PARALLEL SERVICE WITH VACATIONS

SID BROWNE

Columbia University, New York, New York

OFFER KELLA

Hebrew University, Jerusalem, Israel

(Received May 1992; revision received June 1993; accepted July 1993)

We study a system with unlimited service potential where all service requests are served in parallel. The entire system itself becomes unavailable for a random period of time at the first instance that the system becomes idle. A queue builds up while the system is unavailable, and then all waiting customers enter the system simultaneously—each to its own processor—when the system becomes available again. All customers who arrive to find the system in operation proceed directly into service. The analysis of this system entails finding the distribution of the *delayed* busy period of an $M/G/\infty$ queue. The steady-state distribution of the number of customers in the system is obtained for the special cases of exponential and deterministic service times. Among other applications, our results enable us to analyze and solve for the optimal N-policy for the systems with unlimited service potential. We also study a multiclass model of a polling system with exhaustive service.

Many operational problems that arise in determining efficient strategies for the management of service and production systems can be analyzed in a unified manner within the construct of "systems with server vacations." In this class of models (see, e.g., Doshi 1986 for a survey), the server (i.e., system) becomes unavailable for a random period of time (called the *vacation*), that is initiated at the beginning of an idle period. A classical example is the problem of efficient start-up and shut-down policies for production and inventory systems. In the simplest version of this problem (see, e.g., Yadin and Naor 1963, Heyman 1968, Sobel 1969, and Bell 1971), there is a production facility with *one* machine, which produces a single item. Orders arrive according to a simple Poisson process, and the processing time of an individual order is an independent random variable with distribution $G(\cdot)$. There is a cost of "turning on" the machine and starting a production run, as well as a holding cost for keeping the orders unfilled. All orders that arrive while a production run is in progress will be filled during that production run. In many cases, the machine automatically shuts off when the production run is over (i.e., no jobs waiting in the queue). It is clear that under certain conditions it does not pay to turn on the machine when a single order arrives, because there is a good chance that no other orders will arrive during the processing of that first order, and then the machine will shut off, and will have to be restarted for the next (single) order. Instead, it might pay to wait until a given number of orders have arrived, say N, thus assuring a production run that will complete N orders at least. But since the facility will be 'open' for at least N orders, there is now a good chance that many more orders will arrive *during* the production run, thereby eliminating the need to start individual production runs for these newly arriving jobs. This is the well known N-policy for the $M/G/1$ queue. It can be viewed as a very simple vacation model, specifically, where the vacation time is the time from when the production run ends (and the idle period starts) until the N orders arrive. Other variants of this problem directly applicable to determining optimal operating policies for production and inventory systems, such as the T-policy and D-policy are described in Heyman and Sobel (1982, section 11.6). Kella (1989) discusses the optimal threshold for a hybrid system of vacations with the N-policy.

Another classical application of the vacation model is single server multiclass customer queueing systems with nonpreemptive priorities (e.g., Heyman and Sobel, section 11.5). In this model, service to the lower priority customers constitutes vacations with respect to the higher priority customer, because the server cannot begin service to the higher priority customer until the service to the lower priority customer (who is in service at the time the higher priority customer arrives) is completed. One important model of interest in communication as well as production systems is a *polling system*, where a single server sequentially attends to the service at the n individual queues (or stations) in a cyclic pattern (Cooper 1970, Takagi 1986). Under the popular *exhaustive* service discipline, the server cannot move from a queue until that queue is empty of customers. The intercycle times are considered the vacation with respect to an individual queue.

While the literature on the theory and applications of vacation models is quite extensive (Doshi 1986, Kella and Whitt 1991), almost all studies have concentrated on the single server case and ignored the multiple server case.

Subject classifications: Probability: stochastic model applications. Queues: busy period analysis, multichannel, optimization transient results.
Area of review: STOCHASTIC ANALYSIS AND THEIR APPLICATIONS.

0030-364X/95/4305-0870 $01.25

In this paper, we are interested in multiple server systems where all servers are synchronized, in that all servers come from, and go on vacations *simultaneously* as a synchronized unit (see Levy and Yechiali 1976 for a model where exponential servers go on vacations *individually*).

