Journal of the Operational Research Society (1996) 47, 817–828 0160-5682/96 $12.00

The Finite Capacity GI/M/1 Queue with Server Vacations

FIKRI KARAESMEN and SURENDRA M. GUPTA
Northeastern University, Boston, Massachusetts, USA

We consider the GI/M/1/K queue where the server takes exponentially distributed vacations when there are no customers left to serve in the queue. We obtain the queue length distribution at arrival epochs and random epochs for the multiple vacation case. We present heuristic algorithms to compute the blocking probability for this system. Several numerical examples are presented to analyze the behaviour of the blocking probability and to test the performance of the heuristics.

Key words: queueing, stochastic processes

INTRODUCTION

The finite capacity M/G/1 queue with server's vacations has been studied by Courtois[1] and Lee[2]. Lee obtains the performance measures of the queue in terms of Laplace-Stieltjes transforms (LSTs). Keilson and Servi[3,4], reveal important structural properties of the M/G/1/K queue by relating the queue length distribution of the corresponding infinite buffer queue to the finite buffer queue. Blondia[5] presents an analysis of the finite capacity queue with a Markovian Arrival Process (MAP) and arbitrarily distributed service and vacation times.

The analysis of the GI/M/1 queue with vacations is considerably more difficult than the corresponding M/G/1 queue. In the M/G/1 queue with arbitrarily distributed vacations, the state of the system can be completely described by two discrete and one continuous variable. For example, consider a triplet (i, j, x) where i $(i = 0, 1)$ denotes the status of the server, j $(j = 0, 1, 2, \ldots)$ denotes the number of customers in the queue and x $(x \geqslant 0)$ is the elapsed time since the beginning of the last service (vacation) when $i = 1$ $(i = 0)$. This triplet is a complete description of the Markov process for the M/G/1 queue with vacations. On the other hand, for the GI/M/1 queue with general vacations a complete description of the state space requires one discrete but two continuous variables. The triplet describing the state space in this case is (j, x, y) where j $(j = 0, 1, 2, \ldots)$ is the queue length, x $(x \geqslant 0)$ is the elapsed time since the last arrival and y $(y \geqslant 0)$ is the elapsed time since the beginning of the last vacation period. These facts make the M/G/1 queue with vacations much more amenable to analysis and necessitates imposing a restriction on the distribution of the vacation period in the analysis of the GI/M/1 queue.

The infinite capacity GI/M/1 queue with vacations has been independently studied by Chatterjee and Mukherjee[6] and by Tian *et al.*[7]. Chatterjee and Mukherjee incorrectly claim that their solution is valid for generally distributed vacation times (see Appendix 1). Tian *et al.* perform a similar kind of analysis for the case of exponential vacations. Using matrix-geometric solutions, they are able to obtain closed form solutions in terms of a constant r_0 that can be obtained as the solution of a simple equation.

In this paper, we obtain the queue length distribution of the GI/M/1/K under a *multiple* exhaustive vacation scheme. Under this vacation policy, the server takes repeated vacations unless there is a customer to serve upon return from a vacation. However, once service starts the server will keep on serving until the queue is exhausted. This problem is motivated by approximation algorithms for queueing networks with finite buffers where blocking takes place. Many heuristic procedures decompose the network into individual nodes as in Dallery and Frein[8] and Gershwin[9]. The approximation procedures then simplify to iteratively analyzing finite buffered single queues and the blocking effects are incorporated using the blocking probabilities of the decomposed

Correspondence: S. M. Gupta, Department of Mechanical Industrial and Manufacturing Engineering, 334 Snell Engineering Center, Northeastern University, Boston, Massachusetts 02115, USA

queues. The arrival process to the decomposed nodes are usually approximated by a Poisson process. We attempt to remove this restriction by allowing more flexible renewal process arrivals that can capture the variance of the real input process better (see for example, Whitt[10] and Albin[11]).

In addition to its use in approximation algorithms for queueing networks, the system here would also be useful in manufacturing type models in which the arrival process may be deterministic due to automated operations. This is the case when items are transported to special workstations by an automated system such as an assembly line.

Finally, the model may also be useful in certain polling models where the trade-off between service and vacation times is more important to analyze than capturing processing and polling times accurately.

In the sections to follow, we first obtain the queue length distribution at arrival epochs by using the embedded Markov Chain and then relate it to the distribution at random times by using the underlying semi-Markov process. This computation enables us to compute the key performance measures such as the blocking probability and the waiting time distribution. We analyze the blocking probability for this system in detail and present numerical examples that demonstrate the sensitivity of the blocking probabilities to the arrival distribution. Finally, we suggest heuristic procedures for efficient computation of the blocking probability through explicit solutions of special arrival processes.

NOTATION

The following notation will be used throughout the paper. Additional notation will be introduced when necessary.

