General Stochastic Processes in the Theory of Queues

by VÁCLAV E. BENEŠ

General Stochastic Processes in the Theory of Queues

by VÁCLAV E. BENEŠ
Bell Telephone Laboratories

ADDISON-WESLEY PUBLISHING COMPANY, INC.
READING, MASSACHUSETTS • PALO ALTO • LONDON

Library of Congress Catalog Card No. 68-7768

Author's Preface

One of the welcome features of applied mathematics is that it is in a position to appeal to at least two audiences, the mathematicians and the engineers. But when an author tries to present in one work simultaneously a general, rigorous mathematical theory for a given applied topic, and an account of it that will be understandable and useful to practical engineers, he must risk losing both his intended audiences. A mathematician may boggle at the "physical" and "practical" emphases, while an engineer may be left quite bewildered by the theoretical niceties. Nevertheless, if it is accomplished, a simultaneous presentation to both audiences is unquestionably a valuable and challenging task. In this monograph I attempt such a task for the topic of delays in queueing systems with one server.

Delays in queues with one server and order of arrival service are considered without any restrictions on the statistical character of the offered traffic. Elementary methods establish formulas and equations describing probabilities of delay. These methods deemphasize special statistical models and yield a general theory. In spite of the generality of this approach, intuitive proofs and extensive explanations of the physical significance of formulas are given, as well as rigorous derivations. The theory is applied to specific models to obtain illustrative new results. Under mild conditions of stationarity, the asymptotic behavior in time of the delay is studied and is shown to be governed by a functional equation closely analogous to the "fundamental equation" of branching processes, already used in special queueing models. A generalization of the Pollaczek-Khinchin formula is derived for the case in which delays do not build up.

So many monographs, surveys, and books on the theory of queues are currently appearing that I have made no effort to canvass the vast extant literature of queueing. References to it

have been included only insofar as they arose naturally in the text. For the benefit of the reader, therefore, I cite the following books:

J. Riordan, *Stochastic Service Systems*. New York: John Wiley and Sons, 1962.

D. R. Cox and W. L. Smith, *Queues*. New York: John Wiley and Sons, 1961.

T. L. Saaty, *Elements of Queueing Theory*. New York: McGraw-Hill Book Co., 1961.

L. Takács, *Introduction to the Theory of Queues*. New York: Oxford University Press, 1962.

I would like to express my gratitude to Bell Telephone Laboratories for providing a milieu in which advanced theoretical work on practical topics can be pursued, and for supplying all the secretarial work involved in completion of the manuscript. Also, it is a pleasure to acknowledge that a careful reading of the manuscript by my colleague E. Wolman resulted in many corrections and improvements.

Murray Hill, New Jersey V. E. B.
November 5, 1962.

Contents

CHAPTER 1

VIRTUAL DELAY

1. INTRODUCTION

Congestion theory is the study of mathematical models of service systems, such as telephone central offices, waiting lines, and trunk groups. It has two practical uses: first, to provide engineers with specific mathematical results, curves, and tables, on the basis of which they can design actual systems; and second, to establish a general framework of concepts into which new problems can be fitted, and in which current problems can be solved. Corresponding to these two uses, there are two kinds of results: *specific results* pertaining to special models, and *general theorems*, valid for many models.

Most of the present literature of congestion theory consists of specific results resting on particular statistical assumptions about the traffic in the service system under study. Indeed, few results in congestion theory are known which do not depend on *special* statistical assumptions, such as negative exponential distributions, or independent random variables. In this monograph we describe some mathematical results which are free of such restrictions, and so constitute general theorems. These results concern general stochastic processes in the theory of queues with one server and order-of-arrival service.

In this work we have three aims: (1) to describe a new general approach to certain queueing problems; (2) to show that this approach, although quite general, can nevertheless be presented in a relatively elementary way, which makes it widely available; and (3) to illustrate how the new approach yields specific results,

both new and known. What follows is written only partly as a contribution to the mathematical analysis of congestion. It is also, at least initially, a frankly tutorial account aimed at increasing the public understanding of congestion by first steering attention away from special statistical models, and obtaining a general theory. Such a point of view, it is hoped, will yield new methods in problems other than congestion.

When a general theory can be given, it will be useful in several ways. It will (i) increase our understanding of complex systems; (ii) yield new specific results, curves, tables, etc; and (iii) extend theory to cover interesting cases which are known to be inadequately described by existing results. At first acquaintance, the theorems of such a general theory may not resemble "results" at all; that is, they may not seem to be facts which one could obviously and easily use to solve a real problem. A general theory is really a tool or principle, expressing the essence or structure of a system; properly explained and used, this tool will yield formulas and other specifics with which problems can be treated.

