J. King Saud Univ., Vol. 12, *Science* (2), pp. 129-142, Riyadh (A.H. 1420/2000)

Comparative Study of M/G^r/1 and M/G/r Queueing Systems

L. Tadj
Department of Statistics and Operations Research,
King Saud University, P.O. Box 2455,
Riyadh 11451, Saudi Arabia

(Received 24/5/1999; accepted for publication 7/11/1999)

Abstract. We study in this paper the effect of substituting a very little studied queueing system, the M/G/r, by a very well studied one, the quorum system, when such substitution is feasible. We show that it results in higher mean system size and mean queue size, but in lower total expected cost per unit of time.

Introduction

It is well known that the multichannel queue M/G/r with ordinary Poisson input and general service time permits no simple analytical solution; see for example Gross and Harris [1, p. 308]. What if, instead of r parallel servers performing the same task, we place a single server who processes customers in batches of size r. For example, instead of having ten taxicabs drive ten people from a certain location X to another location Y, it may be possible to have a bus drive the ten people altogether from X to Y. Intuitively, this would increase the waiting time of the customers, but is this always true? Intuitively also, this would incur lower costs to the system manager, but again, is this always true? The aim of this study is to compare the performance and the costs incurred by a multichannel queueing system and those of a bulk service queueing system, to help system manager decide whether the trade is worthwhile.

We anticipate applications of this study in transportation systems where we may be better off using a bulk service rather than a multichannel system. Applications to inventory control also seem possible since one may order in bulk or singly.

Queues in which customers are served in bulk are usually called *bulk service queueing systems*. Bulk service queueing systems were introduced by Bailey [2]. Then, they were studied by Downton [3,4], Fabens [5] and Takàcs [6]. Feller [7, p. 196]

reports a paper by Boudreau *et al.* [8] describing *queues for a shuttle train* as follows: A shuttle train with r places for passengers leaves a station every hour on the hour. Prospective passengers appear at the station and wait in line. At each departure the first r passengers in line board the train, and the others remain in the waiting line.

Bulk service queueing systems are also called *quorum systems*; see Chaudhry and Templeton [9]. Quorum systems under various disciplines have been extensively studied in recent years, and tractable solutions are available. A useful bibliography on quorum systems is contained in Dshalalow [10, pp. 61-116]. The idea now is to replace a multichannel queue M/G/r by a quorum system M/G^r/1. Of course, the nature of the multichannel queueing system may not allow such change, but when this is feasible, is it worth making such a change?

The rest of the paper is organized as follows. Section 2 describes the queueing models M/G^r/1 and M/G/r and summarizes relevant results. The computational comparison of the systems is carried out in Section 3. The paper is summarized and conclusions are presented in Section 4.

Previous Results

The M/G^r/1 queueing system

In this model, the arrival process of customers to a single-channel facility is a Poisson process with rate λ. There is no waiting capacity constraint. The server processes customers in batches of a fixed size r except when less than r are in the queue, in which case the server remains idle until the line size reaches r. The service time of a batch of customers has a general probability distribution function $B(t)$ with finite first mean μ^{-1} and second moment b_2. The Laplace-Stieltjes transform of $B(t)$ is denoted $B^*(\theta)$ and is defined by $B^*(\theta) = \int_0^\infty e^{-\theta t} dB(t)$.

The following results may be derived as a special case from the results of Dshalalow and Tadj [11]. The embedded process $Q(T_n)$, where $Q(t)$ denotes the number in the system at an arbitrary instant of time t and T_1, T_2, T_3, ... are the successive times of service completion, is an ergodic Markov chain, provided the traffic intensity $\rho=\lambda/\mu r$ is smaller than 1. Let $p_i^+ = \lim_{n\to\infty} P\{Q_n = i\}$ denote the steady-state system-size probability at a service completion. Also, let $P^+(z) = \sum_{i\geq 0} p_i^+ z^i$. Then

$$P^+(z) = \frac{B^*(\lambda - \lambda z)\sum_{i<r}(z^r - z^i)p_i^+}{z^r - B^*(\lambda - \lambda z)}, \tag{1}$$

and $p_0^+, p_1^+, \ldots, p_{r-1}^+$ are determined by solving the following system of linear

