AD-A038 666 CALIFORNIA UNIV BERKELEY OPERATIONS RESEARCH CENTER F/G 12/2
THE EFFECT OF SERVICE TIME REGULARITY ON SYSTEM PERFORMANCE.(U)
MAR 77 R W WOLFF AF-AFOSR-3213-77
UNCLASSIFIED ORC-77-7 NL

1 OF 1
AD A038666

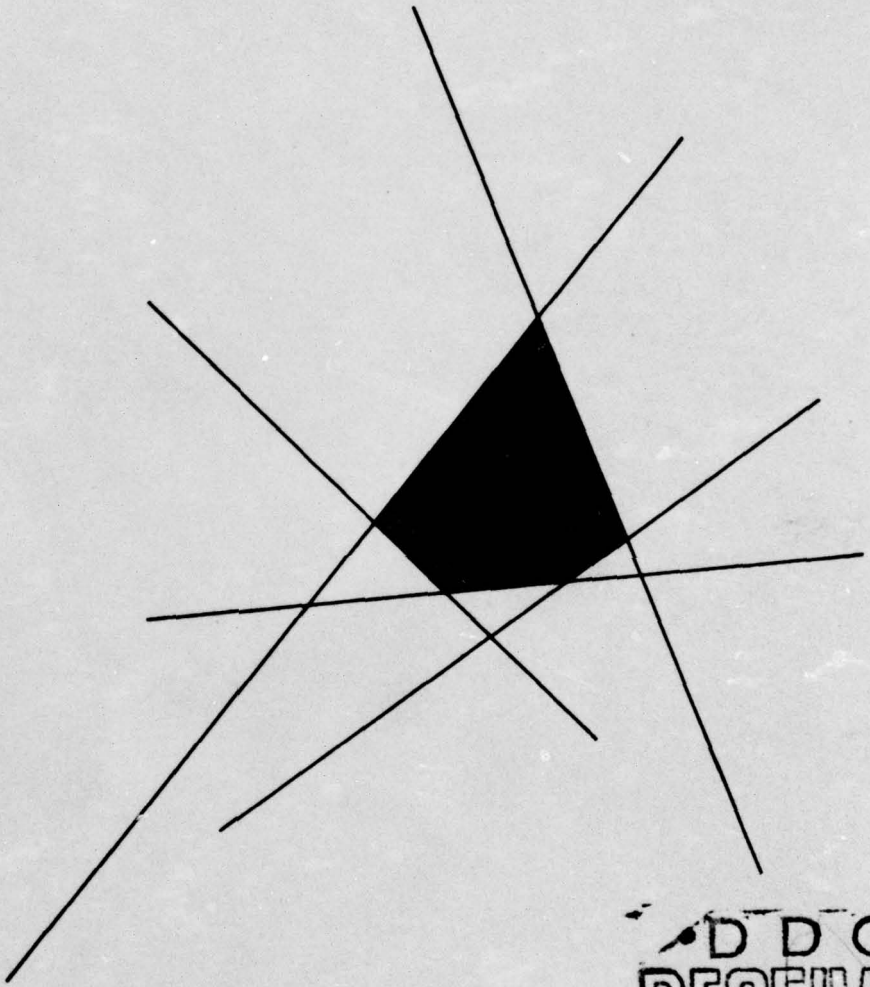
END

DATE FILMED 5-77

ORC 77-7
MARCH 1977

THE EFFECT OF SERVICE TIME REGULARITY ON SYSTEM PERFORMANCE

by
RONALD W. WOLFF



OPERATIONS RESEARCH CENTER

UNIVERSITY OF CALIFORNIA · BERKELEY

THE EFFECT OF SERVICE TIME REGULARITY ON SYSTEM PERFORMANCE[†]

by

Ronald W. Wolff[††]
Department of Industrial Engineering
and Operations Research
University of California, Berkeley

MARCH 1977 ORC 77-7

[†]Partially supported by the Air Force Office of Scientific Research, AFSC, USAF, under Grant AFOSR-77-3213.