$F_X(x)$ the probability distribution function of the random variable X

A random variable denoting the time between arrivals

S random variable denoting the service times

V random variable denoting the vacation times

$1/\lambda$ mean time between arrivals $= \int_0^\infty x\, dF_A(x)$

$1/\mu$ mean service time $= \int_0^\infty x\, dF_S(x) = \int_0^\infty \mu x e^{-\mu x}\, dx$

$1/\theta$ mean vacation time $= \int_0^\infty x\, dF_V(x) = \int_0^\infty \theta x e^{-\theta x}\, dx$

$\hat{V}$ random variable denoting the remaining vacation time distribution (with respect to a random arrival)

$F_{\hat{V}}(x)$ $= \theta \int_0^x (1 - F_V(y))\, dy = 1 - e^{-\theta x}$

A^+ the random variable denoting the amount A exceeds $\hat{V}$

$$F_{A^+}(x) = \frac{\int_{s=0}^{x} \int_{y=0}^{\infty} f_{\hat{V}}(y) f_A(y+s)\, dy\, ds}{\Pr\{A > \hat{V}\}}$$

g_k the probability that there are exactly k service completions in a period of length A

$$= \int_0^\infty \frac{e^{-\mu x}(\mu x)^k}{k!}\, dF_A(x)$$

h_k the probability that there are exactly k service completions in a period of length A^+

$$= \int_0^\infty \frac{e^{-\mu x}(\mu x)^k}{k!}\, dF_{A^+}(x)$$

ω the probability that $\hat{V}$ exceeds

A $$= \int_0^\infty F_A(x) f_{\hat{V}}(x)\, dx$$

ANALYSIS OF PERFORMANCE MEASURES

The embedded Markov Chain

Consider a GI/M/1/K queueing system where the server takes an exponentially distributed vacation when the queue is empty. Arriving customers that find the queue full are assumed to be

blocked and lost to the system. The vacation discipline is such that if, upon return from a vacation, the server finds the queue empty, another vacation is started.

To analyze the above queueing system, we first observe that there is a Markov chain embedded at arrival epochs. The stationary distribution of this embedded Markov chain can be obtained once the transition probabilities are available.

Formally, let the state of the system be denoted by (i, j) where $i = 0$ (or 1) denotes that the server is on vacation (not on vacation) and there are j customers in the system. Then, the following equations hold for the limiting probabilities, $p_{i,j}$, of the embedded Markov chain:

$$p_{1,1} = \sum_{r=1}^{K-1} p_{1,r} g_r + p_{1,K} g_{K-1} + (1-\omega)\left[\sum_{r=0}^{K-1} p_{0,r} h_r + p_{0,K} h_{K-1}\right] \tag{1}$$

$$p_{1,j} = \sum_{r=0}^{K-j} p_{1,j-1+r} g_r + p_{1,K} g_{K-j} + (1-\omega)\left[\sum_{r=0}^{K-j} p_{0,j-1+r} h_r + p_{0,K} h_{K-j}\right] \quad \text{for } 2 \leqslant j \leqslant K-1 \tag{2}$$

$$p_{1,K} = (p_{1,K-1} + p_{1,K}) g_0 + (1-\omega)[(p_{0,K-1} + p_{0,K}) h_0] \tag{3}$$

$$p_{0,0} = \sum_{r=1}^{K-1} p_{1,r} g_{r+1}^c + p_{1,K} g_K^c + (1-\omega)\left[\sum_{r=0}^{K-1} p_{0,r} h_{r+1}^c + p_{0,K} h_K^c\right] \tag{4}$$

$$p_{0,j} = \omega p_{0,j-1} \quad \text{for } 1 \leqslant j \leqslant K-1 \tag{5}$$

$$p_{0,K} = \omega(p_{0,K-1} + p_{0,K}) \tag{6}$$

where $g_j^c = \sum_{r=j}^{\infty} g_r$ and $h_j^c = \sum_{r=j}^{\infty} h_r$.

The above system can be solved together with the boundary condition:

$$\sum_{r=0}^{K} p_{0,r} + \sum_{r=1}^{K} p_{1,r} = 1. \tag{7}$$

Although, we are unable to obtain closed form solutions for the above system as in the case where $K \to \infty$ (see Tian *et al.*[7]), the numerical solution of the above system is straightforward as long as g_i's and h_i's are computable.

The queue length distribution at random times

Let $\{L(t), \varepsilon(t)\}$, be the semi-Markov process that corresponds to the queue length and the server status (where $\varepsilon(t) = 0, 1$) at time t with embedded points at arrival times. Note that this semi-Markov process changes state at arrival times and therefore the time spent in a state at each visit is an interarrival time. If $T_{i,j}$ are the expected sojourn times of this process at state (i, j), then:

$$T_{i,j} = \mathrm{E}[A] \quad \text{for all } i = 0, 1, \quad j = 0, 1, 2, \ldots, K. \tag{8}$$

As all sojourn times are equal, the stationary distribution of the semi-Markov process is equal to the stationary distribution at embedded points that was obtained in the previous section (see Ross[12]).

To pass from the distribution of the semi-Markov process to the distribution at random times, let $\hat{A}$ and $\tilde{A}$ denote the forward and backward recurrence times of an interarrival time respectively. Then:

$$F_{\hat{A}}(x) = F_{\tilde{A}}(x) = \int_0^x \lambda(1 - F_A(y))\, dy. \tag{9}$$

To relate the semi-Markov process to the random time process, also let d_k denote the probability that there are k service completions in the backward recurrence time of an arrival given the server was available at the time of arrival, and d_k^+ denote the same probability given the server

was on vacation at the time of arrival. Then:

$$d_k = \int_0^\infty \frac{e^{-\mu x}(\mu x)^k}{k!}\, dF_{\hat{A}}(x) \tag{10}$$

and

$$d_k^+ = \int_0^\infty \frac{e^{-\mu x}(\mu x)^k}{k!}\, dF_{\hat{A}^+}(x) \tag{11}$$

where $\hat{A}^+$ is the random variable denoting the amount of time $\hat{A}$ exceeds $\hat{V}$ and

$$F_{\hat{A}^+}(x) = \frac{\int_{s=0}^{x}\int_{y=0}^{\infty} f_{\hat{V}}(y) f_{\hat{A}}(y+s)\, dy\, ds}{\Pr\{\hat{A} > \hat{V}\}}. \tag{12}$$