equations:

$$\sum_{i<r} p_i^+ \frac{d^k}{dz^k}\left[A(z)-z^i\right]\Big|_{z=z_s} = 0; s = 1,...,S; k = 0,...,k_s - 1, \tag{2}$$

$$\sum_{i<r} p_i^+ (r-i) = r - \rho r, \tag{3}$$

where z_s; s=1, ..., S, are the roots of the characteristic equation z^r-$B^*(\lambda-\lambda z)$=0 that belong to the closed unit ball $\overline{B}(0,1)$ in C with their multiplicities k_s such that $\sum_{s=1}^{S} k_s = r - 1$.

Another expression is derived by Chaudhry and Templeton [9, p. 191] for $P^+(z)$

$$P^+(z) = r(1-\rho)(z-1)B^*(\lambda-\lambda z)\frac{\prod_{i<r}(z-z_i)/(1-z_i)}{z^r - B^*(\lambda-\lambda z)}, \tag{4}$$

where z_1, z_2, ..., z_{r-1} are the r-1 roots in z of the characteristic equation inside the unit circle $|z|$=1.

Now, let $p_i = lim_{t\to\infty} P\{Q(t) = i\}$ denote the steady-state system-size probability at an arbitrary instant of time. Also, let $P(z) = \sum_{i\geq 0} p_i z^i$. Then

$$P(z) = \frac{1}{r}\cdot\frac{1-z^r}{1-z}P^+(z), \tag{5}$$

that is,

$$p_i = \begin{cases} \frac{1}{r}\sum_{j=0}^{i} p_j^+ z^j & i < r, \\ \frac{1}{r}\sum_{j=i-r+1}^{i} p_j^+ z^j & i \geq r. \end{cases} \tag{6}$$

Our intention is to build a cost model in order to compare the M/G^r/1 and M/G/r queueing systems. Before developing the cost structure, we need to derive some measures of performance of the system.

The mean system size at a departure epoch can be obtained by taking the first derivative of the probability generating function $P^+(z)$, derived in (1), with respect to z and setting z=1. We obtain after considerable algebra

$$L^+ = \frac{1}{2(r-\rho r)}\left\{r^2(M_r - 2\rho^2 - 1) - J_r + \rho r(2r+1) + \lambda^2 b_2\right\}, \tag{7}$$

where

$$M_r = \sum_{i<r} p_i^+, \tag{8}$$

and

$$J_r = \sum_{i<r} i^2 p_i^+. \tag{9}$$

The mean queue size at a departure epoch can be derived using expression (7) and the relation

$$L_q^+ = L^+ - \rho r. \tag{10}$$

The mean system size at an arbitrary instant of time can be obtained by taking the first derivative of the probability generating function $P(z)$, derived in (5), with respect to z and setting $z=1$. We get

$$L = L^+ + \frac{1}{2}(r-1). \tag{11}$$

The mean queue size at an arbitrary instant of time can be derived using expression (11) and the relation

$$L_q = L - \rho r. \tag{12}$$

We also need to derive the mean idle and busy periods, and the mean cycle length. A busy period is defined to begin with the arrival of a customer to the servicing facility with r-1 customers waiting in the queue and to end when the queue size next drops below level r at a service completion. A cycle is the sum of a busy period and an adjacent idle period.

The mean idle period is given by

$$\bar{I} = \frac{1}{\lambda} \cdot \frac{r - \rho r}{M_r}. \tag{13}$$

The mean busy period is given by

$$\bar{B} = \frac{1 - K_i}{K_i} \bar{I}, \tag{14}$$

where

$$K_i = \sum_{i<r} p_i. \tag{15}$$

An equivalent but simpler expression is found in Chaudhry and Templeton [9, p. 324]

$$\bar{B} = \frac{1}{\lambda} \frac{\rho r}{M_r}. \tag{16}$$

The mean cycle length is given by

$$\bar{C} = \frac{1}{\lambda} \frac{r}{M_r}. \tag{17}$$

It is interesting to note that the probability that the service station is idle is equal to the traffic intensity

$$\frac{\bar{B}}{\bar{B} + \bar{I}} = \rho, \tag{18}$$

a result obtained by Heyman [12] for the M/G/1 queueing system.

