as expected because $\rho$ is bigger here.

IGLEHART (1973) also proved the functional generalization of (5.15). The limit process is a natural modification of standard Brownian motion. Look at the largest zero of the Brownian motion in $[0,1]$. Then consider the process from this zero onward rescaled so that this zero time point becomes the origin. The resulting process is the limit process.

Open Problem. It would appear that such conditioned heavy traffic limit theorems can be proved whenever unconditioned heavy traffic

limit theorems have been proved, but several gaps remain. KYPRIANOU's (1972b) results should apply to GI/G/1 queues. (Apparently Iglehart has just done this for $\{W_n\}$.) Functional central limit theorems leading to diffusion approximations should be proved as $\rho_n \uparrow 1$. Multi-server queues should be treated.

Open Problem. Heavy traffic limit theorems could also be proved under different kinds of conditioning.

5.4 Finite Dams. Heavy traffic limit theorems have been proved for finite dams and queues with finite waiting rooms by WHITT (1968, Sections 6.8 and 6.9), BLOMQVIST (1973), and KENNEDY (1973). In the case of a finite dam in discrete time, the content is represented as a random walk between two reflecting barriers. Thus, the content process corresponds to the sequence of waiting times in a queue with an extra barrier. Under the usual normalization (Section 4.3.2), the sequence of unrestricted random walks associated with a sequence of finite dams converges to Brownian motion, possibly with a drift. The two reflecting barriers can be superimposed afterwards. Convergence is preserved because these barriers constitute a continuous mapping on the function space. If the two barriers are put at 0 and a, then because of the normalization by $n^{1/2}$ in the content processes the upper barriers in the sequence of finite dams must be of order $0(n^{1/2})$. This normalization leads to weak convergence to Brownian motion with reflecting barriers at 0 and a and perhaps a drift.

With the upper barrier, steady-state or limiting distributions exist for the content process in each finite dam under all traffic intensities. As expected, there are exponential approximations for this limiting distribution. Detailed arguments appear in BLOMQVIST (1973) and KENNEDY (1973), but the exponential approximation can be understood by looking at it in relation to the diffusion approximation just discussed. The exponential approximation is just the limiting distribution of the diffusion process as $t \to \infty$. For example, let L be the limiting distribution of the Brownian motion with drift b and reflecting barriers at 0 and a. Then

$$P\{L \le x\} = \begin{cases} (1 - e^{2bx})/(1 - e^{2ab}) , & b \neq 0 \\ x/a , & b = 0 , \end{cases} \tag{5.16}$$

which coincides with the limit in Theorem 3.1 of KENNEDY (1973). The limit in (5.16) is easy to calculate directly, for example, by using the approach of MANDL (1968) discussed in Section 5.6.

Open Problem. It is easy to see that the two reflecting barriers must constitute a continuous mapping, but I know of no neat proof. Since weak convergence follows from other processes by the same continuous mapping theorem argument and since the form of the limit has been established by KENNEDY (1973), there is not much more to do for the diffusion approximations. Of course, the exponential approximation still needs to be verified for other processes and other systems. It should be apparent that a direct connection between the diffusion approximations and exponential approximations would solve numerous problems.

Open Problem. A general investigation of barriers as mappings on function spaces is still needed. The paper by LOYNES (1970) is somewhat in this spirit.

## 5.5 Other Queueing Systems

5.5.1 Different Server-Selection Rules. By now, heavy traffic limit theorems have been proved for quite a variety of queueing systems. In addition to the standard multi-channel system, there are the modified systems mentioned in Section 3.3.3 which were introduced primarily to facilitate proofs. Multi-channel queues in which arriving customers select a service channel at random or are assigned in rotation are treated in WHITT (1970a, 1973b). Unlike the modified systems which are designed to behave the same in heavy traffic, these variations do not perform as well in heavy traffic as the standard system. For example, if $W_n$ is the waiting time of the $n^{th}$ customer in a single GI/G/s queue with $\rho = 1$ and one of these server-selection disciplines, then in each case, just as in (3.14),

$$\frac{W_n}{\sigma n^{1/2}} \Rightarrow \text{PN}(0,1) \quad \text{as} \quad n \to \infty \,, \tag{5.17}$$

but the normalizing constant $\sigma$ increases from the standard model to the rotation-server-selection model to the random-server-selection model. Similar orderings are obtained for the means of the exponential approximations which can be derived for the stationary waiting time in each of these systems. These precise comparisons should be useful in estimating the relative efficience of different server-selection rules.