2. THE SYSTEM TO BE STUDIED

There is a queue in front of a single server, and the waiting customers are served in order of arrival, with no defections from the queue. We are interested in the waiting-time of customers.

As a mathematical idealization of the delays to be suffered in the system, we use the virtual waiting-time $W(t)$, which can be defined as the time a customer would have to wait for service if he arrived at time t. $W(\cdot)$ is continuous from the left; at epochs of arrival of customers, $W(\cdot)$ jumps upward discontinuously by an amount equal to the service-time of the arriving customer; otherwise $W(\cdot)$ has slope -1 while it is positive. If it reaches zero, it stays equal to zero until the next jump.

It is usual to define the stochastic process $W(\cdot)$ in terms of the arrival epoch t_k and the service-time S_k of the kth arriving customer, for $k = 1, 2, \ldots$. However, the following procedure is a little more elegant; we describe the service-times and the arrival epochs simultaneously by a single function $K(\cdot)$, which is defined for $t \geq 0$, left-continuous, nondecreasing, and constant between successive jumps. The locations of the jumps are the epochs of arrivals, and the magnitudes are the service-times. It is conven-

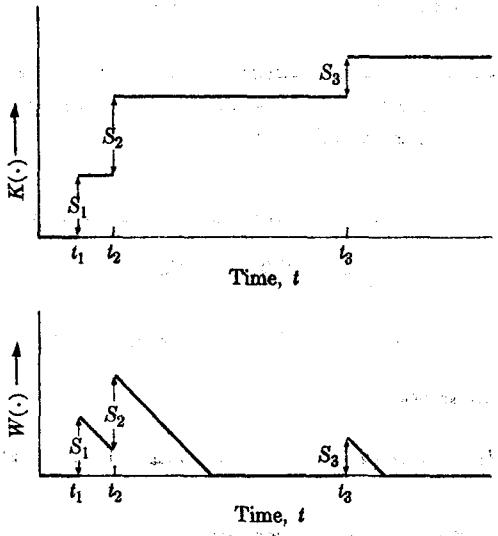

FIG. 1. The load $K(\cdot)$ and the virtual delay $W(\cdot)$. At the epoch t_k of arrival of the kth customer, $W(\cdot)$ jumps upward discontinuously by an amount equal to S_k, the service-time of the kth customer; otherwise, $W(\cdot)$ has slope -1 if it is positive; if it reaches zero it stays equal to zero until the next jump of the load function $K(\cdot)$.

ient to define $K(\cdot)$ to be continuous from the left, except at $t = 0$, where it is continuous from the right. The functions $W(\cdot)$ and $K(\cdot)$ are depicted simultaneously in Fig. 1.

If $K(t)$ is interpreted as the work offered to the server in the interval $[0, t)$, then $W(t)$ can be thought of as the amount of work remaining to be done at time t. In terms of this interpretation, it can be seen that

Work remaining at t = total work load offered up to t
− elapsed time
+ total time during which server was idle in $(0, t)$.

Then formally, $W(\cdot)$ is defined in terms of $K(\cdot)$ by the integral equation

$$W(t) = K(t) - t + \int_0^t U[W(u)]\, du, \qquad t \geq 0, \tag{1}$$

where $U(t)$ is the unit step function, that is, $U(x) = 1$ for $x \geq 0$,

and $U(x) = 0$ otherwise.* For simplicity we have set $W(0) = K(0)$.

It is possible to give an explicit solution of Eq. (1) in terms of $K(\cdot)$ and the supremum functional. This is the content of the following result of E. Reich [1].†

Lemma 1.1. If $K(x) - x$ has a zero in $(0, t)$, then

$$W(t) = \sup_{0<x<t} \{K(t) - K(x) - t + x\}.$$

If $K(x) - x > 0$ for $x \in (0, t)$, then $W(t) = K(t) - t$.

Proof. Let us set

$$z(t) = \sup \{u: 0 < u < t \quad \text{and} \quad W(u) = 0\}.$$

Then

$$\begin{aligned} W(t) &= K(t) - K[z(t)] - t + z(t) \\ &\leq \sup_{0<x<t} \{K(t) - K(x) - t + x\}. \end{aligned}$$

On the other hand, for $0 < x < t$ Eq. (1) gives

$$\begin{aligned} W(t) &= W(x) + K(t) - K(x) + \int_x^t U[-W(u)]\, du - t + x \\ &\geq K(t) - K(x) - t + x. \end{aligned}$$

Lemma 1.1 provides an explicit characterization of the important event $\{W(t) = 0\}$ purely in terms of $K(\cdot)$ and the supremum functional, as follows.