We, now, are ready to define the system's expected total cost per unit of time. We use the same cost structure that has been widely used in the literature for the control of queues. Let

- c_s : start-up cost per cycle, incurred each time the server is turned on.
- c_h : holding cost per unit time, incurred for each customer in the system.
- c_o : operating cost per unit time, incurred for operating the service station.
- c_a : cost per unit time, incurred for performing an auxiliary task by the service station.

Then the total expected cost per unit time is given by

$$TC = c_h L + c_o \frac{\bar{B}}{\bar{C}} + c_a \frac{\bar{I}}{\bar{C}} + c_s \frac{1}{\bar{C}}. \tag{19}$$

Taking expressions (13), (16), and (17) into account, relation (19) reduces to

$$TC = c_h L + c_o \rho + c_a (1 - \rho) + c_s \frac{\lambda M_r}{r}. \tag{20}$$

The M/G/*r* queueing system

In this model, the arrival process of customers to a multichannel facility is a Poisson process with rate λ. There is no waiting capacity constraint. The service time S of a customer has a general probability distribution function $B(t)$ with finite first mean μ^{-1}. It is assumed that the traffic intensity $\rho=/\mu r$ is smaller than one.

As mentioned in the introduction, obtaining the stationary probability generating function for the distribution of the queue size is not possible generally for M/G/*r*. Tijms [13, pp.293-294] obtains useful approximations using a regenerative approach. He first makes an approximation assumption with regard to the behavior of the process at the service completion epochs.

Approximation assumption

(a) If at a service completion epoch k customers are left behind in the system with $1 \le k < r$, then the time until the next service completion epoch is distributed as $min(S_1^e,...,S_k^e)$, where $S_1^e,...,S_k^e$ are independent random variables having each the residual life distribution function

$$B_e(t) = \frac{1}{E(S)} \int_0^t [1 - B(x)]dx; t \ge 0,$$

as probability distribution function.

(b) If at a service completion epoch k customers are left behind in the system with $k \ge r$, then the time until the next service completion epoch is distributed as S/r, where S denotes the original service time of a customer.

Theorem

Under the Approximation Assumption,

$$p_j^{app} = \frac{(r\rho)^j}{j!} p_0^{app}; j = 0,1,...,r-1, \tag{21}$$

$$p_j^{app} = \lambda a_{j-r} p_{r-1}^{app} + \lambda \sum_{k=r}^{j} b_{j-k} p_k^{app}; j = r, r+1,... \tag{22}$$

where the constants a_n and b_n are given by

$$a_n = \int_0^\infty [1 - B_e(t)]^{r-1}[1 - B(t)]e^{-\lambda t}\frac{(\lambda t)^n}{n!}dt; n = 0,1,...$$

$$b_n = \int_0^\infty [1 - B(rt)]e^{-\lambda t}\frac{(\lambda t)^n}{n!}dt; n = 0,1,...$$

As for the previous model, we derive some measures of performance in order to write the system's total expected cost per unit of time explicitly.

The mean queue size is given by

$$L_q^{app} = \left[(1-\rho)\gamma_1 \frac{r}{E(S)} + \frac{1}{2}(1 + c_S^2)\right] L_q(\exp), \tag{23}$$

where $c_S^2 = \sigma^2(S)/E^2(S)$ and

$$\gamma_1 = \int_0^\infty [1 - B_e(t)]^r dt.$$

The quantity L_q(exp) denotes the average queue size in the M/M/r queue.

The mean system size can be derived from expression (23) using the relation

$$L^{app} = L_q^{app} + \rho r. \tag{24}$$

We again need to derive the mean idle period, the mean busy period, and the mean cycle length. Gross and Harris [1] define an i-channel busy period for the multichannel queue M/M/r ($0 \le i \le r$) to begin with an arrival to the system at an instant where there are i-1 in the system to the very next point in time when the system size dips to i-1. The case where i=1 (an arrival to an empty system) defines the system busy period.