Open Problem. Still other systems should be investigated and compared. For example, the system in which each server has his own queue and customers join the shortest line remains to be described.

5.5.2 Priority Classes. One priority system, the single-server queue with a preemptive-resume discipline, has been rather extensively analyzed in heavy traffic. The first work on this system was done by HOOKE and PROBHU (1971) and HOOKE (1969, 1972a,b). They focused on the virtual waiting time process for lower priority customers in a single unstable system. Later HARRISON (1973) verified a diffusion approximation for the same process by considering a sequence of systems. Functional generalizations of these results and heavy traffic limit theorems for other processes are contained in WHITT (1971a,e).

The virtual waiting time process of lower priority customers can be defined in terms of a first passage time function as:

$$W(t) = \inf\{s \geq 0 : Y_h(s+t) - Y_h(t) < -L(t)\}, \quad t \geq 0, \tag{5.18}$$

where $L(t)$ is the total workload of both priorities facing the server at time $t$ and $Y_h(t)$ is the net input process for higher priority customers. For the weak convergence arguments, it is thus convenient to use the D-space of functions with a two-dimensional parameter set, cf. STRAF (1971). There are different limits depending on the traffic intensities of the two priority classes. If the traffic intensity $\rho_h$ of the higher priority customers is greater than one, then with positive probability lower priority customers will never be served at all. If $\rho_h = 1$, then a normalization by $t^2$ is needed in order to get a nondegenerate limit for $W(t)$.

5.5.3 Complex Systems. The possibility of proving heavy traffic limit theorems for queueing systems involving general networks a service facilities was indicated in Section 4 of IGLEHART and WHITT (1970b). It is usually not difficult to verify that a limit exists but it is usually difficult to evaluate the limit in detail. Unless surprising new developments occur, this approach does not appear to be the way to resolve or get around most of the problems in analyzing queueing networks described by DISNEY (1973).

Heavy traffic limit theorems have been proved for other complex systems. Assembly-like queues and queues in series have been studied by HARRISON (1970, 1971a,b) and networks of assembly-like queues and queues in mass-transportation systems have been studied by CRANE (1971).

Open Problem. Queues in computer systems should be, and no doubt will be, investigated in this way.

5.6 Control. The approximations and limit theorems should be used for designing and controlling queueing systems. The functional central limit theorems are promising in this regard because with them it is often possible to obtain convergence for various cost and control features in addition to the underlying probabilistic structures with an application of the continuous mapping theorem. This procedure is illustrated in the quality control setting by IGLEHART and TAYLOR (1968). The forthcoming thesis by RATH (1973) will treat queueing systems in the same way.

It is also possible to consider optimization problems directly for the approximations. Problems of controlling diffusion processes naturally arise corresponding to the various control problems for queues outlined by PRABHU and STIDHAM (1973) and SOBEL (1973) in these proceedings. Work by BATHER (1966, 1968), GIMON (1967), and PUTERMAN (1972, 1973) on diffusion models for dams and inventories is relevant. For example, a diffusion model for a queue with a removable server is easily treated by Puterman's approach. The diffusion process in this model is Brownian motion with a positive drift when the server is working and a negative drift when the server is not working. A reflecting barrier at the origin is included. A reasonable stationary policy switches the server on when the process reaches some level $S$ and switches the server off when the level reaches 0. With fixed costs for switching on and off plus a linear waiting cost, the long-run average cost and the limiting distribution as functions of $S$ and the drift and diffusion coefficients are computed in WHITT (1973d). The proof, following PUTERMAN (1972, 1973), is an immediate application of the general control theory for diffusion processes in MANDL (1968), in particular Theorem 1 on p. 149. Long-run average costs, expected discounted costs, and limiting distributions can be calculated for such models by solving second-order linear differential equations related to the generator of the diffusion process.

The deterministic version of the diffusion models above (corresponding to zero variances) was analyzed by WHITT (1973c). For the deterministic version, it is easy to show that stationary $(s,S)$ policies are optimal among all switching policies and that the long-run average cost is a convex function of $s$ and $S$ if the waiting or holding and shortage costs are convex and increasing away from the origin ($s$ is allowed to be negative in the inventory model).