[††]A portion of this research was completed while the author was a visiting consultant to the Operations Research Projects Department, Bell Telephone Laboratories, Holmdel, N. J. 07733

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(14) REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ORC-77-7	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): (6) THE EFFECT OF SERVICE TIME REGULARITY ON SYSTEM PERFORMANCE.		5. TYPE OF REPORT & PERIOD COVERED: (9) Research Report.
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) Ronald W. Wolff		8. CONTRACT OR GRANT NUMBER(s): AFOSR-77-3213 new
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Operations Research Center, University of California, Berkeley, California 94720		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: United States Air Force, Air Force Office of Scientific Research, Bolling AFB, D.C. 20332		(11) 12. REPORT DATE: Mar[illegible]77; 13. NUMBER OF PAGES: 18
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (12) 19p.		15. SECURITY CLASS. (of this report): Unclassified; 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

(15) AF-AFOSR-3213-77

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

(16) 2304 (17) A5

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Queue
Service Time
Variance
Performance Measures

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(SEE ABSTRACT) next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

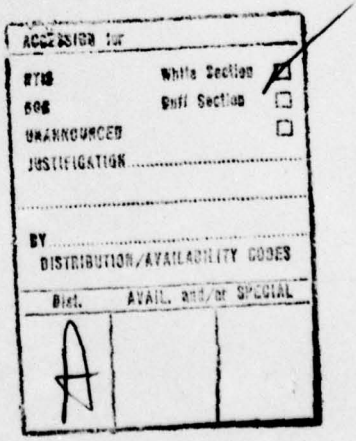
Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

270 750

JB

ABSTRACT

Conventional wisdom holds that the more regular the arrival process and/or service times are, the better system performance will be. Examples of contrary behavior are presented in three cases: loss systems, processor sharing, and multi-channel queues without losses. In each case, it is shown that making service times more regular can make system performance worse.



THE EFFECT OF SERVICE TIME REGULARITY ON SYSTEM PERFORMANCE

by

Ronald W. Wolff

0. INTRODUCTION

In undersaturated queues, queueing occurs solely because of the stochastic variation of the arrival process and/or service times. Thus, it is generally believed that the more regular (in some appropriate sense) each of these processes is, the better any of the usual performance measures will be.

This "conventional wisdom" is very useful when true. For example, we may be able to improve system performance when the arrival and/or service processes are under some control. Alternatively, performance measures for systems which are difficult to analyze can be bounded by corresponding measures for easier ones. If we are lucky, these bounds may even be good approximations.

In this paper, we briefly review published results in support of conventional wisdom and cite one published result which is contrary. Then, in a series of examples, we exhibit other results which are contrary to conventional wisdom. All of our examples compare performance measures for different service distributions under specified, sufficiently irregular arrival processes. In some cases, not only does a performance measure move in the "wrong" direction, but also the magnitude of the change is large. In the process, our understanding of the nature of these effects should improve.

1. RESULTS IN SUPPORT OF CONVENTION WISDOM

Kingman's upper bound [7] for the expected delay in a GI/G/1 queue is in terms of the sum of the variance of the inter-arrival and service distributions. Marshall [13] obtained lower bounds on expected delay for this queue when the inter-arrival distribution is more regular than the exponential (under either bounded mean residual life or increasing failure rate). The closeness of the upper and lower bounds shows that, in certain cases, the upper bound is a good approximation.

For heavy traffic,* Kingman [8], [10] showed that the distribution of delay in queue is approximately an exponential distribution with mean equal to his upper bound in [7]. Köllerström [12] extended Kingman's result to the GI/G/c queue: In heavy traffic, the delay distribution is approximately an exponential distribution with mean equal to Kingman's GI/G/1 upper bound with the variance of service, $V(S)$, replaced by $V(S/c)$. That is, in heavy traffic, a multi-server queue behaves like a "fast" single server queue. Thus, the lower bound obtained by Brumelle [1] for the expected delay in a GI/G/c queue is a good approximation in heavy traffic. Similarly, the heavy traffic approximations of Iglehart and Whitt [5] for the number of customers in a GI/G/c queue are consistent with conventional wisdom.