Since the arrival process is Poisson, *the mean system idle period* is given by

$$\bar{I} = \frac{1}{\lambda}. \tag{25}$$

The mean system busy period is derived by writing the ratio of the percentage of time the server is busy to the percentage of time he is idle:

$$\frac{\bar{B}}{\bar{I}} = \frac{1 - p_0}{p_0}, \tag{26}$$

which yields

$$\bar{B} = \frac{1 - p_0}{p_0} \bar{I}. \tag{27}$$

The mean system cycle length is therefore

$$\bar{C} = \frac{1}{p_0} \bar{I}. \tag{28}$$

Employing the same cost structure as for the previous model, the system's total expected cost per unit of time is given by

$$TC = c_h L + c_o (1 - p_0) + c_a p_0 + c_s \frac{p_0}{\bar{I}}. \tag{29}$$

Computational Study

We carry out the computational study by assuming that the service times are exponentially distributed, that is, $B(t) = 1\text{-}e^{-\mu t}$, $t \geq 0$. The Laplace-Stieltjes transform of the service time distribution is given by $B^*(\theta) = \mu/(\mu+\theta)$.

The M/M^r/1 queueing system

To compute the steady-state system-size probabilities at a service completion $p_0^+, p_1^+, ..., p_{r-1}^+$, we first solve the characteristic equation $z^r\text{-}B^*(\lambda-\lambda z)=0$, which reduces, in this case, to:

$$\rho r z^r - z^{r-1} - ... - z - 1 = 0. \tag{30}$$

It is well known that equation (30) has r-1 simple roots inside the closed unit ball. They are denoted z_s, s=1, ..., r-1. Next, we solve the system of r linear equations (2) and (3). Finally, the steady-state system-size probabilities at an arbitrary instant of time p_0, ..., p_{r-1} are computed using relations (6). The various measures of performance and the total expected cost per unit of time derived in Section 2 are computed using their respective relations.

The M/M/r queueing system

The approximations given in the previous section become exact when the service time is assumed to be exponentially distributed. The well known results of the M/M/r

queue may be used. We get, see e.g., Gross and Harris [1]:

$$p_0 = \left[\sum_{n=0}^{r-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^r}{r!} \left(\frac{r\mu}{r\mu - \lambda} \right) \right]^{-1}, \tag{31}$$

$$p_i = \begin{cases} \dfrac{(\lambda/\mu)^i}{i!} p_0 & 1 \le i < r, \\ \dfrac{(\lambda/\mu)^i}{r^{i-r} r!} p_0 & i \ge r. \end{cases} \tag{32}$$

The mean queue length is given by

$$L_q = \left[\frac{(\lambda/\mu)^r \lambda\mu}{(r-1)!(r\mu - \lambda)^2} \right] p_0. \tag{33}$$

The mean system size is given by

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^r \lambda\mu}{(r-1)!(r\mu - \lambda)^2} \right] p_0. \tag{34}$$

The mean idle period is given by

$$\bar{I} = \frac{1}{\lambda}. \tag{35}$$

The mean busy period is given by

$$\bar{B} = \frac{1 - p_0}{p_0} \bar{I}. \tag{36}$$

The mean cycle length is given by

$$\bar{C} = \frac{1}{p_0} \bar{I}. \tag{37}$$

Computational Comparison

Most computations were done using version 5 of the Matlab mathematical package on a PC running at 90 Mhz. They took only fractions of a second to complete. Three different values of the traffic intensity were chosen: ρ=0.1 (light traffic), ρ=0.5 (average

traffic), ρ=0.9 (heavy traffic). In all three cases, the steady-state system-size probabilities were computed. To compute L and L_q, *we set* $\lambda^2 b_2=2(\rho r)^2$, since the second moment of an exponential distribution is given by $b_2=2/\mu^2$.

Figures representing the variations of the mean system size, the mean queue size, the mean idle period, the mean busy period, the mean cycle length, the system "turned-off" probability, and the total expected cost per unit of time, as a function of r were drawn for all three values of the traffic intensity. For brevity, only the graphs representing the costs are exhibited and only a few words will be said about the others.

As expected, for for all three kinds of traffic, the quorum mean system size is much higher than the multichannel mean system size, and the difference gets higher and higher as r increases.

The same conclusions can be made for the mean queue size as in the case of the mean system size in all three kinds of traffic. This agrees with our prediction that the mean waiting time of a customer is much higher in the quorum system than in the multichannel.