One major component in the analysis of vacation systems is the *delayed busy period* caused to the (primary) customers by the vacation time. In the case of *serial* processing, i.e., a single server, and Poisson arrivals, this delayed busy period is relatively easy to analyze (in fact, they are compound Poisson with respect to the delays; see Prabhu 1980, page 81), due to the essential linearity of the system in that the server works at the unit rate. However, little is known about the case of *parallel* processing—with a service unit being comprised of multiple servers—even though many current production and service systems utilize many processors (servers) working together in parallel. To motivate our results, consider a production facility with k machines capable of producing a single item. Orders for this item are still assumed to arrive according to a Poisson process. The machines are coupled together and synchronized, in that the machines are not turned on and off individually, but rather simultaneously, as an entire facility. If there is a significant cost to turning on the facility, one central operational issue is again to determine efficient start-up and shut-down policies. An N policy is again a good candidate, i.e., a policy which waits until N orders have arrived (and are enqueued) before turning on the facility for a production run. The facility is then turned off at the termination of the production run, when there are no longer any orders outstanding. Under this policy, the length of a production run is equivalent to a busy period in an $M/G/k$ queue that is initiated with N customers, i.e., a delayed busy period. This is a simple $M/G/k$ vacation model with the vacation again being the time until N customers arrive. To determine the operating characteristics of this policy, one therefore needs to know, among other things, the distribution of the length of a delayed busy period in the $M/G/k$. (A note on terminology: We use the term busy period in multiple server systems to mean the time until *all* servers are idle.) Unfortunately, except for very special service time distributions (e.g., exponential), the even simpler *ordinary* busy period of the $M/G/k$ system (see e.g., deSmit 1973, Weins 1989), is already quite complicated (because the output process during the busy period is no longer linear as many jobs are worked on simultaneously), which, in turn, leaves the exact analysis of the more complicated delayed busy period (and hence the vacation model) almost completely unattainable. Therefore, the development of approximations are in order.

Typically, approximations of performance measures for systems involving the $M/G/k$, involve a related model with *unlimited* service potential, i.e., an $M/G/\infty$ system (see e.g., Tijms 1986, subsection 4.4.3). While the $M/G/\infty$ system is still highly nonlinear, in contrast to the $M/G/k$ system, the distribution of the ordinary busy period is available in explicit form (Takács 1956, Shanbag 1966, and Stadje 1985), which makes the delayed busy period distribution amenable for analysis. Therefore, with a view toward eventually developing some approximations for (more realistic) finite systems, in this paper we study the $M/G/\infty$ model with system vacations, where all servers go on vacation together when the system becomes empty. This paper is obviously just a first step in that direction, although we believe our results are of interest in their own right. A system with infinite servers who utilize a *gate* between serving successive stages of customers, and then go on vacation together when the system is empty is analyzed in Browne et al. (1992a, b).

The remainder of the paper is organized as follows. Section 1 introduces our model and notation. The Laplace transform for a busy period in an $M/G/\infty$ queue that is initiated by k customers is explicitly solved for. From this we obtain the Laplace transform of the cycle time for the vacation model. While the transform cannot, in general, be explicitly inverted to obtain the corresponding distribution function (this is also true for the simpler ordinary $M/G/\infty$ system; see Stadje), we are able to use our results to obtain the first two moments of the cycle time explicitly. Section 2 returns to the start-up and shut-down problem described above, and solves for the optimal N-policy for a system with an unlimited number of servers, where there is a cost to start a production run, as well as holding costs. We characterize the optimal threshold in this case, which is related to a classical EOQ model. As noted, similar problems have received much attention in the literature for the case of *serial* processing; see, for example, Bell and Heyman and Sobel, where one of the objectives is to find the optimal number of customers in the queue at which to start the system.

In Sections 3 and 4 we obtain some more explicit forms by analyzing the case of exponential service, as well as the deterministic case. Note that the case of exponential service times also corresponds to a *single* server model with *state-dependent service rates* $\mu_n = n\mu$, and that our results are the first on general vacations (to our knowledge) that pertain to such models. (For vacations with a state-dependent arrival rate, see Shanthikumar 1988, and for state-dependent vacations, see Harris and Marchal 1988). Section 5 analyzes a model of a polling system with two customers classes, namely the *alternating priority* infinite server queue.

1. THE MODEL, NOTATION, AND BUSY PERIODS

Consider an $M/G/\infty$ system in which whenever there are no customers in the system, service throughout the system is shut down for a random length of time, which we call a *vacation*. When a vacation ends, if there are customers waiting, the system is turned on and remains that way until it is empty. Otherwise, a second vacation takes

place. We assume that interarrival times, service times, and vacation times are independent, that each sequence is identically distributed, and the service times and vacations have finite first and second moments, respectively. To study this system, we use the following notation:

V = a random variable having the vacation time distribution;
S = a random variable having the service time distribution;
λ = the arrival rate;
$F_X(t) = P[X \leq t]$ = the distribution function (d.f.) of X;
EX = the expected value of X;
$\mathrm{Var}(X)$ = the variance of X;
$\tilde{X}(\alpha) = Ee^{-\alpha X}$ = the Laplace-Stieltjes transform (LST) of a nonnegative random variable X;
R_X = a random variable having the stationary forward recurrence time distribution of a nonnegative random variable X, i.e.,