Bounds on the delay distribution for the GI/G/1 queue [9], [16] are also consistent with conventional wisdom.

In several papers, with the main results summarized in [15], it is shown that for the GI/G/1 queue under various definitions of regularity, the stationary delay distribution becomes more regular as the inter-

---

*Definitions of heavy traffic vary, but for most purposes, we mean heavily loaded systems where server utilization is less than but near 1.

arrival and/or service distributions become more regular. For two of these definitions, the mean delay decreases.

Finally, in a classic paper by Kiefer and Wolfowitz [6], it is shown that for a stable GI/G/c queue with positive arrival rate, the r^{th} moment of the stationary delay distribution is finite if and only if the $(r + 1)^{st}$ moment of the service time distribution is finite.

2. A CONTRARY RESULT FOR THE INFINITE SERVER QUEUE

The evidence in favor of conventional wisdom is impressive. However, it should be noted that most of these results pertain to the GI/G/1 queue without losses (all customers are served). With the exception of [6], all results pertaining to the GI/G/c queue are heavy traffic approximations.

In an unpublished paper by Haji and Newell [4], summarized in Newell ([14], p. 32), an infinite server queue is analyzed for the mean and variance of N, the stationary number of busy servers.

For arrival rate λ and service time S with $E(S) = 1/\mu$ and $P(S \leq t) = G(t)$, $E(N)$ obviously depends only on these rates:

$$E(N) = \lambda/\mu \ . \tag{1}$$

For the variance to mean ratio, $V(N)/E(N)$, Haji and Newell obtain the approximation:

$$V(N)/E(N) \simeq 1 + (I - 1)v \ , \tag{2}$$

where I is (roughly) the ratio of the variance to the mean number of arrivals in an interval and

$$v = \int_0^\infty [1 - G(t)]^2 dt/E(S) \ . \tag{3}$$

Now v is a measure of service time regularity (increasing v means greater regularity). In fact, v is maximized ($v = 1$) uniquely in the case of constant service, where $(1 - G)^2 = 1 - G$ for all t.

The effect of server regularity on (2) depends on the sign of $I - 1$. For a Poisson process, $I = 1$ and v has no effect. For less regular arrival processes, e.g., batch Poisson arrivals, $I > 1$, and making the service more *regular* in the sense of (3) *increases* $V(N)$.

While contrary to conventional wisdom (and, at first, quite a shock) this observation has an intuitive explanation: For arrivals which occur in batches and constant service, either an entire batch is present or none of it is at any time t. Irregular service spreads out the departure times, permiting the observance of "partial" batches, i.e., the number present from any batch is somewhere in-between "all" and "none." Thus, we expect larger deviations from the fixed $E(N)$, and hence a larger variance, when service is regular.

3. LOSS SYSTEMS

By a loss system, we mean a system with c servers in parallel (c channels) such that an arrival finding all servers busy departs immediately without receiving service (the arrival is lost). An important performance measure for these systems is the fraction of arrivals lost.

A classic result, with a long history in the literature, is called *Erlang's Loss Formula*, e.g., [18]: For Poisson arrivals and general independent service (the M/G/c loss system), the stationary distribution of the number of busy servers $\{p_n , n = 0,1, \ldots, c\}$ is the unique solution to:

$$\lambda p_{n-1} = n\mu p_n , \; n = 1,2, \ldots, c \quad \text{and} \quad \sum_{n=0}^{c} p_n = 1 , \tag{4}$$

where λ and μ are the arrival and service rates.

Of course, we could exhibit the solution to (4) explicitly. It is written in the form above to emphasize that (a) the solution depends *only* on the arrival and service rates and (b) knowing this, (4) can be written down immediately by assuming that G is exponential. The fraction of arrivals lost in this case is also the fraction of time there are c busy servers, p_c .