The graphs representing the mean idle period, the mean busy period, and the mean cycle length presented no surprise. The quorum mean idle period was always larger than the multichannel mean idle period, the quorum mean busy period was always smaller than the multichannel mean busy period, and the quorum mean cycle length was larger than the multichannel mean cycle length for ρ=0.1, and smaller for ρ=0.5 andρ=0.9.

The quorum system is turned off when the server is idle and the probability of this to happen is just $\sum_{i<r} p_i$. The multichannel system is turned off when all severs are idle, that is with probability p_0. In all three kinds of traffic, the quorum "system turned-off" probability gets higher and higher than the multichannel "system turned-off" probability as r increases.

Finally, we came to compute the total expected cost per unit time, relation (20) for the quorum system, and (29) for the multichannel system. We set λ=1 to compute the mean idle and busy periods, and the mean cycle length. We also chose the following unit costs: c_h=10, c_0=0.1, c_a=0.1, and c_s=1000. Figure 1 represents the variations of the total expected cost per unit time for both systems, as a function of r. As expected, in all three kinds of traffic: light (Fig. 1a), average (Fig. 1b), and heavy (Fig. 1c), the quorum cost is lower than the multichannel cost. In other words, the system manager is better off using a bulk service model instead of a multichannel model, if feasible. Of course, other values of the parameters may bring completely different results. A sensitivity analysis on the unit costs may reveal how sensitive our results are? We do not conduct such analysis since our computations are merely for illustrative purposes. Note that the fixed unit costs c_h, c_o, c_a, and c_s need not be the same for both systems since starting up a bulk system,

for example, may be more costly than starting up a multichannel one.

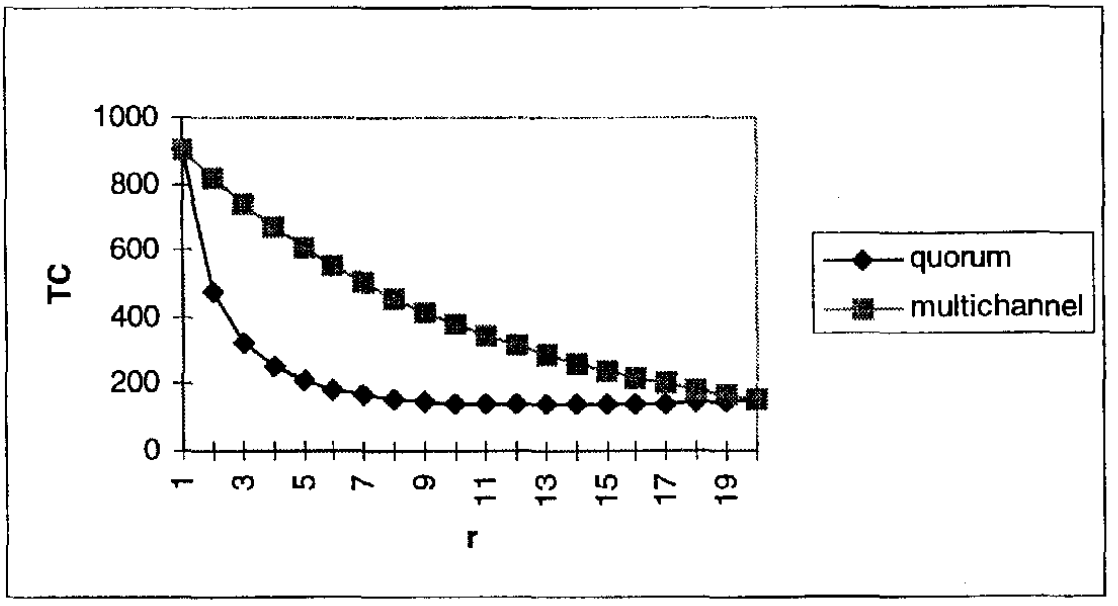

Fig. 1a. Total cost comparison (ρ=0.1).