Open Problem. Obviously, there is much to be done here. For the diffusion models, more general controls than switching should be

considered. But, even among switching policies, the optimality of stationary policies with special structure still needs to be verified.

## REFERENCES

In addition to the sources mentioned in the paper, this bibliography includes all the work I have seen on queues in heavy traffic. There are a few references on related topics (e.g., other approximations for queues and other limit theorems for queueing processes) but I have made no attempt to be complete outside of heavy traffic.


[1] ARJAS, E. and DE SMIT, J. H. A. (1973) On the total waiting time during a busy period of the single server queue. Discussion Paper No. 7312, Center for Operations Research and Econometrics, Louvain, Belgium.

[2] AVI-ITZHAK, B. (1971) Heavy traffic characteristics of a circular data network. Bell System Tech. J. 50 2521-2549.

[3] BATHER, J. A. (1966) A continuous time inventory model. J. Appl. Prob. 3 538-549.

[4] _____ (1968) A diffusion model for the control of a dam. J. Appl. Prob. 5 55-71.

[5] _____ (1969) Diffusion models in stochastic control theory. J. Roy. Stat. Soc. Ser. A. 132 335-352.

[6] BILLINGSLEY, P. (1968) Convergence of Probability Measures. John Wiley and Sons, New York.

[7] _____ (1971) Weak Convergence of Measures: Applications in Probability. Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

[8] BLOMQVIST, N. (1973) A heavy traffic result for the finite dam. J. Appl. Prob. 10 223-228.

[9] BLOOMFIELD, P. and COX, D. R. (1972) A low traffic approximation for queues. J. Appl. Prob. 9 832-840.

[10] BOROVKOV, A. (1964) Some limit theorems in the theory of mass service, I. Theor. Probability Appl. 9 550-565.

[11] _____ (1965) Some limit theorems in the theory of mass service, II. Theor. Probability Appl. 10 375-400.

[12] _____ (1967a) On limit laws for service processes in multi-channel systems. Siberian Math. J. 8 746-763.

[13] _____ (1967b) Convergence of weakly dependent processes to the Wiener process. Theor. Probability Appl. 12 159-174.

[14] _____ (1967c) On the convergence to diffusion processes. Theor. Probability Appl. 12 405-431.

[15] _____ (1972) Stochastic Processes in the Theory of Mass Service. Science Publishers, Moscow (in Russian).

[16] BRODY, S. (1963) On a limit theorem in the theory of mass service. Ukrain. Mat. Zh. 15 76-79 (in Russian).

[17] BRUMELLE, S. L. (1971) Some inequalities for parallel server queues. Opns. Res. 19 402-413.

[18] _____ (1973) Bounds on the wait in a GI/M/k queue. Man. Sci. 19 773-777.

[19] CHERNOFF, H. (1968) Optimal stochastic control. Sankhyā 30 221-252.

[20] CHUNG, K. L. (1968) A Course in Probability Theory. Harcourt, Brace and World, New York.

[21] ÇINLAR, E. (1972) Superposition of point processes. Stochastic Point Processes. Ed. P. A. W. Lewis, John Wiley and Sons, New York, 549-606.

[22] COHEN, J. W. (1969) The Single Server Queue. North Holland Publ. Co., Amsterdam.

[23] _____ (1972) Asymptotic relations in queueing theory. Stoch. Proc. Appl. 2 107-124.

[24] CRANE, M. A. (1971) Limit theorems for queues in transportation systems Ph.D. thesis and TR No. 16, Department of Operations Research, Stanford University.

[25] _____ and IGLEHART, D. L. (1973) Simulating stable stochastic systems, I-III. J. Assoc. Comput. Mach. To appear.

[26] DALEY, D. J. (1969) The total waiting time in a busy period of a stable single server-queue, I. J. Appl. Prob. 6 550-564.

[27] _____ and JACOBS, D. R., Jr. (1969) The total waiting time in a busy period of a stable single-server queue, II. J. Appl. Prob. 6 565-572.

[28] DARLING, D. (1956) The maximum of sums of stable random variables. Trans. Amer. Math. Soc. 83 164-169.