Finally, if $\pi_{i,j}$'s denote the stationary probabilities at random times, then we can proceed as in Ross[12] to obtain:

$$\pi_{1,1} = \sum_{r=1}^{K-1} p_{1,r} d_r + p_{1,K} d_{K-1} + (1-\kappa)\left[\sum_{r=0}^{K-1} p_{0,r} d_r^+ + p_{0,K} d_{K-1}^+\right] \tag{13}$$

$$\pi_{1,j} = \sum_{r=0}^{K-j} p_{1,j-1+r} d_r + p_{1,K} d_{K-j} + (1-\kappa)\left[\sum_{r=0}^{K-j} p_{0,j-1+r} d_r^+ + p_{0,K} d_{K-j}^+\right]$$

$$\text{for } 2 \leqslant j \leqslant K-1 \tag{14}$$

$$\pi_{1,K} = (p_{1,K-1} + p_{1,K}) d_0 + (1-\kappa)[(p_{0,K-1} + p_{0,K}) d_0^+] \tag{15}$$

$$\pi_{0,0} = \sum_{r=1}^{K-1} p_{1,r} d_{r+1}^c + p_{1,K} d_K^c + (1-\kappa)\left[\sum_{r=0}^{K-1} p_{0,r} (d_{r+1}^+)^c + p_{0,K} (d_K^+)^c\right] \tag{16}$$

$$\pi_{0,j} = \kappa p_{0,j-1} \quad \text{for } 1 \leqslant j \leqslant K-1 \tag{17}$$

$$\pi_{0,K} = \kappa(p_{0,K-1} + p_{0,K}) \tag{18}$$

where $d_j^c = \sum_{r=j}^{\infty} d_r$, $(d_j^+)^c = \sum_{r=j}^{\infty} d_r^+$ and

$$\kappa = \int_0^\infty F_{\hat{A}}(x) f_{\hat{V}}(x)\, dx \tag{19}$$

(i.e. κ is the probability that $\hat{V}$ exceeds $\hat{A}$).

Blocking probabilities

A particularly important measure of performance for a finite buffer queue is the blocking probability. In our case, as the Markov Chain is embedded at arrival epochs, the blocking probabilities can be immediately obtained as:

$$\mathrm{P}\{\text{arriving customer is blocked}\} = p_{1,K} + p_{0,K}. \tag{20}$$

We will denote the blocking probability by P_b.

Waiting time distribution

Having computed the probability distribution at arrival epochs, it is straightforward to obtain the waiting time distribution. Due to the memoryless property of the service and vacation distributions, a customer who arrives when the system is in state $(0, j)$, $0 \leqslant j < K$ will have to wait for a vacation completion and j service completions. Similarly, if the customer arrives when the system is in state $(1, j)$, $1 \leqslant j < K$, the waiting time will consist of j service completions. These

translate into the following expression for the LST of the waiting time distribution, $W(s)$:

$$W(s) = \sum_{j=0}^{K-1} p_{0,j} \left(\frac{\mu}{\mu + s}\right)^j \left(\frac{\theta}{\theta + s}\right) + \sum_{j=1}^{K-1} p_{1,j} \left(\frac{\mu}{\mu + s}\right)^j. \tag{21}$$

NUMERICAL EXAMPLES

In this section, we present numerical examples using the exact solution of the embedded Markov chain. The purpose of the examples is two fold. Firstly, we are interested in the effects of the arrival process on the blocking probability. The renewal arrival process enables us to model non-Poisson arrivals with different variations. However, the motivation to use alternative arrival processes exists only if the performance of the system differs significantly from the Poisson arrival case. The second purpose of the numerical experimentation is the necessity to develop a practically usable heuristic for general arrival processes that is computationally more efficient than explicit solution of the embedded chain.

In addition to the deterministic and exponential arrival cases, we also experiment with mixtures of exponential distributions to obtain different coefficients of variation (CV). In particular, we use a hyperexponential distribution obtained by mixing two exponential distributions (denoted H_2) and the two stage generalized Erlang distribution (denoted E_2).

In the H_2 distributions used, we assume balanced means (as in Whitt[10]) which corresponds to the following density function:

$$f(x) = q\lambda_1 e^{-\lambda_1 x} + (1 - q)\lambda_2 e^{-\lambda_2 x} \quad x \geqslant 0 \tag{22}$$

where $0 \leqslant q \leqslant 1$ and $\lambda^{-1} = q\lambda_1^{-1} + (1 - q)\lambda_2^{-1}$. Note that, the assumption of balanced means corresponds to $q\lambda_1^{-1} = (1 - q)\lambda_2^{-1}$.

With the above definitions, it is possible to obtain different CVs (where CV $\geqslant 1$) by altering q. The CV of the distribution is given by:

$$\text{CV} = ((1 + (2q - 1)^2)/(1 - (2q - 1)^2))^{1/2} \tag{23}$$

As for E_2 distributions, we alter λ_1 and use the fact that $\lambda^{-1} = \lambda_1^{-1} + \lambda_2^{-1}$ to obtain λ_2. The density will be given by:

$$f(x) = \frac{\lambda_1 \lambda_2 (e^{-\lambda_1 x} - e^{-\lambda_2 x})}{(\lambda_2 - \lambda_1)} \quad x \geqslant 0 \tag{24}$$

when $\lambda_1 \neq \lambda_2$. When $\lambda_1 = \lambda_2 = 2\lambda$, this reduces to the regular 2 stage Erlang distribution with mean λ (i.e. $f(x) = (2\lambda)^2 x e^{-2\lambda x}$, $x \geqslant 0$). Note that, the CV of the E_2 distribution is given by:

$$\text{CV} = ((\lambda_1^2 + \lambda_2^2)/(\lambda_1 + \lambda_2)^2)^{1/2} \tag{25}$$

and is between $1/\sqrt{2}$ and 1.