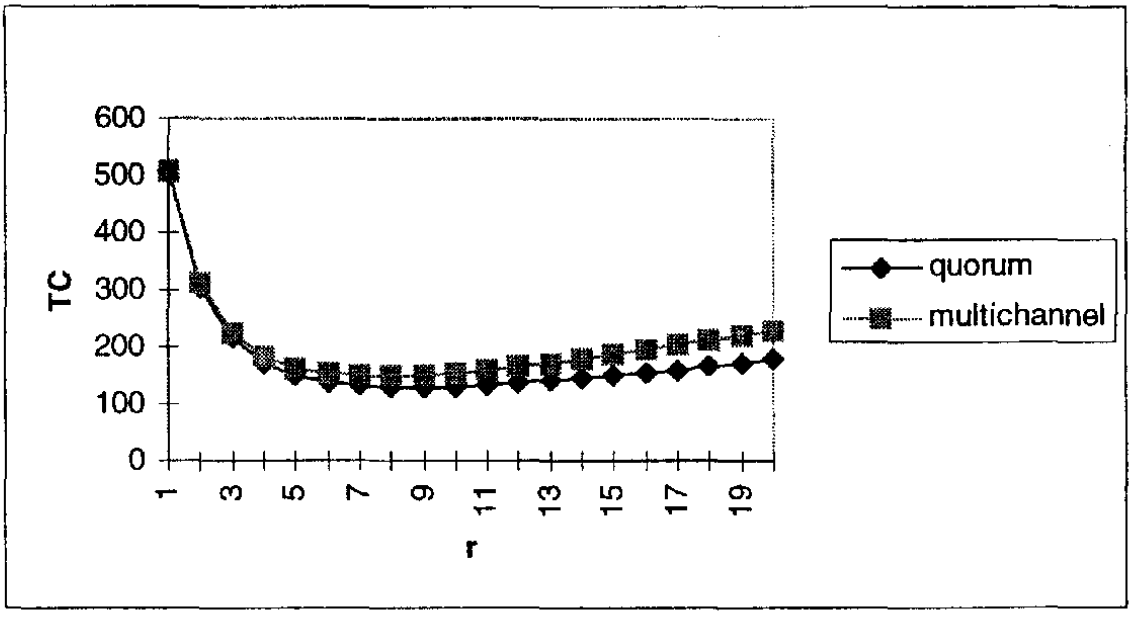

Fig. 1b. Total cost comparison (ρ=0.5).

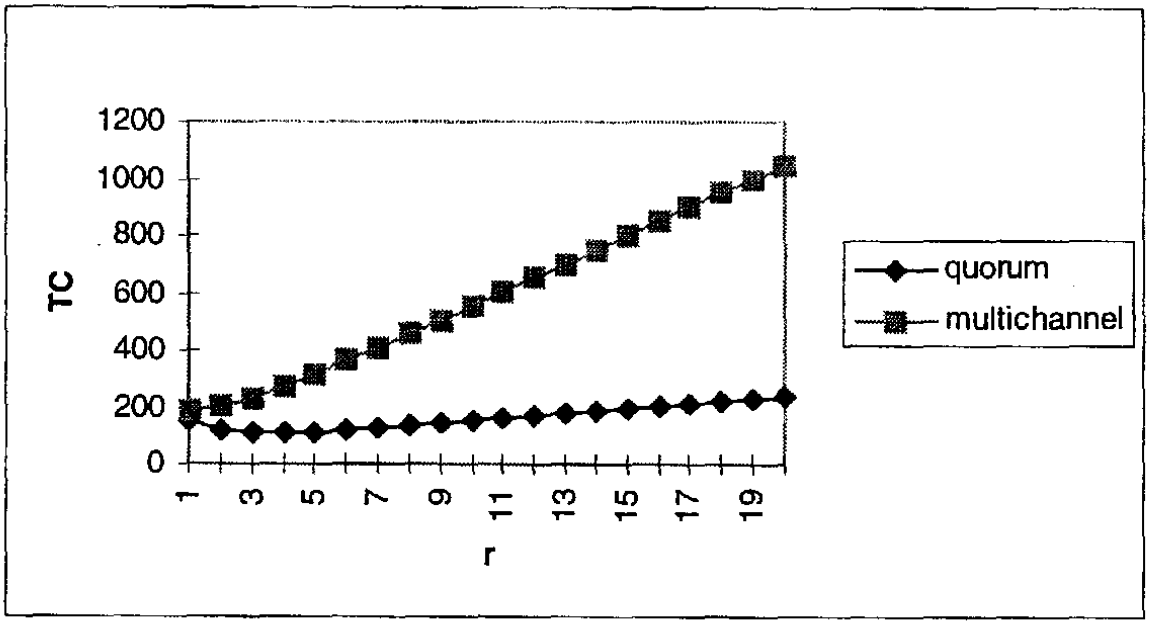

Fig. 1c. Total cost comparison (ρ=0.9).