Lemma 1.2. $W(t) = 0$ if and only if

$$K(t) \leq t, \tag{2}$$

and

$$\sup_{0<x<t} \{K(t) - K(x) - t + x\} \leq 0. \tag{3}$$

---

* Formulas are numbered sequentially in each chapter. Thus (1) refers to the first formula of the current chapter, and (4.2) refers to the second formula of Chapter 4, etc.

† Numbers in brackets are keyed to the references at the end of the book.

Proof. Suppose that $W(t) = 0$. Then Eq. (1) implies

$$K(t) - t + \int_0^t U[-W(u)]\,du = 0,$$

so that $K(t) \leq t$. Also, as in Lemma 1.1, for $0 < x < t$,

$$\begin{aligned} W(t) &= W(x) + K(t) - K(x) + \int_x^t U[-W(u)]\,du - t + x \\ &\geq K(t) - K(x) - t + x. \end{aligned}$$

Hence $W(t) = 0$ implies

$$\sup_{0<x<t} \{K(t) - K(x) - t + x\} \leq 0.$$

Conversely, assume the conditions (2) and (3) of Lemma 1.2.

Case 1. $K(x) = x$ for some $x \in (0, t)$. Then by Lemma 1.1,

$$W(t) = \sup_{0<x<t} \{K(t) - K(x) - t + x\} \leq 0.$$

Since $W(\cdot)$ is nonnegative, we have $W(t) = 0$.

Case 2. $K(x) = x$ for no $x \in (0, t)$. Then by Lemma 1.1,

$$W(t) = K(t) - t \leq 0.$$

We note that Lemma 1.2 takes into account the initial value $W(0) = K(0)$, as it should.

Lemma 1.1 may be interpreted physically in the following manner. The quantity in braces $\{K(t) - K(x) - t + x\}$ is, if positive, the *excess* of arriving load in the interval $[x, t)$ over the elapsed time $t - x$; it is therefore the *overload* in $[x, t)$. Reich's formula then says, essentially, that

$$\text{Delay at } t = \sup_{0<x<t} \{\text{overload in } [x, t)\}.$$

The relationship between the waiting-time $W(\cdot)$ and the offered traffic $K(\cdot)$ can be further elucidated graphically by reference to Fig. 2. The light solid line shows $K(t) - t$, the traffic offered up to time t minus the traffic that could have been served if the server had been kept busy throughout the interval $(0, t)$.

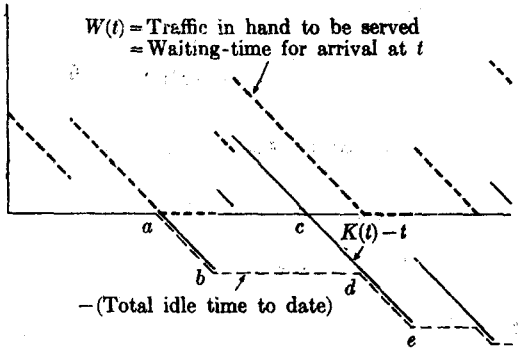

FIG. 2. Relationship between waiting-time and offered traffic.

It is assumed, in Fig. 2, that the server starts busy at $t = 0$. It is busy until $t = a$. At this point the server becomes idle, and $K(t) - t$ turns negative; its negative value is the negative of the idle time. At $t = b$, more traffic is offered, and $K(t) - t$ jumps up.

The heavy dashed line represents the waiting-time at t, $W(t)$; $W(t)$ can also be thought of as the traffic in hand and yet to be served at time t. This line can never be negative. It is equal to $K(t) - t$ before a and is zero from a to b. At b it jumps up, remaining above and parallel to $K(t) - t$ until $t = d$, when the server becomes idle again. $W(t)$ is above $K(t) - t$ by exactly the amount by which $K(t) - t$ was most negative at b. At d, when $W(t)$ reaches zero, $K(t) - t$ is just reaching its previous local infimum, and $K(d) - d = K(b) - b$.