Thus, in the examples, we start with a deterministic arrival process (D) that has CV = 0. Next we use the E_2 processes for CVs less than 1, the Poisson process (M) that has CV = 1 and the H_2 processes for CVs greater than 1. We will fix the buffer size K at 10, and the service rate μ at 1 and analyze the effect of different loads by varying λ. To have an understanding of the vacations effect we experiment with two vacation levels, the long vacations ($\theta = \mu/4$) and short vacations ($\theta = 4\mu$). The results are summarized in Table 1.

Figure 1 is a brief summary of Table 1. This is a typical display of the behaviour of the blocking probability which is monotonically increasing in ρ and the CV of the arrival process in all of the examples we ran.

In addition to supporting the monotonicity properties of the blocking probability, the numerical experiments indicate that the changes in the coefficient of variation of the arrival process can alter the blocking probability by a factor larger than 300,000 as is evident in row 1 of Table 1. This observation highlights the importance of estimating the arrival process correctly.

TABLE 1. *Exact blocking probabilities*

Short Vacations ($\theta = 4\mu$)									
ρ	D(0)*	E(0.71)	E(0.73)	E(0.77)	E(0.85)	M(1)	H(1.29)	H(1.71)	H(3.09)
0.5	8.9×10^{-8}	0.00004	0.00005	0.00007	0.00013	0.00056	0.00179	0.00555	0.02841
0.6	7.1×10^{-6}	0.0004	0.0005	0.0006	0.0010	0.0027	0.0069	0.0173	0.0683
0.7	0.0002	0.0026	0.0029	0.0035	0.0049	0.0094	0.0191	0.0392	0.1198
0.8	0.0023	0.0112	0.0119	0.0133	0.0165	0.0249	0.0414	0.0713	0.1739
0.9	0.0144	0.0330	0.0341	0.0363	0.0414	0.0527	0.0748	0.1109	0.2257
1.0	0.0496	0.0720	0.0733	0.0758	0.0813	0.0930	0.1168	0.1548	0.2731
1.1	0.1071	0.1243	0.1254	0.1275	0.1321	0.1419	0.1638	0.1999	0.3159
1.2	0.1716	0.1815	0.1823	0.1838	0.1870	0.1941	0.2122	0.2443	0.3543
1.3	0.2323	0.2374	0.2378	0.2388	0.2408	0.2455	0.2595	0.2867	0.3888
1.4	0.2862	0.2887	0.2890	0.2895	0.2907	0.2937	0.3041	0.3264	0.4199
1.5	0.3335	0.3347	0.3349	0.3352	0.3359	0.3378	0.3452	0.3633	0.4481
Long Vacations ($\theta = (1/4)\mu$)									
ρ	D(0)*	E(0.71)	E(0.73)	E(0.77)	E(0.85)	M(1)	H(1.29)	H(1.71)	H(3.09)
0.5	0.0052	0.0100	0.0103	0.0109	0.0125	0.0171	0.0248	0.0384	0.0800
0.6	0.0106	0.0186	0.0191	0.0202	0.0227	0.0300	0.0422	0.0634	0.1286
0.7	0.0172	0.0303	0.0311	0.0328	0.0368	0.00472	0.0650	0.0945	0.1825
0.8	0.0265	0.0471	0.0483	0.0507	0.0564	0.0699	0.0933	0.1303	0.2359
0.9	0.0433	0.0718	0.0734	0.0765	0.0835	0.0990	0.1267	0.1690	0.2854
1.0	0.0749	0.1063	0.1081	0.1115	0.1189	0.1345	0.1642	0.2087	0.3299
1.1	0.1227	0.1496	0.1513	0.1544	0.1611	0.1752	0.2041	0.2481	0.3695
1.2	0.1790	0.1980	0.1993	0.2018	0.2072	0.2186	0.2448	0.2862	0.4048
1.3	0.2354	0.2472	0.2481	0.2499	0.2539	0.2625	0.2849	0.3225	0.4361
1.4	0.2875	0.2943	0.2949	0.2961	0.2989	0.3051	0.3235	0.3567	0.4642
1.5	0.3340	0.3379	0.3382	0.3391	0.3409	0.3452	0.3598	0.3887	0.4895

* The values in parentheses denote the CVs.