$$F_{R_X}(t) = \int_0^t [1 - F_X(x)]\, dx/EX,$$

$$\tilde{R}_X(\alpha) = [1 - \tilde{X}(\alpha)]/(\alpha EX),$$

and

$$ER_X^k = EX^{k+1}/[(k+1)EX];$$

θ_k = a random variable having the distribution of the busy *period* (the time until the system is empty) for a system starting with k customers at time zero, all of whom begin service at the same instant (implicitly, $\theta_0 \equiv 0$);
θ_V = a random variable having the distribution of the busy *period*, starting from the *end* of a vacation (can be zero if no customers are present);
$B_V = V + \theta_V$ = a random variable having the distribution of a busy *cycle* starting from the *beginning* of a vacation;
$\rho(t) = \lambda \int_0^t [1 - F_S(x)]\, dx = \lambda E \min(S, t)$;
$\rho = \lim_{t\to\infty} \rho(t) = \lambda ES$;
$p_k(t) = e^{-\rho(t)}[\rho(t)^k/k!]$ = the probability that there are k customers in a standard $M/G/\infty$ system at time t, starting from an empty system.

If we consider a standard $M/G/\infty$ queue in which at time zero there are k customers all starting service at the same time, then the probability that at time t there are n customers present is given by

$$\pi_{kn}(t) = \sum_{i=0}^{\min(n,\,k)} b_{ki}(t)p_{n-i}(t), \tag{1}$$

where

$$b_{ki}(t) = \binom{k}{i}[1 - F_S(t)]^i F_S(t)^{k-i}. \tag{2}$$

In particular, for $n = 0$ we have that

$$\pi_{k0}(t) = b_{k0}(t)p_0(t) = F_S(t)^k e^{-\rho(t)}. \tag{3}$$

We begin with the following result.

Lemma 1. *For every $k \geq 0$ and $\alpha > 0$ we have*

$$\tilde{\theta}_k(\alpha) = \frac{A_k(\alpha)}{A_0(\alpha)}, \tag{4}$$

where

$$A_k(\alpha) = \int_0^\infty e^{-\alpha t}\pi_{k0}(t)\, dt, \tag{5}$$

whence

$$E\theta_k = e^\rho \lim_{\alpha \downarrow 0}[A_0(\alpha) - A_k(\alpha)]$$

$$= e^\rho \int_0^\infty [1 - F_S(t)^k]e^{-\rho(t)}\, dt, \tag{6a}$$

and

$$E\theta_k^2 = 2e^\rho[E\theta_k \int_0^\infty (e^{-\rho(t)} - e^{-\rho})\, dt + \int_0^\infty t(1 - F_S(t)^k)e^{-\rho(t)}\, dt]. \tag{6b}$$

Proof. Although we can imitate the ideas in Takács (1956) (where the LST of θ_1 is obtained in the context of the dead time in a type-2 particle counter), and in Browne and Steele (1992) (where the remaining busy period in an ordinary $M/G/\infty$ system is studied), the following seems more direct and elementary. Starting at time 0 with k customers, all beginning their service at the same time, we let the system run without interruption, so that all subsequent busy periods will start with the arrival of a *single* customer. Let $\{\theta_{1i} | i \geq 1\}$ be the sequence of these subsequent busy periods and let $\{I_i | i \geq 1\}$ be the sequence of empty periods (idle times). Then θ_k, $\{\theta_{1i} | i \geq 1\}$ and $\{I_i | i \geq 1\}$ are all independent, and for every $i \geq 1$, θ_{1i} is distributed like θ_1 and I_i has an exponential distribution with parameter λ. For the purpose of this proof only set $T_0 = \theta_k$ and

$$T_n = \theta_k + \sum_{i=1}^{n} (I_i + \theta_{1i}) \tag{7}$$

for $n \geq 1$. (Note that in this notation, we have suppressed the dependency of T_0 and T_n on k.) Then, with $X(t) = 1$ if the system is empty at time t, and $X(t) = 0$ otherwise, we have that

$$A_k(\alpha) = \int_0^\infty e^{-\alpha t}\pi_{k0}(t)\,dt = E\int_0^\infty e^{-\alpha t}X(t)\,dt$$
$$= E\sum_{n=0}^{\infty}\int_{T_n}^{T_n+I_{n+1}} e^{-\alpha t}\,dt$$
$$= \sum_{n=0}^{\infty} Ee^{-\alpha T_n}\frac{1-Ee^{-\alpha I_{n+1}}}{\alpha} \tag{8}$$
$$= \sum_{n=0}^{\infty}\tilde{\theta}_k(\alpha)\left(\frac{\lambda}{\lambda+\alpha}\tilde{\theta}_1(\alpha)\right)^n\frac{1}{\lambda+\alpha}$$
$$= \frac{\tilde{\theta}_k(\alpha)}{\lambda+\alpha-\lambda\tilde{\theta}_1(\alpha)}.$$