Loss systems when the arrival process is not Poisson occur naturally in telephone systems, in particular, when the arrival process in a loss system is itself the overflow process of lost calls at some other loss system. The *equivalent random method*, described in [2], is an approximation technique developed to estimate the fraction of calls lost when the arrival process is a composite of overflow processes and service is exponential.

One might expect loss systems and infinite server systems to be closely related. The equivalent random method is based on this notion. More general approximation methods are presently under investigation [3] which attempt to do this explicitly for general arrival processes and service time distributions in terms of the ratio $V(N)/E(N)$ in Section 2. This ratio is called *peakedness*.

While evaluating the accuracy of various approximation methods is of considerable interest, that is not our purpose here. Instead, we will show that loss systems can also exhibit contrary behavior.

Our example will be for a loss system with c channels, batch Poisson arrivals (denoted by BM) with batch arrival rate λ and constant batch size b, and three different service distributions at rate μ: constant (D), exponential (M), and a special case of hyperexponential (H),

$$H = (1 - \alpha)U_o + \alpha \exp(\alpha\mu) ,$$

i.e., H is a mixture of an exponential with mean $1/\alpha\mu$ and a unit step at the origin.

The exponential and hyperexponential cases can be solved from balance equations. If we choose b and c so that c/b is an integer, then the constant service case can be solved using Erlang's loss formula because groups of b servers are busy and idle together. Thus, the constant service case behaves "like" an M/D/(c/b) loss system.

For $b = c = 2$, the percent of calls lost for constant service is easily shown to be identical with that for exponential service. That is, it does not exhibit contrary behavior. However, this case is

very special. The irregular arrivals (2 at a time) match perfectly the number of channels.

Example 1:

We now exhibit contrary behavior for the case: batch size $b = 2$ and $c = 4$ channels. In this case, we did not explicitly compute results for H.* In the limit $(\alpha \to 0)$ it can be shown that the hyperexponential case behaves like an M/M/4 loss system with the same offered load: $2\lambda E(S)$. Results are as follows:

$2\lambda E(S)$	BM/D/4 % loss	BM/M/4 % loss	M/M/4 % loss
.5	2.4	2.0	0.16
1	7.7	6.6	1.5
2	20.0	18.4	9.5
3	31.0	29.5	20.6
4	40.0	38.7	31.1
6	52.9	52.1	47.0
8	61.5	61.0	57.5

Notice that the direction of the effect of service regularity is independent of the offered load. The effect, at least in this case, appears to be greater to the right (less regular) side of the exponential, and can be substantial.

While this example is consistent with infinite server results, the $b = c = 2$ case shows that peakedness alone may be an inadequate measure, particularly if the number of servers is small.

*Explicit results for the hyperexponential case will be presented in the example in Section 4.

4. PROCESSOR SHARING

Priority rules for processing jobs at a computer system central processor often permit interrupting jobs. An idealized version of such a rule is called *Round Robin:* jobs (customers) join the end of a single queue at a single server (the CPU). On entering service, each job is allocated an amount of CPU time, $\delta > 0$. The job either completes service during δ or, if not, is interrupted after receiving δ and joins the end of the same queue. This is repeated until each job completes service and departs.

The limiting version of the round robin rule as $\delta \to 0$ is called *Processor Sharing*. Under a work-conservation assumption, it was first shown in [17] that the M/G/1 Processor Shared (PS) queue possesses a remarkable property analogous to Erlang's loss formula: the stationary distribution of the number of customers in systems, $\{p_n\}$ is

$$p_n = (1 - \rho)\rho^n \ , \ n = 0,1, \ \ldots \ , \text{ and} \tag{5}$$

$$L = \sum np_n = \frac{\rho}{1 - \rho} \ , \tag{6}$$

where $\rho = \lambda E(S)$, *independent* of the form of the service distribution.

For non-Poisson arrivals, can the PS rule exhibit contrary behavior? By now, it should be clear how to proceed.