[29] DISNEY, R. L. (1973) Some Topics in Queueing Network Theory. These proceedings.

[30] DUDLEY, R. M. (1972) Speeds of metric probability convergence. Z. Wahrscheinlichkeitstheorie verw. Geb. 22 323-332.

[31] ERDÖS, P. and KAC, M. (1946) On certain limit theorems in the theory of probability. Bull. Amer. Math. Soc. 52 292-302.

[32] FAHADY, K. S., QUINE, M. P., and VERE-JONES, D. (1971) Heavy traffic approximations for the Galton-Watson process. Adv. Appl. Prob. 3 282-300.

[33] FISHMAN, G. (1972a) Output analysis for queueing simulations. TR No. 56, Department of Administrative Sciences, Yale University.

[34] _____ (1972b) Estimation in multiserver queueing simulations. TR No. 58, Department of Administrative Sciences, Yale University.

[35] _____ (1973) Statistical analysis in multiserver queueing simulations. TR No. 64, Department of Administrative Sciences, Yale University.

[36] FREEDMAN, D. (1971) Brownian Motion and Diffusion. Holden-Day, San Francisco.

[37] GAFARIAN, A. V., MUNJAL, P. K., and PAHL, J. (1971) An experimental validation of two Boltzmann-type statistical models for multi-lane traffic flow. Transp. Res. 5 211-224.

[38] GAVER, D. P., Jr. (1968) Diffusion approximations and models for certain congestion problems. J. Appl. Prob. 5 607-623.

[39] _____ (1969) Highway delays resulting from flow-stopping incidents. J. Appl. Prob. 6 137-153.

[40] _____ (1971) Analysis of remote terminal backlogs under heavy demand conditions. J. ACM. 18 405-415.

[41] _____ and SHEDLER, G. S. (1973) Processor utilization in multi-programming systems via diffusion approximations. Opns. Res. 21 569-576.

[42] GIMON, J. G. (1967) A continuous time inventory model. TR No. 29, Department of Statistics, Stanford University.

[43] GNEDENKO, B. and KOVALENKO, I. (1968) Introduction to Queueing Theory. Israel Program for Scientific Translations, Ltd., Jerusalem.

[44] HARRISON, J. M. (1970) Queueing models for assembly-like systems. Ph.D. thesis and TR No. 7, Department of Operations Research, Stanford University.

[45] _____ (1971a) Assembly-like queues. Graduate School of Business, Stanford University. (J. Appl. Prob. 10 354-367)

[46] _____ (1971b) The heavy traffic approximation for single server queues in series. TR No. 24, Department of Operations Research, Stanford University. (J. Appl. Prob. 10 613-629)

[47] _____ (1973) A limit theorem for priority queues in heavy traffic. J. App. Prob. 10. To appear.

[48] HEATHCOTE, C. (1965) Divergent single server queues. Proc. of the Symp. on Congestion Theory. Eds. W. Smith and W. Wilkinson, The University of North Carolina Press, Chapel Hill, 108-129.

[49] _____ and WINER, P. (1969) An approximation for the moments of waiting times. Opns. Res. 17 175-186.

[50] HEYDE, C. C. (1967) A limit theorem for random walks with drift. J. Appl. Prob. 4 144-150.

[51] _____ (1969a) On extended rate of convergence results for the invariance principle. Ann. Math Statist. 40 2178-2179.

[52] _____ (1969b) On the maximum of sums of random variables and the supremum functional for stable processes. J. Appl. Prob. 6 419-429.

[53] _____ (1970) On some mixing sequences in queueing theory. Opns. Res. 18 312-315.

[54] _____ (1971) On the growth of the maximum queue length in a stable queue. Opns. Res. 19 447-452.

[55] _____ and SCOTT, D. J. (1969) A weak convergence approach to some limit results with mixing which have applications in queueing theory. Australian National University.

[56] HEYMAN, D. P. (1974) An approximation for the busy period of the M/G/1 queue using a diffusion model. J. Appl. Prob. 11. To appear.

[57] _____ (1972) Diffusion approximations for congestion models. Bell Telephone Laboratories, Holmdel, New Jersey.

[58] HOOKE, J. A. (1969) Some Limit Theorems for Priority Queues. Ph.D. thesis and TR No. 91, Department of Operations Research, Cornell University.