It is not hard to see that the same holds for $k = 0$ (with $\tilde{\theta}_0(\alpha) \equiv 1$) so that

$$\tilde{\theta}_k(\alpha) = \frac{\tilde{\theta}_k(\alpha)}{\tilde{\theta}_0(\alpha)} = \frac{A_k(\alpha)}{A_0(\alpha)}. \tag{9}$$

To obtain (6a) we first observe that a change of variables and dominated convergence give

$$\alpha A_0(\alpha) = \int_0^\infty e^{-(t+\rho(t/\alpha))}\,dt \to e^{-\rho} \tag{10}$$

as $\alpha \downarrow 0$ (the integrand is dominated by e^{-t}, and $\lim_{\alpha\downarrow 0}\rho(t/\alpha) = \lim_{t\to\infty}\rho(t) = \rho$). Since $E\theta_k = \lim_{\alpha\downarrow 0}[1-\tilde{\theta}_k(\alpha)]/\alpha$, (6a) is obtained when we use the fact that $\lim_{\alpha\downarrow 0}(f(\alpha)/g(\alpha)) = \lim_{\alpha\downarrow 0} f(\alpha)/\lim_{\alpha\downarrow 0} g(\alpha)$, where $f(\alpha) = A_0(\alpha) - A_k(\alpha)$, and $g(\alpha) = \alpha A_0(\alpha)$, and both limits exist. Note that $E\theta_1 = (e^\rho - 1)/\lambda$, which is well known for the ordinary $M/G/\infty$ busy period.

To obtain (6b) first note that since $1 - \tilde{\theta}_k(\alpha) = \alpha E\theta_k[1 - \alpha E(\theta_k^2)/2E\theta_k + o(\alpha)]$, we have

$$\frac{1}{1-\tilde{\theta}_k(\alpha)} = \frac{A_0(\alpha)}{A_0(\alpha)-A_k(\alpha)}$$
$$= \frac{1}{\alpha E\theta_k} + \frac{E(\theta_k^2)}{2(E\theta_k)^2} + o(1). \tag{11}$$

Next, observe that $A_0(\alpha) - A_k(\alpha)$ admits an expansion around zero which, using (6a), can be expressed as $e^{-\rho}E\theta_k - \alpha\int_0^\infty te^{-\rho(t)}(1 - F_S(t)^k)\,dt + o(\alpha)$, and therefore we have

$$\frac{A_0(\alpha)}{A_0(\alpha)-A_k(\alpha)} = \frac{e^\rho A_0(\alpha)}{E\theta_k} + \frac{e^{2\rho}\alpha A_0(\alpha)}{(E\theta_k)^2}$$
$$\cdot\int_0^\infty te^{-\rho(t)}(1-F_S(t)^k)\,dt + o(1). \tag{12}$$

Now equate (11) and (12), to get

$$\frac{E(\theta_k^2)}{2(E\theta_k)^2} = \frac{1}{E\theta_k}\left[e^\rho A_0(\alpha) - \frac{1}{\alpha}\right] + \frac{e^{2\rho}\alpha A_0(\alpha)}{(E\theta_k)^2}$$
$$\cdot\int_0^\infty te^{-\rho(t)}(1-F_S(t)^k)\,dt + o(1),$$

and then use $e^\rho A_0(\alpha) - 1/\alpha = \int_0^\infty e^{-\alpha t}(e^{\rho-\rho(t)} - 1)\,dt$ and (10) to see that (6b) is obtained as $\alpha \downarrow 0$.

With Lemma 1 we are now prepared to study the LSTs of θ_V and B_V. In the following theorem $\rho'(t) = \lambda[1 - F_S(t)]$, that is, the derivative of $\rho(t)$.

Theorem 1. *For every* $\alpha > 0$,

$$\tilde{\theta}_V(\alpha) = \frac{1}{A_0(\alpha)}\int_0^\infty \tilde{V}[\rho'(t)]e^{-[\alpha t+\rho(t)]}\,dt \tag{13}$$

and

$$\tilde{B}_V(\alpha) = \frac{1}{A_0(\alpha)}\int_0^\infty \tilde{V}[\alpha+\rho'(t)]e^{-[\alpha t+\rho(t)]}\,dt \tag{14}$$

whence

$$E\theta_V = EB_V - EV = EV\int_0^\infty \tilde{R}_V(\rho'(t))e^{\rho-\rho(t)}\rho'(t)\,dt \tag{15a}$$

and

$$E\theta_V^2 = 2e^\rho\left[E\theta_V\int_0^\infty (e^{-\rho(t)} - e^{-\rho})\,dt + EV\int_0^\infty t\tilde{R}_V(\rho'(t))e^{-\rho(t)}\rho'(t)\,dt\right] \tag{15b}$$

Proof. Equation (13) follows from

$$\tilde{\theta}_V(\alpha) = \sum_{k=0}^{\infty} Ee^{-\lambda V}\frac{(\lambda V)^k}{k!}\tilde{\theta}_k(\alpha) \tag{16}$$

and some manipulations, while (14) follows from

$$\tilde{\theta}_V(\alpha) = \sum_{k=0}^{\infty} Ee^{-\lambda V}\frac{(\lambda V)^k}{k!}e^{-\alpha V}\tilde{\theta}_k(\alpha). \tag{17}$$

For (15a, b) repeat the argument in Lemma 1 or observe that for $i = 1, 2$,

$$E\theta_V^i = \sum_{k=0}^{\infty} Ee^{-\lambda V}\frac{(\lambda V)^k}{k!}E\theta_k^i. \tag{18}$$

The proof is complete.