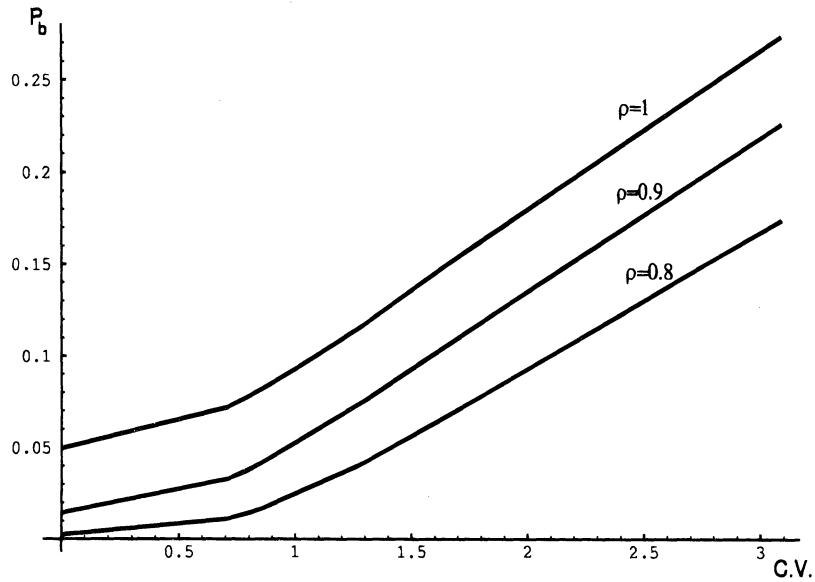

FIG. 1. *The blocking probability as a function of the CV.*

HEURISTICS FOR THE BLOCKING PROBABILITY

When arrival processes are modelled by more complicated probability distributions than the two stage exponential mixtures considered in the previous section, it is difficult to compute the transition probabilities in the embedded chain. In this case, two alternative strategies emerge. The

first strategy is modelling the arrival process with an E_2 or H_2 distribution with the same coefficient of variation as the original and computing the blocking probability exactly. The second strategy is directly approximating the blocking probability without fitting the distribution to E_2 or H_2 distributions. In this section, we employ the second strategy and develop an approximation scheme for the blocking probability.

As a first step in approximation, we propose heuristic bounds for the blocking probability. The numerical experiments suggest that the blocking probability is monotonically increasing in the CV of the arrival process. Therefore, special cases for CV for which the blocking probabilities are easy to compute give heuristic bounds for the more general cases. Below, we study the blocking probability for the case of Poisson and deterministic arrival processes. The heuristic argument is that when the CV of the arrival process is between 0 and 1, the blocking probability will be bounded by the blocking probabilities in the deterministic arrival and the Poisson arrival case. Similarly, when
the CV > 1, the blocking probability is bounded from below by the blocking probability of the Poisson arrival case.

The blocking probability for the M/M/1/K queue with vacations

Gupta[13] obtains the probability distribution of the M/M/1/K queue with multiple vacations in closed form. As the probability distribution is not needed here, an alternative approach to obtain the blocking probability will be used.

Vinod[14] uses matrix-geometric methods (see Neuts[15]) to give the following probability distribution for the infinite capacity M/M/1 queue with vacations:

$$p^{\infty}_{0,0} = \frac{\theta(\mu - \theta)}{\mu(\lambda + \theta)} \tag{26}$$

$$\boldsymbol{p}^{\infty}_1 = \left(\frac{\lambda\theta(\mu - \lambda)}{(\lambda + \theta)^2\mu)}, \frac{\lambda\theta(\mu - \lambda)}{(\lambda + \theta)\mu^2)}\right) \tag{27}$$

$$\boldsymbol{p}^{\infty}_j = \boldsymbol{p}^{\infty}_1 \boldsymbol{R}^{j-1} \quad \textit{for } j \geqslant 2 \tag{28}$$

where $\boldsymbol{p}^{\infty}_j = (p^{\infty}_{0,j}, p^{\infty}_{1,j})$ (for $j = 1, 2, \ldots$) and

$$\boldsymbol{R} = \begin{bmatrix} \dfrac{\lambda}{\lambda + \theta} & \dfrac{\lambda}{\mu} \\ 0 & \dfrac{\lambda}{\mu} \end{bmatrix}. \tag{29}$$

We, now relate the above probability distribution of the infinite capacity queue to that of the finite capacity queue using scaling properties of the queues with Poisson arrivals as studied by Keilson and Servi[3,4]. Keilson and Servi give the following relation that relates the probabilities of the infinite capacity queue to the blocking probability in the finite capacity queue:

$$\mathrm{P}\{\text{arriving customer is blocked}\} = (1 - \rho)\kappa/(1 - \rho\kappa). \tag{30}$$

where κ is the probability that there are K or more customers in the infinite capacity queue. Now, using the above results for the infinite capacity queue, we obtain:

$$\kappa = \sum_{i=K}^{\infty} \boldsymbol{p}^{\infty}_1 \boldsymbol{R}^{i-1}(1, 1)^T \tag{31}$$

At this point noting that:

$$p^{\infty}_{0,0} + \sum_{i=1}^{\infty} \boldsymbol{p}^{\infty}_1 \boldsymbol{R}^{i-1}(1, 1)^T = 1, \tag{32}$$

we get the following expression for κ:

$$\kappa = 1 - \left(p^{\infty}_{0,\,0} + \sum_{i=1}^{K-1} \boldsymbol{p}_1^{\infty} \boldsymbol{R}^{i-1}(1,\,1)^T \right) \tag{33}$$

$$= 1 - (p^{\infty}_{0,\,0} + \boldsymbol{p}_1^{\infty}(\boldsymbol{I} - \boldsymbol{R}^{K-1})(\boldsymbol{I} - \boldsymbol{R})^{-1}(1,\,1)^T) \tag{34}$$

Substituting the right hand side of equation (34) in (30), we have a formula for the blocking probability in the M/M/1/K queue. Since computing the Kth power of the matrix $\boldsymbol{R}$ requires a computation of the order of log K, the computational complexity of obtaining the blocking probability is $O(\log K)$.