[59] _____ (1970) On some limit theorems for the GI/G/1. J. Appl. Prob. 7 634-640.

[60] _____ (1972a) A priority queue with low-priority arrivals general. Opns. Res. 20 373-380.

[61] _____ (1972b) Some heavy-traffic limit theorems for a priority queue with general arrivals. Opns. Res. 20 381-388.

[62] _____ and PRABHU, N. U. (1971) Priority queues in heavy traffic. Opsearch 8 1-9.

[63] IGLEHART, D. L. (1965a) Limit diffusion approximations for the many-server queue and the repairman problem. J. Appl. Prob. 2 429-441.

[64] _____ (1965b) Limit theorems for queues with traffic intensity one. Ann. Math Statist. 36 1437-1449.

[65] _____ (1967) Diffusion approximations in applied probability. Lectures in Applied Mathematics, Vol. 12: Mathematics of the Decision Sciences, Part 2, 234-254.

[66] _____ (1969) Diffusion approximations in collective risk theory. J. Appl. Prob. 6 285-292.

[67] _____ (1971a) Multiple channel queues in heavy traffic, IV: law of the iterated logarithm. Z. Wahrscheinlichkeitstheorie verw. Geb. 17 168-180.

[68] _____ (1971b) Functional limit theorems for the queue GI/G/1 in light traffic. Adv. Appl. Prob. 3 269-281.

[69] _____ (1972a) Extreme values in the GI/G/1 queue. Ann. Math. Statist. 43 627-635.

[70] _____ (1972b) Weak convergence in applied probability. TR No. 26, Department of Operations Research, Stanford University.

[71] _____ (1972c) Weak convergence in queueing theory. TR No. 27, Department of Operations Research, Stanford University.

[72] _____ (1973a) Weak convergence of compound stochastic processes. Stochastic Processes and Their Applications 1 11-31.

[73] _____ (1973b) Functional central limit theorems for random walks conditioned to stay positive. TR No. 28, Department of Operations Research, Stanford University.

[74] _____ and KENNEDY, D. P. (1970) Weak convergence of the average of flag processes. J. Appl. Prob. 7 747-753.

[75] IGLEHART, D. L. and LEMOINE, A. J. (1972) Approximations for the repairman problem with two repair facilities, I and II. TR No. 266-2, 4, Control Analysis Corporation, Palo Alto, California.

[76] IGLEHART, D. L. and TAYLOR, H. M. (1968) Weak convergence of a sequence of quickest detection problems. Ann. Math. Statist. 39 2149-2153.

[77] IGLEHART, D. L. and WHITT, W. (1970a) Multiple channel queues in heavy traffic, I. Adv. Appl. Prob. 2 150-177.

[78] _____ (1970b) Multiple channel queues in heavy traffic, II: sequences, networks and batches. Adv. Appl. Prob. 2 355-369.

[79] JACOBS, D. R., Jr. and SCHACH, S. (1972) Stochastic order relationships between GI/G/k systems. Ann. Math. Statist. 43 1623-1633.

[80] KARLIN, S. and McGREGOR, J. (1964) On some stochastic models in genetics. Stochastic Models in Medicine and Biology, ed. J. Gurland, University of Wisconsin Press, Madison, 245-279.

[81] KENDALL, D. G. (1957) Some problems in the theory of dams. J. Roy. Stat. Soc. Ser. B 19 207-212.

[82] _____ (1964) Some recent work and further problems in the theory of queues. Theor. Probability Appl. 9 1-13.

[83] KENNEDY, D. P. (1972a) The continuity of the single server queue. J. Appl. Prob. 9 370-381.

[84] _____ (1972b) Rates of convergence for queues in heavy traffic, I and II. Adv. Appl. Prob. 4 357-391.

[85] _____ (1973) Limit theorems for finite dams. Stochastic Processes and Their Applications 1 269-278.

[86] KHINTCHINE, A. Y. (1960) Mathematical Methods in the Theory of Queueing. Griffin, London.

[87] KINGMAN, J. F. C. (1961) The single server queue in heavy traffic. Proc. Camb. Phil. Soc. 57 902-904.

[88] _____ (1962) On queues in heavy traffic. J. Roy. Statist. Soc., Ser. B 25 383-392.