Remark. If we denote the inverse function of $f(t) = 1 - e^{-\rho(t)}$ by $g(t)$, then substitution in (15a) gives

$$E\theta_V = e^\rho EV\int_0^{1-e^{-\rho}} \tilde{R}_V\left(\frac{1}{g'(x)(1-x)}\right)dx. \tag{19}$$

In particular, since $\tilde{R}_V(\cdot) \le 1$ and $[g'(x)(1-x)]^{-1} = \rho'(g(x)) \le \lambda$, this implies simple upper and lower bounds for the expected length of a busy cycle:

$$EV[1 + \tilde{R}_V(\lambda)(e^\rho - 1)] \le EB_V \le e^\rho EV. \tag{20}$$

2. THE EXPECTED NUMBER OF CUSTOMERS IN THE SYSTEM AND OPTIMIZATION

Here we will consider two models. The first model starts with an empty system with the service station being shut off. When the kth customer arrives ($k \geq 1$) the service station is turned on and keeps working until the system empties, after which this procedure repeats itself. This kind of policy is often referred to in the literature as an N-policy, but we will use the index k. It has been studied mostly with respect to the $M/G/1$ model (e.g., Bell 1971, Heyman and Sobel 1982, and Kella 1989). The second model is the one with vacations, as described in Section 1. Since the process that describes the number of customers in the system for each of these models is regenerative with finite mean nonarithmetic cycle times (in fact, due to the Poisson arrivals, the cycle time distributions are absolutely continuous), the existence of a limiting distribution is assured. In particular, if we let $L_k(t)$, $L_V(t)$ be the number of customers in the system at time t for the first and second models, respectively, and we let B_k be the cycle time for the first model (we have already denoted the second by B_V), then the expected value of the limiting distributions of these processes is given by

$$l_i = \frac{1}{EB_i} E \int_0^{B_i} L_i(t)\, dt \quad i = k, V. \tag{21}$$

We already know the value of EB_V from (15a). Also, it is clear that $EB_k = (k/\lambda) + E\theta_k$, where $E\theta_k$ is given by (6a).

Theorem 2. *For $k \geq 1$ and V such that $EV < \infty$,*

$$l_k = \rho + \frac{(k-1)}{2}\pi_k, \qquad l_V = \rho + \lambda E R_V \pi_V \tag{22}$$

where

$$\pi_k = \frac{k}{k + \lambda E\theta_k}, \qquad \pi_V = \frac{EV}{EV + E\theta_V} \tag{23}$$

are the corresponding limiting probabilities that each system is shut down.

Remark. It is interesting to note that there is considerable resemblance between (22) and corresponding first moment results for the $M/G/1$ queue. In fact, for the latter system, if $\rho < 1$, a similar result holds except that we replace ρ with $\lambda^2 ES^2/2(1-\rho)$ (the steady-state expected number of customers in the system for an $M/G/1$ queue) and π_k and π_V by 1 (e.g., Fuhrmann and Cooper 1985).

Proof. We start by noting that the integral(s) in (21) can be written as a sum of two integrals. The first is the integral of $L_i(t)$ until the system is turned on, and the second is the integral from that epoch until the end of the cycle. Let $N(t)$ denote the Poisson arrival process. For $k \geq 1$ it is clear that the contribution to the expected value from the first integral is

$$\sum_{i=1}^{k} \frac{i-1}{\lambda} = \frac{k(k-1)}{2\lambda}, \tag{24}$$

while the corresponding contribution for the model with vacations is

$$E\int_0^V N(t)\, dt = E\int_0^V (\lambda t)\, dt = \lambda EV^2/2, \tag{25}$$

where the first equality follows by conditioning on V and applying Fubini's theorem. The second integral is simply the sum of all the service times of the customers that arrived during the entire busy cycle (from time zero). Wald's identity immediately implies that this gives $ESEN(B_i)$ for $i = k, V$. To complete the derivation we use the fact that since $N(t) - \lambda t$ is a martingale (with respect to the filtration generated by the process $L_i(\cdot)$) it follows by Doob's optional sampling theorem that $EN(\min(B_i, t)) = \lambda E \min(B_i, t)$ and by monotone convergence that $EN(B_i) = \lambda EB_i$. With (21) and some straightforward manipulations, the proof is complete.