Remarks

1. Note that both Vinod's[14] and Keilson and Servi's[3] relations are valid for $\rho < 1$. However, the formula (34) holds independently of this assumption. The matrix $\boldsymbol{R}$ has no probabilistic meaning.
2. Computing the blocking probability using equation (34) has a computational advantage over using the result in Gupta[13] which has a computational complexity of $O(K)$.
3. As the constant κ is obtained, the whole queue length distribution is explicitly available through the scaling property. In fact, this seems to be the only case where the scaling property leads to an explicit solution for queues with vacations.

The complexity of the deterministic arrival case

When the arrival process is deterministic, the transition probabilities in the embedded chain can be computed through simple exponential integrals. In particular:

$$g_i = \frac{\mathrm{e}^{-1/\rho}}{\rho^i i!}, \tag{35}$$

$$\omega = \mathrm{e}^{-\theta/\lambda} \tag{36}$$

and

$$h_i = \frac{\theta \mu^i}{(\mathrm{e}^{\theta/\lambda} - 1) i!} \int_0^{1/\lambda} \mathrm{e}^{(\theta-\mu)x} x^i \, dx. \tag{37}$$

Note that, h_i ($i = 0, 1, 2, \ldots, K$) can be computed recursively (see Appendix 2 for the explicit computation), therefore the computational complexity of obtaining the blocking probability is determined by the solution of the embedded chain. Ignoring the special structure of the Markov Chain and solving a $K \times K$ linear system of equations leads to a computational complexity of $O(K^3)$.

Heuristic bounds for the blocking probability

In general, obtaining the blocking probability for the GI/M/1/K queue with server vacations may require numerical integration. However, it was seen that for the special cases of Poisson and deterministic arrival processes the blocking probability can be computed very efficiently. On the other hand, numerical experimentation suggests that the blocking probability is monotonically increasing with respect to the CV of the arrival process. Combining these two pieces of information, one can establish bounds on the blocking probability utilizing the Poisson and deterministic arrival streams.

Table 1 displays this bounding behaviour. Whenever the CV of the arrival process is between the CVs of the deterministic and Poisson arrival processes, the blocking probability also stays between the corresponding blocking probabilities. Furthermore, whenever the arrival process is more variable than the Poisson process, the corresponding blocking probability is greater than

that of the Poisson. Figure 2 is extracted from Table 1 and displays the bounds. Note that the bounds become tighter when $\rho > 1$ and the blocking probability can be approximated through the bounds in practice. When ρ is close to zero the blocking probability becomes very small and the bounds seem to be tight numerically, however the percentage error is large.

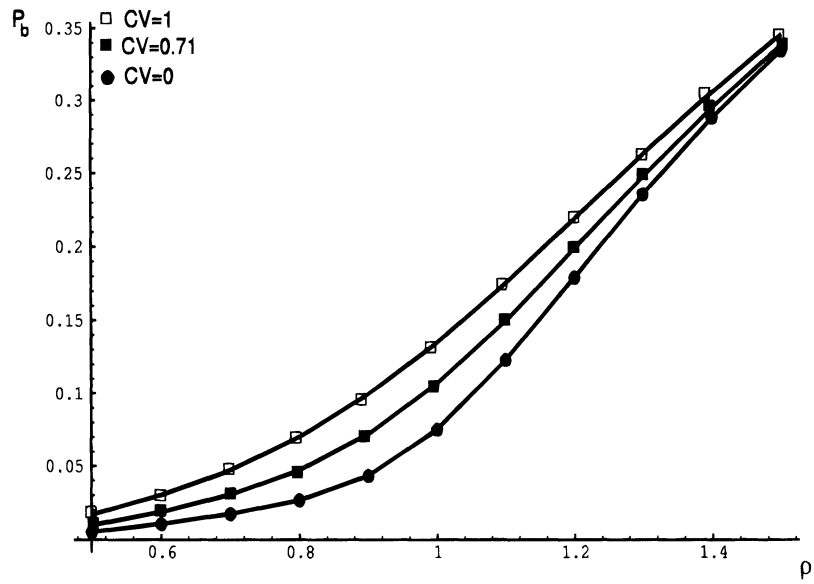

FIG. 2. *Typical behaviour of the bounds.*

Approximations using the infinite capacity queue

Although the proposed heuristic bounds provide a rough approximation of the blocking probability, better approximations may be useful in practice. A promising strategy is to relate the corresponding infinite capacity queue to the finite capacity queue on hand. In this case, explicit results for the infinite capacity queue are readily available which ensures the computational efficiency of the approximation. On the other hand, a drawback of this type of approximation is that it is only valid for the cases where the infinite capacity queue is stable ($\rho < 1$). Nevertheless, the approximation range ($0 < \rho < 1$) covers most models of practical interest.

It was recently reported in Tijms[16] that the following formula provides a good approximation for the blocking probability in the GI/G/c/K queue:

$$\mathrm{P}\{\text{arriving customer is blocked}\} = (1 - \rho)\kappa/(1 - \rho\kappa). \tag{38}$$

where κ is the probability that an arriving customer sees K or more customers in the infinite capacity GI/G/c queue with the same parameters. Numerical experimentation suggests the above approximation performs well for GI/G/c/K queues. On the other hand, by the results of Keilson and Servi[3], equation (38) is exact for the vacation system when the arrival process is Poisson. Encouraged by this fact, we tested the performance of the Tijms heuristic for the GI/M/1/K queue with exponential vacations.