During the interval (d, e), $W(\cdot)$ remains at zero, and $K(t) - t$ becomes more negative, establishing new local infima as it goes, and building up more idle time. At $t = e$, $K(t) - t$ and $W(\cdot)$ both jump up. Again $W(\cdot)$ is parallel to $K(t) - t$, but is now above it by an amount equal to the negative of the last infimum, $K(e) - e$.

In Fig. 2,

$$\inf_{a<x<t} [K(x) - x]$$

is shown as a light dashed line. It is a monotone, nonincreasing function of t, and is the negative of the total idle time up to time t. To account for the period $t < a$, when $K(t) - t$ has not yet become negative, and the server has not yet been idle, we write

$$W(t) = K(t) - t - \min \{0, \inf_{0<x<t} [K(x) - x]\},$$

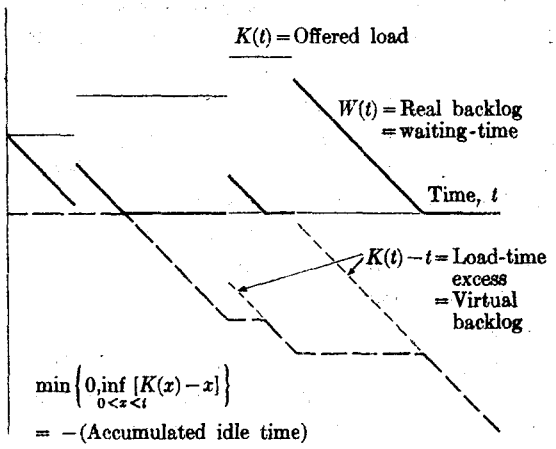

FIG. 3. Relationship among offered load, waiting-time, negative of accumulated idle time, and "load-time excess."

and thus obtain another representation for the delay; i.e., a solution of Eq. (1).

In a manner similar to that of Fig. 2, Fig. 3 depicts, simultaneously, the offered load $K(\cdot)$ in a light solid line, the waiting-time $W(\cdot)$ in a heavy solid line, the negative of the accumulated idle time in a heavy dashed line, and the "load-time excess" $K(t) - t$ in a light dashed line, when it does not coincide with the negative of the idle time. The terminology in Fig. 3 has been purposely chosen to suggest an interpretation in terms of inventory or storage theory. $W(t)$ is the *real backlog* (of orders, say), $K(t)$ is the *cumulative amount ordered*, and $K(t) - t$ might be termed the load-time excess or the *virtual backlog*. Then

$$\text{Real backlog} = \text{virtual backlog} + \text{accumulated idle time.}$$

3. CHARACTER OF THE RESULTS TO BE PROVEN

Models of waiting lines usually contain explicit assumptions about the statistical nature of the offered load $K(\cdot)$. For instance, the simplest models amount to assuming that the interarrival times $(t_n - t_{n-1})$ are all independent, with the same negative exponential distribution; a similar assumption is made for the

service-times. These assumptions give a class of models parametrized by the means of the negative exponential distributions.

A broader class of models is specified by retaining the assumptions that the interarrival times be independent and identically distributed (and similarly for service-times), but allowing any distribution, not just the negative exponential. The interarrival and service-time distributions may still be said to "parametrize" this broader class of models, since their choice determines a model in the class. In the papers of Khinchin [2], Kendall [3], Bailey [4], Takács (5), and Beneš [6], the arrivals form a Poisson process, and the service-times are independent of each other and of the arrival process. In the work of Smith [7] and Lindley [8], it is assumed that the interarrival-times and the service-times are (independent) renewal processes. The references just cited are merely representative; we make no attempt to give an adequate bibliography of the subject.

Indeed, the literature of applied probability theory contains many investigations of waiting-times (for one server); however, these studies have depended essentially on assumptions of statistical independence or special distributions. Many useful and interesting results have been obtained under these assumptions, which probably include most cases of practical interest. We believe, though, that the assumptions have tended to obscure the stochastic process of interest (the waiting-time) with analytical detail, since it is not always possible to separate the essential features of the stochastic process from those which only reflect the strength and analytic nature of the hypotheses.

As assumptions more general even than those of Lindley [8] are considered, it becomes extremely laborious to specify the model first, and then compute interesting quantities, such as distributions of delay, probabilities of loss, etc. So instead of looking for ways of exactly characterizing the model, we can try to search directly for simple ways of expressing the quantities of interest in terms of the model. Since the probability

$$\Pr\{W(t) \leq w\} \tag{4}$$

is what we actually wish to compute from the model, the question arises whether this calculation can be made without first specifying the entire probabilistic structure of $K(\cdot)$. The following intuitive argument can be adduced for answering "yes." $W(\cdot)$ is defined in terms of the load $K(\cdot)$ by a very special relationship, expressed

in the integral equation (1); hence, no matter what are the statistical features of $K(\cdot)$, it is likely that the distribution of $W(\cdot)$ depends only on some very particular, physically interpretable, statistical functions associated with $K(\cdot)$. It is not obvious that such an economy can be made in the generality we desire.