We conclude this section with an optimization problem. Consider the model with N-policy. Assume that there is a holding cost for each customer of h per-unit time and a setup cost of K for each cycle. How this setup cost is charged is of no consequence. For example, part of it could be charged when the station is turned off and the rest when it is turned on. We would like to find k which minimizes our long-run average costs

$$C_k = hl_k + \frac{K}{EB_k}. \tag{26}$$

Interestingly enough, this problem has a nice simple answer.

Theorem 3. *Let*

$$k^* = \min\left\{k \middle| \frac{k(k+1)}{2\lambda} \geq \frac{K}{h}\right\} \tag{27}$$

then for every $k \geq k^*$ we have that $C_k < C_{k+1}$, hence

$$\inf_{k \geq 1} C_k = \min_{1 \leq k \leq k^*} C_k. \tag{28}$$

Remark. It is interesting to note the following well-known related problem. Consider a discrete EOQ inventory model with demands that form a renewal counting process with rate λ, that is, the interdemand times are i.i.d. with mean λ^{-1}. Assume that the holding cost is h and the fixed ordering cost is K. Then it is well known and easy to check that k^* from (27) is precisely the value which minimizes the long-run average costs.

Proof. For $k \geq 0$, letting

$$a_k = \frac{1}{\lambda} + e^{\rho} \int_0^{\infty} F_S(t)^k [1 - F_S(t)] e^{-\rho(t)}\, dt \tag{29}$$

it is clear that $\{a_k | k \geq 0\}$ forms a decreasing sequence. Noting that $EB_k = \sum_{i=0}^{k-1} a_i$ for $k \geq 1$, it is straightforward (but somewhat messy) to verify that $C_k < C_{k+1}$ if and only if

$$\frac{1}{2}\left(1 + \frac{K}{h}\frac{2\lambda}{k(k+1)}\right) < \frac{\sum_{i=0}^{k} a_i}{(k+1)a_k}. \tag{30}$$

Finally, if $k \geq k^*$ then the left side is ≤ 1 and, as the right side is always greater than one, and the result follows.

3. EXPONENTIAL SERVICE TIMES

When the service times are exponential, say with parameter μ, it is possible to obtain a somewhat more explicit expression for $E\theta_i$, $i = k, V$. For this case $\rho = \lambda/\mu$, $\rho(t) = 1 - e^{-(1-e^{-\mu t})\rho}$ and $\rho'(t) = \lambda e^{-\mu t}$. Hence, substituting $x = \rho e^{-\mu t}$ in (15a) we have that

$$E\theta_V = EV \int_0^{\rho} \bar{R}_V(\mu x) e^x \, dx. \tag{31}$$

If in addition V is exponentially distributed with parameter ν, then simple manipulation gives

$$E\theta_V = \mu^{-1} e^{-(\nu/\mu)}[Ei((\nu/\mu) + \rho) - Ei(\nu/\mu)], \tag{32}$$

where $Ei(\cdot)$ is the *exponential integral* defined via Cauchy's principal value as:

$$Ei(t) = \int_{-\infty}^{t} \frac{e^x}{x} dx \equiv \lim_{\epsilon \downarrow 0} \int_{(-\infty, -\epsilon) \cup (\epsilon, t)} \frac{e^x}{x} dx$$

$$= \gamma + \log x + \sum_{n=1}^{\infty} \frac{x^n}{nn!}, \quad t > 0, \tag{33}$$

where γ is the Euler constant (e.g., Abramowitz and Stegun 1972).

Note that the busy periods of both the vacation model and the N-policy model correspond in this case directly to an equivalent system with a *single* exponential server who works at the state-dependent rate $\mu_n = n\mu$ when there are n customers in the system.

For $k \geq 1$ it is easy to check that a simple change of variables in (6a) gives

$$E\theta_k = \mu^{-1} \int_0^1 \frac{1 - x^k}{1 - x} e^{(1-x)\rho} \, dx. \tag{34a}$$

This can be rewritten in terms of the Erlang loss function, $B(n, x) = (\sum_{i=0}^{n} x^i/i!)^{-1} x^n/n!$ as

$$E\theta_k = \lambda^{-1}\left(e^{\rho} - 1 + \sum_{i=1}^{k-1}\left[\frac{i! e^{\rho}}{\rho^i} - \frac{1}{B(i, \rho)}\right]\right). \tag{34b}$$

In this special case, we can solve for the distribution of the steady-state number of customers in the system, for both the vacation model and the N-policy model. We begin by considering an $M/M/\infty$ queue, starting with k customers in the system, such that every time the system is about to become empty, it is instantly restarted with k customers. This gives a Markov chain with the state space being the positive integers (zero is excluded) and transition matrix $Q = q_{ij}$ where

$$q_{ij} = \begin{cases} \lambda & \text{if } j = i + 1, \\ i\mu & \text{if } i \geq 2 \text{ and } j = i - 1, \\ \mu & \text{if } i = 1 \text{ and } j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