As in the case of Poisson arrivals, the following expression can be obtained for κ:

$$\kappa = 1 - (p^{\infty}_{0,0} + \boldsymbol{p}^{\infty}_{1}(\boldsymbol{I} - \boldsymbol{R}^{K-1})(\boldsymbol{I} - \boldsymbol{R})^{-1}(1, 1)^{T}) \tag{39}$$

Tian *et al.*[7] give the matrix $\boldsymbol{R}$, $p^{\infty}_{0,0}$ and $\boldsymbol{p}^{\infty}_{1}$. It is required to solve a simple nonlinear equation to obtain the explicit expression but that presented no computational difficulty in the examples considered.

Table 2 displays the approximate blocking probabilities for the identical systems as in Table 1. Figure 3 displays a comparison of the approximation with the exact results for the cases of short and long vacations respectively.

TABLE 2. *Approximate blocking probabilities*

Short Vacations ($\theta = 4\mu$)						
ρ	E(0.73)*	E(0.77)	E(0.85)	H(1.29)	H(1.71)	H(3.09)
0.5	0.00005	0.00007	0.00013	0.00193	0.00667	0.04740
0.6	0.0005	0.0006	0.0010	0.0074	0.0206	0.1124
0.7	0.0029	0.0034	0.0048	0.0203	0.0458	0.1850
0.8	0.0117	0.0131	0.0164	0.0437	0.0810	0.2503
0.9	0.0337	0.0361	0.0412	0.0778	0.1228	0.3050
Long Vacations ($\theta = (1/4)\mu$)						
ρ	E(0.73)*	E(0.77)	E(0.85)	H(1.29)	H(1.71)	H(3.09)
0.5	0.0101	0.0107	0.0124	0.02623	0.04448	0.1286
0.6	0.0187	0.0198	0.0226	0.0447	0.0738	0.2090
0.7	0.0305	0.0323	0.0366	0.0687	0.1093	0.2844
0.8	0.0475	0.0501	0.0561	0.0982	0.1487	0.3464
0.9	0.0724	0.0758	0.0831	0.1326	0.1895	0.3957

* The values in parentheses denote the CVs.

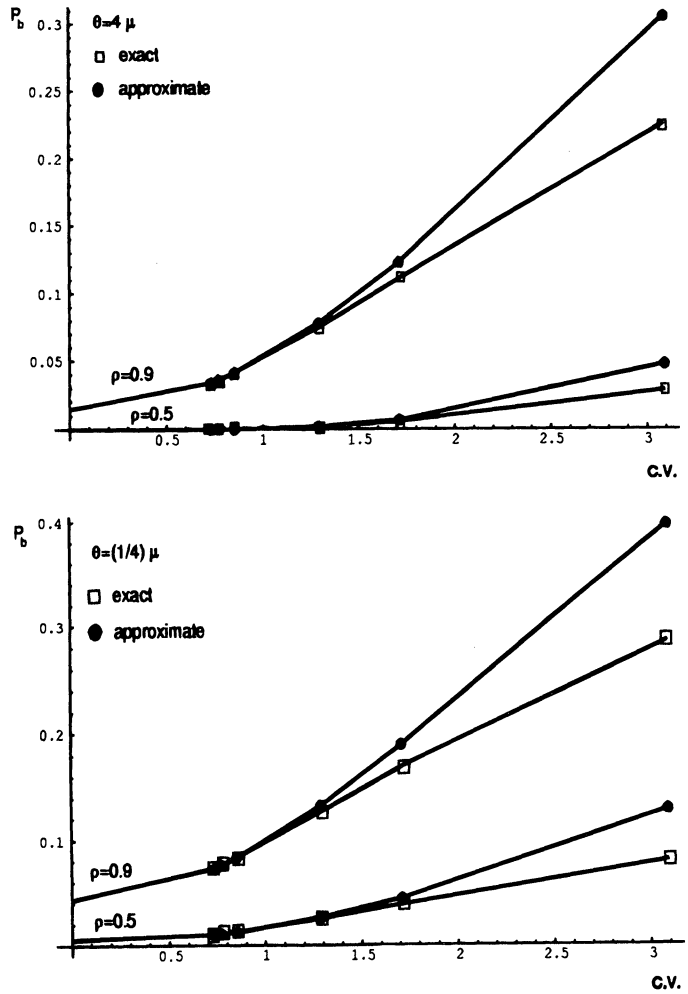

FIG. 3. *A comparison of the approximation with the exact results.*

Our numerical experimentation suggests that the approximation is very good when the CV of the arrival process is either less than 1 or close to 1. Furthermore, all numerical results indicate that the approximation underestimates the blocking probability when CV < 1 and overestimates the blocking probability when CV > 1. This property is conjectured in Miyazawa and Tijms[17] for GI/G/c/K queues. Our numerical experiments support this conjecture.

CONCLUSIONS

Research in performance analysis of queueing systems has focused on finding general ways to model tractable service processes while restricting the arrival process. This is considered a good trade-off as in many cases service processes may not be exponentially distributed whereas the Poisson approximation to arrival processes is acceptable. On the other hand, there are cases where the arrival process is non-Poisson such as in the case of automated arrivals to a manufacturing centre. Our results show that poor estimation of the arrival process to such nodes can cause large errors in performance measures. This implies that in certain cases it might be better to use a general model of the arrival process and restrict the service processes including vacations and interruptions.

The model used here captures a reasonable approximation of the service *vs.* vacations trade-off. This enables the decision maker to experiment with different service/maintenance (vacation) ratios to develop a better schedule. In that direction the approximation procedures are fast enough to be useful in practical *what if* type analysis and the exact procedures provide accurate solutions in reasonable time.