Conclusion

In this paper, we study the effect of substituting a very little studied queueing system, the M/G/r, by a very well studied one, the quorum system. This substitution is not always feasible, depending on the nature of the queueing system on hand, but when it is, it may result in lower total expected cost per unit of time. Given fixed values for the system parameters and given a distribution function for the service time (which may not be exponential, but the analysis would be conducted along the same lines), a system analyst may help a decision maker decide whether a substitution renders the system more efficient.

Note that all the graphs representing the total expected cost per unit of time are convex. It may be worth investigating whether this is always true. This suggests that one may find a "best" r for which the total cost is minimum. One such r would satisfy $TC(r) \leq TC(r+1)$.

References

[1] Gross, D. and Harris, C.M. *Fundamentals of Queueing Theory*. New York: Wiley, 1974.

[2] Bailey, N.T.J. "On Queueing Processes with Bulk Service." *J. Roy. Stat. Soc., Ser. B*, 16 (1954), 80-87.

[3] Downton, F. "Waiting Times in Bulk Service Queues." *J. Roy. Stat. Soc., Ser. B*, 17 (1955), 256-261.

[4] ————. "On Limiting Distributions Arising in Bulk Service Queues." *J. Roy. Stat. Soc., Ser. B*, 18 (1956), 265-274.

[5] Fabens, A.J. "The Solution of Queueing and Inventory Models by Semi-Markov Processes." *J. Roy Stat.*

Soc., Ser. B, 23 (1961), 113-117.

[6] Takàcs, L. *Introduction to the Theory of Queues.* New York: Oxford University Press, 1962.

[7 Feller, W. *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York: Wiley, 1971.

[8] Boudreau, P.E., Griffin Jr., J.S., and Kac, M. "An Elementary Queueing Problem." *Amer. Math. Monthly*, 69 (1962), 713-724.

[9] Chaudhry, M.L. and Templeton, J.G.C. *A First Course in Bulk Queues.* New York: Wiley, 1983.

[10] Dshalalow, J.H. "Queueing Systems with State Dependent Parameters." In: Dshalalow, J.H. (Ed.). *Frontiers in Queueing: Models and Applications in Science and Engineering*.. Boca Raton: CRC Press, 1997.

[11 ———— and Tadj, L. "Queueing System with a Fixed Accumulation Level, Random Server Capacity, and Capacity Dependent Service Time." *Intern. J. Math. and Math. Sci.*, 15, No.1 (1992), 189-194.

[12] Heyman, D.P. "Optimal Operating Policies for M/G/1 Queueing System." *Opns Res.*, 16 (1968), 362-382.

[13] Tijms, H.C. *Stochastic Models: An Algorithmic Approach.* Hichester:Wiley, 1998.

مقارنة بين نموذجي الصفوف
$M/G^r/1, M/G/r$

لطفي تاج

قسم الإحصاء وبحوث العمليات، كلية العلوم، جامعة الملك سعود ، ص.ب ٢٤٥٥ ،

الرياض ١١٤٥١ ، المملكة العربية السعودية

(استلم للنشر في ١٤٢٠/٢/٩هـ؛ وقبل للنشر في ١٤٢٠/٧/٢٧هـ)

ملخص البحث. ندرس في هذا البحث أثر تبديل نموذج الصفوف ذي القنوات المتعددة بنموذج تكون فيه الخدمة جماعية، إذا كان ذلك ممكناً، وذلك لأن النموذج الأول درس قليلاً بينما درس النموذج الثاني دراسة جيدة . يؤدي هذا التبديل إلى ارتفاع في متوسط عدد الزبائن في الجهاز ومتوسط طول الصف، وإلى انخفاض في التكلفة الكلية المتوقعة في وحدة الزمن.