This system clearly has a limiting distribution with state probabilities satisfying the equations:

$$\begin{aligned} \mu\pi_1^k + \lambda\pi_{i-1}^k &= i\mu\pi_i^k \quad 1 \leq i \leq k, \\ \lambda\pi_{i-1}^k &= i\mu\pi_i^k \quad k < i, \end{aligned} \tag{36}$$

where $\pi_0^k = 0$. Letting $\Pi^k(z) = \sum_{i=1}^{\infty} z^i \pi_i^k$ be the corresponding generating function we have that

$$\lambda\Pi^k(z) + \mu\Pi^k(z)\frac{1 - z^k}{1 - z} = \mu \frac{d}{dz}\Pi^k(z), \tag{37}$$

with initial condition $\Pi^k(1) = 1$, hence the unique solution is

$$\Pi^k(z) = \frac{H_k(z)}{H_k(1)}, \tag{38}$$

where

$$H_k(z) = \int_0^z \frac{1 - x^k}{1 - x} e^{(z-x)\rho} \, dx = \sum_{i=0}^{k-1} \int_0^z x^i e^{(z-x)\rho} \, dx. \tag{39}$$

Note that we can factor out $e^{-(1-z)\rho}$ from (38), which is the steady-state generating function of the number of customers in the system for a standard $M/M/\infty$ model. However, it is not hard to show that the remaining factor is *not* a generating function, so that a desired decomposition result, which holds under very general conditions for the $M/G/1$ queue (e.g., Fuhrmann and Cooper 1985, Shanthikumar 1988, and Kella and Whitt 1991) does *not* hold in this case.

To proceed, let $Q_k(t)$ denote the number of customers at time t for the system just considered. Then, by regenerative theory, it also holds (with a slight abuse of notation) that

$$\Pi^k(z) = \frac{1}{E\theta_k} E \int_0^{\theta_k} z^{Q_k(t)} \, dt. \tag{40}$$

Therefore, with (34) we have that

$$E \int_0^{\theta_k} z^{Q_k(t)} \, dt = H_k(z)/\mu. \tag{41}$$

This allows us to find the steady-state distributions of the number of customers in the model with N-policy and the one with interruptions. Denote by $P_i(\cdot)$, where $i = k$, V, the generating functions of these distributions.

Theorem 3. *When the service times are exponentially distributed with parameter* μ

$$P_k(z) = \frac{\dfrac{1-z^k}{1-z} + \rho H_k(z)}{k + \rho H_k(1)}, \tag{42}$$

and

$$P_V(z) = \frac{\tilde{R}_V(\lambda(1-z)) + \rho H_V(z)}{1 + \rho H_V(1)}, \tag{43}$$

where

$$H_V(z) = \int_0^z \tilde{R}_V(\lambda(1-x))e^{(z-x)\rho}\,dx. \tag{44}$$

Proof. To obtain (42) we first compute the expected value of the integral of $z^{L_k(t)}$ from time zero, until the first time that there are k customers present. It is easy to check that this quantity is given by

$$\lambda^{-1}\sum_{i=1}^{k} z^{i-1} = \lambda^{-1}\frac{1-z^k}{1-z}. \tag{45}$$

Adding (45) to $\mu^{-1}H_k(z)$, dividing by $EB_k = \lambda^{-1}k + \mu^{-1}H_k(1)$ and multiplying the numerator and denominator by λ gives (42).

To get (43) we have that

$$E\int_V^{B_V} z^{L_V(t)}\,dt = E\sum_{k=0}^{\infty} e^{\lambda V}\frac{(\lambda V)^k}{k!}\mu^{-1}H_k(z) = \mu^{-1}H_V(z). \tag{46}$$

To this we add (with, as before, $N(\cdot)$ being our Poisson process)

$$E\int_0^V z^{N(t)}\,dt = E\int_0^V e^{\lambda t(1-z)}\,dt = EV\tilde{R}_V(\lambda(1-z)), \tag{47}$$

and divide by EB_V, which gives (43). This completes the derivation.

4. DETERMINISTIC SERVICE TIMES

Assume now that service times are deterministic, and of length D. Note that in this case $\rho = \lambda D$ and $\rho(t) = \lambda \min(t, D)$. In this case as well, we will be able to find the steady-state distribution of the number of customers in the system, both for the N-policy and the model with vacation. The following is the main result of this section. $P_i(z)$, $i = k, V$ are defined in exactly the same way as in Theorem 3.