To approximate networks consisting of nodes with interruptions, the blocking probability for the GI/M/1/K queue with vacations has to be computed efficiently. We showed that the blocking probability can be obtained explicitly for the M/M/1/K case. Note that, the approach used to obtain the blocking probability immediately leads to explicit expressions for the stationary queue length distribution. This is of practical interest as it provides an easily obtained benchmark and can be used in approximations. Further numerical results suggest that explicit solutions for special cases can be used as heuristic bounds which may be useful in practical computation. Finally, the performance of an efficient approximation algorithm for the blocking probability was tested. It was seen that the approximation is accurate for the practically interesting range of parameters of the arrival process. Therefore the approximation algorithm also has the potential to be integrated in manufacturing network approximation algorithms.

APPENDIX 1

Consider equation (2b) of Chatterjee and Mukherjee[6]:

$$p_{0,j} = \mathrm{P}(\hat{V} > A)p_{0,j-1} \tag{40}$$

This equation fails to hold when there are more than one arrivals during a vacation period because the arrival process does not constitute a random incidence for the vacation distribution anymore. For example, if one arrival has already occured during the vacation period, the second arrival occurs during the forward recurrence time of the vacation which has the distribution $F_{\hat{V}}$ (not F_V). In this case, the relationship in equation (40) has to be corrected as:

$$p_{0,j} = \mathrm{P}(\hat{V}_{(2)} > A)p_{0,j-1}. \tag{41}$$

where $\hat{V}_{(2)}$ denotes the forward recurrence time of $\hat{V}$. Note that, the relationship (40) holds for the special case where vacations are exponentially distributed because in that case the forward recurrence time of a vacation is equal to itself due to the memoryless property of exponential distribution, i.e.:

$$\hat{V} = V. \tag{42}$$

APPENDIX 2

In the deterministic arrival case, h_i ($i = 0, 1, 2, \ldots, K$) can be computed recursively. Below, we give the explicit expressions for this recursion. Let

$$C_i = \frac{\theta\mu^i}{(e^{\theta/\lambda} - 1)i!} \tag{43}$$

and

$$v_i = \int_0^{1/\lambda} e^{(\theta-\mu)x} x^i \, dx. \tag{44}$$

then

$$h_0 = C_0 v_0 \tag{45}$$

$$= C_0 \frac{e^{-(\mu+\theta)/\lambda} - 1}{(\theta - \mu)} \tag{46}$$

Using integration by parts, for $i \geqslant 1$, we have:

$$v_i = \frac{(1/\lambda)^i e^{(\theta-\mu)/\lambda} - iv_{i-1}}{(\theta - \mu)}. \tag{47}$$

Finally, for $i = 1, 2, \ldots, K$, h_i can be obtained as:

$$h_i = C_i v_i \tag{48}$$

$$= C_i \frac{(1/\lambda)^i e^{(\theta-\mu)/\lambda} - iv_{i-1}}{(\theta - \mu)}. \tag{49}$$

REFERENCES

1. P. Courtois (1980) The M/G/1 finite capacity queue with delays. *IEEE Trans. on Comm.* **COM-28**, 165–172.
2. T. Lee (1984) The M/G/1/N queue with vacation time and exhaustive service discipline. *Opns Res.* **32**, 774–784.
3. J. Keilson and L. Servi (1989) Blocking probability for M/G/1 vacation systems with occupancy level dependent schedules. *Opns Res.* **37**, 134–140.
4. J. Keilson and L. Servi (1993) The M/G/1/K blocking formula and its generalizations to state-dependent vacation systems and priority systems. *Que. Sys.* **14**, 111–123.
5. C. Blondia (1991) Finite capacity vacation models with non-renewal input. *J. Appl. Prob.* **28**, 174–197.
6. U. Chatterjee and S. Mukherjee (1990) GI/M/1 queue with server vacation. *J. Opl Res. Soc.* **41**, 83–87.
7. N. Tian, D. Zhang and C. Cao (1989) GI/M/1 queue with exponential vacations. *Que. Sys.* **5**, 331–344.
8. Y. Dallery and Y. Frein (1989) A decomposition method for the approximate analysis of closed queueing networks with blocking. In *First International Workshop on Queueing Systems with Finite Buffers.* (H. Perros and T. Altiok, Eds.) pp 193–215. Elsevier North Holland, Amsterdam.
9. S. Gershwin (1989) An efficient decomposition algorithm for unreliable tandem queueing systems with finite buffers. In *First International Workshop on Queueing Systems with Finite Buffers.* (H. Perros and T. Altiok, Eds.) pp 127–146. Elsevier North Holland, Amsterdam.
10. W. Whitt (1981) Approximating a point process by a renewal process. *Mgmt. Sci.* **27**, 619–636.
11. S. Albin (1984) Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Opns Res.* **32**, 1133–1162.
12. S. Ross (1983) *Stochastic Processes.* Wiley, New York.
13. S. M. Gupta (1996) Machine interference problem with warm spares, server vacations and exhaustive service. *Perf. Eval.*, accepted for publication.
14. B. Vinod (1986) Exponential queues with server vacations. *J. Opl Res. Soc.* **37**, 1007–1014.
15. M. Neuts (1981) *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore.
16. H. Tijms (1992) Heuristics for finite buffer queues. *Prob. Eng. Info. Sci.* **6**, 277–285.
17. M. Miyazawa and H. Tijms (1993) Comparisons of two approximations for the loss probability in finite buffer queues. *Prob. Eng. Info. Sci.* **7**, 19–27.

Received May 1995; accepted November 1995 after one revision