Example 2:

Consider a batch Poisson arrival process with batch rate λ and random batch size ν .

(a) For exponential service, L is independent of the rule, i.e., for BM/M/1 ,

$$L_{PS} = L_{FIFO} = \frac{\rho}{1 - \rho} \cdot \frac{E\{(\nu)(\nu + 1)\}}{2E(\nu)} \tag{7}$$

(b) For hyperexponential service with $H = (1 - \alpha)U_o + \alpha \exp(\alpha\mu)$, the customers with zero service time go through immediately under a PS rule. This leaves customers with exponential service. Thus, this system is equivalent to a BM/M/1 FIFO queue composed only of "long" service time customers. Therefore,

$$L_{PS} = \frac{\rho}{1 - \rho} \cdot \frac{E\{(\nu_\ell)(\nu_\ell + 1)\}}{2E(\nu_\ell)} , \tag{8}$$

where ν_ℓ is the number of "longs" in a batch of size ν . It is easily shown that

$$\frac{E\{(\nu_\ell)(\nu_\ell + 1)\}}{2E(\nu_\ell)} = \frac{2E(\nu) + \alpha[E(\nu^2) - E(\nu)]}{2E(\nu)} , \tag{9}$$

and

$$\lim_{\alpha \to 0} L_{PS} = \frac{\rho}{1 - \rho} . \tag{10}$$

(c) For constant service, restrict the batch size to be a constant $\nu = b$. Sharing customers is equivalent to sharing batches, where the expected number of batches in system is given by (6). Since each batch in system contains exactly b customers,

(11) $$L_{PS} = \frac{\rho b}{1 - \rho}$$

In this example, the effect of service regularity can be large. If b is large, (11) is nearly double (7). For fixed ρ, we can make (7) as large as we want by making batches large and/or irregular. The effect of hyperexponential service is to remove the batch effect, i.e., for fixed ν and sufficiently small α, ν_ℓ is likely to be 0 or 1.

We also remark that quite apart from the arrival process, Processor Sharing is a terrible rule when service is sufficiently regular. For constant service, *every* customer departs later under PS than under FIFO (strictly later, except for those jobs that end busy periods).

When interruptions are permitted, one needs to interpret conservation laws with care, e.g., for the conservation law on pg. 199 of [11], it is not true that reducing the delays of some jobs can only be achieved by increasing the delays of others.

5. MULTI-CHANNEL FIFO QUEUES WITHOUT LOSSES

In Section 1, we observed that in heavy traffic, multi-channel queues behave like fast single channel queues. Contrary behavior is possible only for sufficiently "light" traffic.

In fact, one might expect that for a sufficiently irregular arrival process, there will be a "crossover" point of server utilization such that irregular service is preferred below that point and regular service is preferred above. (The crossover point would presumably depend on the distributions being compared.) The following example exhibits this behavior.

Example 3:

Consider a batch poisson arrival process with constant batch size $b = 4$ and $c = 2$ channels, with server utilization $\rho = 2\lambda/\mu$.

(a) For the BM/M/2 queue, the generating function of the state probabilities $P(z) = \sum_n p_n z^n$ was found and differentiated, yielding

$$L = \frac{\rho}{4 + \rho} + \frac{5\rho}{2(1 - \rho)} . \tag{12}$$

(b) The BM/D/2 queue may be analyzed as two BM/D/1 queues with batch arrival size $b = 2$ at each. Finding L is now easy:

$$L = 2\rho + \frac{\rho(\rho + 1)}{(1 - \rho)} . \tag{13}$$

There is a crossover point at $\rho = \rho_o \approx .35$, with exponential service preferred for $\rho > \rho_o$ and constant service preferred for $\rho < \rho_o$.

For the case $c = b = 2$, no crossover point exists. Constant service is preferred to exponential service for all ρ .

6. CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The results of this paper are provacative rather than definitive.