[89] _____ (1965a) The heavy traffic approximation in the theory of queues. Eds. W. Smith and W. Wilkinson, Proc. of the Symp. on Congestion Theory. The University of North Carolina Press, Chapel Hill, 137-159.

[90] _____ (1965b) Approximations for queues in heavy traffic. Queueing Theory: Recent Developments and Applications. Ed. R. Cruon, Elsevier Publishers, New York.

[91] _____ (1970) Inequalities in the theory of queues. J. Roy. Stat. Soc., Ser. B 32 102-110.

[92] KYPRIANOU, E. K. (1971a) On the quasi-stationary distribution of virtual waiting time in queues with Poisson arrivals. J. Appl. Prob. 8 494-507.

[93] _____ (1971b) The virtual waiting time of the GI/G/1 queue in heavy traffic. Adv. Appl. Prob. 3 249-268.

[94] _____ (1972a) On the quasi-stationary distribution of the GI/M/1 queue. J. Appl. Prob. 9 117-128.

[95] _____ (1972b) The quasi-stationary distributions of queues in heavy traffic. J. Appl. Prob. 9 821-831.

[96] LALCHANDANI, A. (1967) Some limit theorems in queueing theory. Ph.D. thesis and TR No. 29, Department of Operations Research, Cornell University.

[97] LE GALL, P. (1962) Les systemes avec ou sans attente et les processus stochastiques. Tome I, Dunod. (in French).

[98] LEWIS, P. A. W. (1972) Stochastic Point Processes: Statistical Analysis, Theory, and Applications. John Wiley and Sons, New York.

[99] LOULOU, R. J. (1971) Weak convergence for multichannel queues in heavy traffic. Ph.D. thesis and TR No. 71-31, Operations Research Center, College of Engineering, University of California, Berkeley.

[100] _____ (1973a) Multichannel queues in heavy traffic. J. Appl. Prob. 10. To appear.

[101] _____ (1973b) On the extension of some heavy-traffic theorems to multiple-channel systems. These proceedings.

[102] LOYNES, R. (1962) The stability of a queue with non-independent interarrival and service times. Proc. Camb. Phil. Soc. 58 497-520.

[103] _____ (1965) Extreme values in uniformly mixing stationary stochastic processes. Ann. Math. Statist. 36 993-999.

[104] _____ (1970) Stopping times on Brownian motion: some properties of Root's construction. Z. Wahrscheinlichkeitstheorie verw. Geb. 16 211-218.

[105] MANDL, P. (1968) Analytic Treatment of One-Dimensional Markov Processes. Springer-Verlag, Berlin and New York.

[106] MAXUMDAR, S. (1970) On priority queues in heavy traffic. J. Roy. Stat. Soc. Ser. B. 32 111-114.

[107] NAGAEV, S. V. (1970) On the speed of convergence in a boundary problem, I, II. Theor. Probability Appl. 15 179-199 and 419-441.

[108] NEWELL, G. F. (1965) Approximation methods for queues with application to the fixed-cycle traffic light. SIAM Review 7 223-239.

[109] _____ (1968a) Queues with time-dependent arrival rates; I. The transition through saturation. J. Appl. Prob. 5 436-451.

[110] _____ (1968b) Queues with time-dependent arrival rates; II, The maximum queue and the return to equilibrium. J. Appl. Prob. 5 579-590.

[111] _____ (1968c) Queues with time-dependent arrival rates; III: A mild rush hour. J. Appl. Prob. 5 591-606.

[112] _____ (1971) Applications of Queueing Theory. Chapman and Hall, London.

[113] _____ (1973) Approximate stochastic behavior of n-server service systems with large n. Lecture Notes in Economics and Mathematical Systems, No. 87. Springer-Verlag, Berlin, Heidelberg, and New York.

[114] PARTHASARATHY, K. R. (1967) Probability Measures in Metric Spaces. Academic Press, New York.

[115] PIPES, L. A. (1968) Topics in the hydrodynamic theory of traffic flow. Transp. Res. 2 143-149.

[116] PLEDGER, G. and SERFLING, R. (1971) The waiting time of a vehicle in queue. TR No. M203, Department of Statistics, Florida State University.