A principal result [formulas (2.4) and (2.5) or (3.16) and (3.17)], described in later sections, states that the probability (4) can, in fact, be given a fairly simple expression which is generally valid for any load process. This expression depends only on two special functions obtainable from the statistical structure of the load $K(\cdot)$. Each function has a definite intuitive or physical significance, given later. These statistical functions achieve the desired economy of description because we can state that the desired probabilities depend only on the features of $K(\cdot)$ expressed in the functions. For the purposes of calculating (4), we do not need the entire probabilistic structure of $K(\cdot)$, but only a relatively small relevant part.

From a theoretical viewpoint, the assumptions made in the literature have been inadmissibly strong. For indeed, Eq. (1) defines a transformation of a stochastic process $K(\cdot)$ of service-times and arrival epochs into another stochastic process $W(\cdot)$ of waiting-times. For each t, there is an operator or formula which gives the distribution of $W(t)$ in terms of suitable fundamental statistical functions associated with $K(u)$ for $u \leq t$. *The principal problem is to find the form of the operator and the character of the fundamental functions.* The answer to this problem should depend only on the integral equation (1) and on the fact that $K(\cdot)$ is a nondecreasing step function. It should depend on no special features of the probability measure for $K(\cdot)$ except those implied by this last property.

Some inkling of the nature of this answer can be given briefly here. The operator we seek is linear and operates only on the distribution of $K(t) - t$, and, for each $u \leq t$, on the conditional distribution of $K(t) - K(u) - t + u$ relative to the knowledge that

$$K(u) - u \leq 0, \quad \text{and} \quad \sup_{0<y<u} [K(u) - K(y) - u + y] \leq 0.$$

Accordingly, the present work involves (at first) no assumptions of independence and no special distributions. We shall assume at first that $K(\cdot)$ is a random, nondecreasing step function; its *only* statistical peculiarity is that it is a nondecreasing step function.

Another way of putting the problem we have attempted to solve is as follows: For general load processes $K(\cdot)$, what is a small amount of information about the statistical nature of $K(\cdot)$ with which one can nevertheless compute $\Pr\{W(t) \leq w\}$ for all t and w, and how is this calculation to be made? The required statistical functions are then the *information about* $K(\cdot)$; the formula for the probability (4) indicates the *method of calculation.*

One approach to the waiting-times, used really in all the papers cited previously, is to solve the Kolmogorov equations for the distributions of a Markov process. Since our assumptions do not necessarily give rise to a Markov process, this approach is not sufficiently general to solve our problem.

Another possibility is to first solve Eq. (1) and then try to express the distribution of the solution $W(t)$ in terms of the probability measure for $K(u)$, $u \leq t$. However, the solution of (1) involves the supremum functional, and so, although it is adequate in some cases, this approach incurs directly the notorious difficulties associated with the distribution of a supremum.

Our approach is, in a sense, an inversion of the usual method described above. The latter consists in first doing probability theory to set up Kolmogorov or renewal equations, and then doing analysis to solve the equations. We can, however, achieve greater generality by taking maximum advantage of the fact that the process $W(\cdot)$, of interest already, satisfies Eq. (1). In Chapter 2, we do this by a careful analysis of $W(\cdot)$ itself, by performing, in effect, some of our analysis in the domain of random functions and taking averages only at convenient points. In Chapter 3, our procedure is as follows: We first obtain a representation of the random variable $\exp\{-sW(t)\}$, $\mathrm{Re}\,(s) > 0$; the expectation of this variate is the Laplace-Stieltjes transform of the distribution of $W(t)$. From this expectation, we derive a formula for $\Pr\{W(t) \leq w\}$ by inversion, and the formula expresses the functional which we seek.

The present work is not a complete monograph on queues with one server. Rather, it is an account of principal results deduced by methods that are relatively new in queueing theory. Several of these results have been included because they show that the structure of the problems in the most general case is the same as in the special cases considered to date by means of Markov processes. Our effort to dispense with assumptions of independence and special distributions was originally stimulated by the