In all three examples, we compared service distributions for specified sufficiently irregular arrival processes. Exhibiting contrary behavior in the converse situation appears to be much more difficult. If examples of this kind exist, it is this author's judgment that they would depend on very special structural relationships between the arrival process and service facility. For example, batches of size three might perform better than batches of size two for a loss system with three servers.

The Poisson process appears to be a boundary between irregular arrival processes which can exhibit contrary behavior and regular arrival processes which cannot. Contrary behavior also depends on structure, e.g., it occurs in processor sharing under circumstances in which it does not occur in loss systems.

The competing effects which account for the crossover in the multi-channel case (Example 3) tend to diminish the effect of service time variability on system performance. Thus, the mathematically convenient assumption of exponential service may result in surprisingly good approximations for system performance in moderately loaded multi-channel queues with irregular arrivals.

For arrival processes which are more regular than Poisson in some appropriate sense (including the Poisson itself) it is conjectured that contrary behavior cannot occur. In particular, this author would expect results similar to those in [15] to be true.

REFERENCES

[1] Brumelle, S. L., [1971(a)], "Some Inequalities for Parallel-Server Queues," Operations Research, 19, pp. 402-413.

[2] Cooper, R. B., Introduction to Queuing Theory, Macmillan, New York, [1972].

[3] Eckberg, A. E., and A. A. Fredericks [1976] personal communication, Bell Telephone Laboratories, Holmdel, NJ 07733.

[4] Haji, R., and G. F. Newell, "Variance of the Number of Customers in an Infinite Channel Server," University of California, [1971]. Unpublished.

[5] Iglehart, D. L. and W. Whitt, [1970], "Multiple Channel Queues in Heavy Traffic I," Adv. Appl. Prob., 2, pp. 150-177.

[6] Kiefer, J., and J. Wolfowitz [1956], "On Characteristics of the General Queuing Process with Applications to Random Walk," Ann. Math. Stat., 27, pp. 147-161.

[7] Kingman, J. F. C., [1962(a)], "Some Inequalities for the GI/G/1 Queue," Biometrika, 49, pp. 315-324.

[8] Kingman, J. F. C., [1962(b)], "On Queues in Heavy Traffic," J. Roy. Stat. Soc., B24, pp. 383-392.

[9] Kingman, J. F. C., "A Martingale Inequality in the Theory of Queues," Proc. Camb. Phil. Soc., 59, pp. 395-461.

[10] Kingman, J. F. C., [1965], "The Heavy Traffic Approximation in the Theory of Queues," Ch. 6 in Proceedings of the Symposium on Congestion Theory, W. L. Smith and W. E. Wilkinson (eds) University of North Carolina Monograph Series in Probability and Statistics.

[11] Kleinrock, L., Queuing Systems Vol. II: Computer Applications, John Wiley and Sons, New York [1976].

[12] Köllerström, J., "Heavy Traffic Theory for Queues with Several Servers. I," J. Appl. Prob., 11, pp. 544-552, [1974].

[13] Marshall, K. T., [1968(a)], "Some Inequalities in Queuing," Operations Research 16, No. 3, pp. 651-665.

[14] Newell, G. F., [1973], Approximate Stochastic Behavior of n-Server Service Systems with Large n, Springer-Verlag, New York.

[15] Rolski, T. and D. Stoyan, "On the Comparison of Waiting Times in GI/G/1 Queues," Operations Research 24, pp. 197-200, [1976].

[16] Ross, S. M., "Bounds on the Delay Distribution in GI/G/1 Queues," J. Appl. Prob., 11, pp. 417-421, [1974].

[17] Sakata, M., S. Noguchi and J. Oizumi, "Analysis of a Processor-Shared Queueing Model for Time-Sharing Systems," Proc. Second Hawaii International Conference on System Sciences, University of Hawaii, Honolulu, Hawaii, pp. 625-628, [1969].

[18] Takács, L., "On Erlang's Loss Formula," Ann. Math. Statist., 40, pp. 71-78, [1969].