[117] PRABHU, N. U. (1968) Some new results in storage theory. J. Appl. Prob. 5 452-460.

[118] _____ (1969) The simple queue in non-equilibrium. Opsearch 6 118-128.

[119] _____ (1970a) The queue GI/M/1 with traffic intensity one. Studia Sci. Math. Hung. 5 89-96.

[120] _____ (1970b) Limit theorems for the single server queue with traffic intensity one. J. Appl. Prob. 7 227-233.

[121] _____ and STIDHAM, S., Jr. (1973) Optimal control of queueing systems. These proceedings.

[122] PRESMAN, E. (1965) On the waiting time for many-server queueing systems. Theor. Probability Appl. 10 63-73.

[123] PROHOROV, Yu. (1956) Convergence of random processes and limit theorems in probability theory. Theor. Probability Appl. 1 157-214.

[124] _____ (1963) Transient phenomena in processes of mass service. Litovsk. Mat. Sb. 3 199-205. (in Russian)

[125] PUTERMAN, M. L. (1972) On the optimal control of diffusion processes. Ph.D. thesis and TR NO. 14, Department of Operations Research, Stanford University.

[126] _____ (1973) A diffusion process model for a production facility. Graduate School of Business, University of Massachusetts. To appear in Man. Sci.

[127] RATH, J. (1973) Limit theorems for controlled queues. Ph.D. thesis, Department of Operations Research, Stanford University.

[128] ROSENKRANTZ, W. (1967) On rates of convergence for the invariance principle. Trans. Amer. Math. Soc. 129 542-552.

[129] ROSS, S. M. (1973) Bounds on the delay distribution in GI/G/1 queues. Operations Research Center, University of California, Berkeley.

[130] SAMADAROV, E. (1963) Service systems in heavy traffic. Theor. Probability Appl. 8 307-309.

[131] SAWYER, S. (1972) Rates of convergence for some functionals in probability. Ann. Math. Statist. 43 273-284.

[132] _____ (1973a) The Skorohod Representation. Department of Mathematics, Belfer Graduate School, Yeshiva University, New York.

[133] SCHASSBERGER, R. (1970) On the waiting time in the queueing system GI/G/1. Ann. Math. Statist. 41 182-187.

[134] _____ (1972) On the work load process in a general preemptive resume priority queue. J. Appl. Prob. 9 588-603.

[134a] _____ (1973b) Forthcoming book, Springer, Berlin. (in German)

[135] SKOROHOD, A. V. (1965) Studies in the Theory of Random Processes. Addison-Wesley, Reading, Massachusetts.

[136] SMITH, W. L. (1955) Regenerative stochastic processes. Proc. Roy. Soc. A 232 6-31.

[137] _____ (1958) Renewal theory and its ramifications. J. Roy. Stat. Soc., Ser. B. 20 243-302.

[138] SOBEL, M. J. (1973) Optimal operation in queues. These proceedings.

[139] SPEED, T. P. (1973) Some remarks on a result of Blomqvist. J. Appl. Prob. 10 229-232.

[140] STIDHAM, S., Jr. (1970) On the optimality of single-server queueing systems. Opns. Res. 18 708-732.

[141] STOYAN, D. (1972) Monotonicity in stochastic models. ZAMM 52 23-30. (in German)

[142] STRAF, M. L. (1971) Weak convergence of stochastic processes with several parameters. Proc. Sixth Berk. Symp. Math. Stat. Prob. 2 187-221.

[143] SUZUKI, T. and YOSHIDA, Y. (1970) Inequalities for many-server queues and other queues. J. Opns. Res. Soc. of Japan 13 59-77.

[144] TAKACS, L. (1967) Combinatorial Methods in the Theory of Stochastic Processes. John Wiley and Sons, New York.

[145] _____ (1973) Occupation time problems in the theory of queues. These proceedings.

[146] TOMKO, J. (1972) The rate of convergence in limit theorems for service systems with finite queueing capacity. J. Appl. Prob. 9 87-102.

[147] VERVAAT, W. (1972) Functional central limit theorems for processes with positive drift and their inverses. Z. Wahscheinlichkeitstheorie verw. Geb. 23 245-253.

[148] _____ (1973) Limit theorems for sample maxima and record values: a review. To appear.

