Lecture Notes
in Economics and
Mathematical Systems

Operations Research, Computer Science, Social Science

Edited by M. Beckmann, Providence, G. Goos, Karlsruhe, and
H. P. Künzi, Zürich

87

G. F. Newell

# Approximate Stochastic Behavior of n-Server Service Systems with Large n

Springer-Verlag
Berlin · Heidelberg · New York 1973

Dr. Gordon F. Newell
University of California
Institute of Transportation and
Traffic Engineering
109 McLaughlin Hall
Berkeley, CA 94720/USA

AMS Subject Classifications (1970): 60K25, 60K30, 62M10, 90B99, 94A20

ISBN-13: 978-3-540-06366-7 e-ISBN-13: 978-3-642-65651-4
DOI: 10.1007/978-3-642-65651-4

# PREFACE

For many stochastic service systems, service capacities large enough to serve some given customer demand is achieved simply by providing multiple servers of low capacity; for example, toll plazas have many toll collectors, banks have many tellers, bus lines have many buses, etc. If queueing exists and the typical queue size is large compared with the number $n$ of servers, all servers are kept busy most of the time and the service behaves like some "effective" single server with mean service time $1/n$ times that of an actual server. The behavior of the queueing system can be described, at least approximately, by use of known results from the much studied single-channel queueing system. For $n \gg 1$, however, (we are thinking particularly of cases in which $n \gtrsim 10$), the system may be rather congested and quite sensitive to variations in demand even when the average queue is small compared with $n$. The behavior of such a system will, generally, differ quite significantly from any "equivalent" single-server system.

The following study deals with what, in the customary classification of queueing systems, is called the $G/G/n$ system; $n$ servers in parallel with independent service times serving a fairly general type of customer arrival process. The arrival rate of customers may be time-dependent; particular attention is given to time dependence typical of a "rush hour" in which the arrival rate has a single maximum possibly exceeding the capacity of the service.

The methods of analysis exploit a postulate that $n \gg 1$, that all relevant counts of customers are made on a scale which is also large compared with 1 (typically of order $n$), and that stochastic fluctuations in arrival counts of order $n$ are of order $n^{1/2}$. Graphs of cumulative counts of customer arrivals and available servers are used to represent the evolution of stochastic realizations of the system. A combination of graphical and analytic methods are then used to estimate queue distributions for various typical types of behaviors.

If the traffic intensity (arrival rate/service rate), $\rho(t)$, increases toward a maximum, the system behaves like an $\infty$-channel system until $1 - \rho(t)$ is of order

$n^{-1/2}$ . When $\rho(t)$ come sufficiently close to 1 or exceeds 1 so that queueing becomes virtually certain, the system then behaves essentially like some effective single-channel service system. While $\rho(t)$ is between these two extremes, when it is uncertain whether or not a queue exists, several different types of behavior can exist depending upon how long the system stays in this transition state, and how $\rho(t)$ behaves during this time.

Chapter I describes the graphical representations and the postulates upon which the approximations will be based. Chapter II deals mostly with situations in which the above transition lasts for a time which is small compared with an expected service time. This is particularly appropriate for systems such as bus routes in which the service time (trip time) is comparable with the duration of the rush hour. In this case the distribution of the number of customers in the system (server plus the queue) remains approximately normally distributed at all times with appropriate time-dependent means and variances.

Chapter III treats the cases in which the transition lasts for a time large compared with the service time. The behavior through the transition can, for the most part, be described by "diffusion approximations." This is described very qualitatively because it is quite similar to previously analysed behaviors of a single-channel service system. Chapter IV deals with extremely slow changes in $\rho(t)$ , equilibrium distributions. Approximate distributions are compared with exact distributions for the M/M/n and M/D/n systems (Poisson arrivals and either exponentially distributed or deterministic service times).

Despite the length of this report, it contains very few detailed results. It describes mostly a classification of types of behaviors and the qualitative properties of these types, and methods which could be used to obtain more quantitative results. It would have been desirable to add a few "case studies" to test the accuracy, but, as yet, none have been made. The only numerical comparisons are with the known equilibrium distributions mentioned above. It would appear, however, that for $n \gtrsim 10$ , any person skilled in the art of making rough calculations should be able to estimate average lengths of queues, delays, etc., to within about 10% in most

situations he is likely to encounter; this is the level of accuracy needed in most design applications.

Most of the preliminary studies upon which the following is based were made over a period of several years while I was trying to teach transportation engineering students some uses of mathematical methods of analysis. The style clearly reflects the consequences of attempting to show both theoreticians and "hard-nosed" engineers how to find useful answers to real problems. I am indebted to my colleagues and students of diverse backgrounds for forcing me to address my remarks to all of them simultaneously.

This work was not directly supported by any research grants, however, student support subsidiary to this work has been financed by the National Science Foundation under Grants GP 9323 and GP 24617. Miscellaneous expenses have also been financed in part by a grant from General Motors Corporation. Most of the following was composed while I was on sabbatical leave January-July 1972 in residence at Union College, Schenectady, N. Y. I am indebted to Union College, particularly to Professor Gilbert Harlow and the Department of Civil Engineering, for providing space for me and a couple of students, and a pleasant atmosphere in which to work.

The final editing was done at Berkeley; the typing was done by Phyllis De Fabio in the Institute of Transportation and Traffic Engineering.

CONTENTS