[149] VISKOV, O. (1964) Two asymptotic formulae in the theory of mass service. Theor. Probability Appl. 9 177-178.

[150] _____ and PROHOROV, Yu. (1964) The probability of loss of calls in heavy traffic. Theor. Probability Appl. 9 99-104.

[151] WHITT, W. (1968) Weak convergence theorems for queues in heavy traffic. Ph.D. thesis, Department of Operations Research, Cornell University and TR. 2, Department of Operations Research, Stanford University.

[152] _____ (1970a) Multiple channel queues in heavy traffic, III: random server selection. Adv. Appl. Prob. 2 370-375.

[153] _____ (1970b) A guide to the application of limit theorems for sequences of stochastic processes. Opns. Res. 18 1207-1213.

[154] _____ (1971a) Weak convergence theorems for priority queues: preemptive-resume discipline. J. Appl. Prob. 8 74-94.

[155] _____ (1971b) Classical limit theorems for queues. Department of Administrative Sciences, Yale University.

[156] _____ (1971c) The continuity of queues. Revision to appear in Adv. Appl. Prob. 6.

[157] _____ (1971d) Representation and convergence of point processes on the line. To appear in Ann. Prob.

[158] _____ (1971e) Heavy traffic limit theorems for priority queues.

[159] _____ (1972a) Complements to heavy traffic limit theorems for the GI/G/1 queue. J. Appl. Prob. 9 185-191.

[160] _____ (1972b) Embedded renewal processes in the GI/G/s queue. J. Appl. Prob. 9 650-658.

[161] _____ (1972c) Continuity of several functions on the function space D. To appear in Ann. Prob.

[162] _____ (1972d) On the quality of Poisson approximations. To appear in Z. Wahrscheinlichkeitstheorie verw. Geb.

[163] _____ (1972e) Preservation of rates of convergence under mappings. To appear in Z. Wahr.

[164] _____ (1973a) A converse to the continuous mapping theorem for composition.

[165] _____ (1973b) A broad structural analysis of multi-server queueing models.

[166] _____ (1973c) Deterministic inventory and production models.

[167] _____ (1973d) A diffusion model for a queue with removable server.

[168] _____ (1973e) Exponential heavy traffic approximations for multi-server queues.

[169] WICHURA, M. J. (1973) On the functional form of the law of the iterated logarithm for the partial maxima of independent identically distributed random variables. Department of Statistics, University of Chicago.

[170] WORTHINGTON, F. (1967) Ladder processes in continuous time. Ph.D. thesis and TR No. 34, Department of Operations Research, Cornell University.

[171] YU, O. S. (1972) On the steady-state solution of a GI/$E_k$/r queue with heterogeneous servers. TR No. 38, Department of Operations Research, Stanford University.

[172] _____ (1973) Stochastic bounding relations for a heterogeneous-server queue with Erlang service times. Stanford Research Institute. Menlo Park, California.

## ADDITIONAL REFERENCES


[173] BARTFAI, P. (1970) Limsuperior theorems for the queueing model. Studia Sci. Math. Hung. 5 317-325.

[174] KÖLLERSTRÖM, J. (1973) Heavy traffic theory for queues with several servers, I. Department of Mathematics, The University of Oxford.

[175] LEMOINE, A. J. (1972) Delayed random walks and limit theorems for generalized single server queues. Ph.D. thesis and Technical Report No. 21, Department of Operations Research, Stanford University.

[176] _____ (1973) Limit theorems for generalized single server queues: the exceptional system. Department of Mathematics, Clemson University.

[177] McNEIL, D. R. (1973) Diffusion limits for congestion models. J. Appl. Prob. 10 368-376.

[178] STOYAN, D. (1972) Continuity theorem for single-server queues. Math. Operationsforsch. Statist. 3 103-111.

[179] _____ (1973) A continuity theorem for queue size. Bull. Acad. Sci. Polon. To appear.

[180] _____ (1974) Some bounds for many-server systems. Math. Operationsforsch. Statist. To appear.

[181] STOYAN, H. (1973) Monotonicity and continuity properties of many server queues. Math. Operationsforsch. Statist. 4 155-163.

[182] ROLSKI, T. and STOYAN, D. (1973) Two classes of semi-orderings and their application to queueing theory. ZAMM. To appear.