Theorem 4. *When the service times are deterministic, and of length* D:

$$P_k(z) = \left[\pi_k\frac{1-z^k}{k(1-z)} + (1-\pi_k)\right]e^{-(1-z)\rho} \tag{48}$$

and

$$P_V(z) = [\pi_V\tilde{R}_V(\lambda(1-z)) + (1-\pi_V)]e^{-(1-z)\rho}, \tag{49}$$

where π_i, $i = k, V$ *are given in* (23) *where, in this case, they can be written as*

$$\pi_k = \frac{k}{k + e^{\rho} - 1} \qquad \pi_V = \frac{\lambda EV}{\lambda EV + \tilde{V}(\lambda)(e^{\rho}-1)}. \tag{50}$$

Remark. Note that in this case we do get a full distributional decomposition of the same flavor as is obtained for the $M/G/1$ queue. This is in comparison to Theorem 2, where this was the case only for the expected value. Observe that $(1-z^k)/[k(1-z)]$ is the generating function of the uniform mass distribution on $0, \ldots, k-1$, $\tilde{R}_V(\lambda(1-z))$ is the generating function of the number of customers that arrive during a random length of time which has the stationary forward recurrence time distribution of a vacation, and $e^{-(1-z)\rho}$ is the generating function of a Poisson random variable with parameter ρ, which is the corresponding steady-state distribution of the number of customers in a standard $M/D/\infty$ system.

Proof. Let us begin with the model with N-policy. As in the proof of Theorem 3,

$$P_k(z) = \frac{E\displaystyle\int_0^{B_k} z^{L_k(t)}\,dt}{EB_k}. \tag{51}$$

Note that since the service times are deterministic it immediately follows that $E\theta_k = E\theta_1 = (e^{\rho}-1)/\lambda$ for all $k \geq 1$, so that $EB_k = (k + e^{\rho} - 1)/\lambda$. To compute the expected integral in (51) first note that the expected integral from time zero until the first time that there are k customers present does not depend on the service time and is given by (45). For the remaining part we let $Q_k(t)$ be as before (see (40)), the number of customers in an $M/D/\infty$ system, with k customers present at time zero all starting (and, in this case, also ending) service at the same time. Then it should be clear that the remaining part of the integral in (51) is given by $E\int_0^{\theta_k} z^{Q_k(t)}\,dt$. Since the service times are deterministic it should be clear that

$$\begin{aligned} E\int_0^{\theta_k} z^{Q_k(t)}\,dt &= E\int_0^{D} z^{Q_1(t)+n-1}\,dt \\ &\quad + E\int_D^{\theta_k} z^{Q_1(t)}\,dt \\ &= E\int_0^{\theta_1} z^{Q_1(t)}\,dt \\ &\quad - (z-z^n)E\int_0^{D} z^{Q_1(t)-1}\,dt. \end{aligned} \tag{52}$$

Now since the LST of the limiting distribution of the number of customers in an ordinary $M/G/\infty$ queue is given by

$$e^{-(1-z)\rho} = \frac{\lambda^{-1} + E\int_0^{\theta_k} z^{Q_1(t)}\,dt}{\lambda^{-1} + E\theta_1} \tag{53}$$

we have that

$$E\int_0^{\theta_1} z^{Q_1(t)}\,dt = \frac{e^{\rho z} - 1}{\lambda}. \tag{54}$$

On the interval $(0, D)$, $Q_1(\cdot) - 1$ is a Poisson process with rate λ, hence

$$E\int_0^D z^{Q_1(t)-1}\,dt = \int_0^D e^{-\lambda(1-z)}\,dt = \frac{1 - e^{-(1-z)\rho}}{\lambda(1-z)}. \tag{55}$$

Putting it all together, straightforward manipulations give (48). To obtain (49) we can repeat the same argument as in (46). Since there is no new insight here we omit this derivation.

For the deterministic case we can give a more precise statement than is given by Theorem 2 for the related minimization problem considered in Section 2. In particular, note that with a_k as in (29), we have that

$$a_k = \begin{cases} \lambda^{-1}e^{\rho} & \text{for } k = 0, \\ \lambda^{-1} & \text{for } k \geq 1. \end{cases} \tag{56}$$

hence, from (30), $C_k < C_{k+1}$ if and only if

$$\frac{1}{2}\left(1 + \frac{K}{h}\frac{2\lambda}{k(k+1)}\right) < \frac{k + e^{\rho}}{k+1}. \tag{57}$$

Hence, we immediately obtain the following corollary.

Corollary 1. *C_k (see (26)) is minimized with*

$$k = \left\lceil \left[\left(e^{\rho} - \frac{1}{2}\right)^2 + 2\lambda\,\frac{K}{h}\right]^{1/2} - \left(e^{\rho} - \frac{1}{2}\right)\right\rceil. \tag{58}$$

5. AN ALTERNATING PRIORITY INFINITE SERVER QUEUE

One application that our previous results have direct bearing upon is a particular example of a *polling* system with two classes of customers that utilize a common service center, which is comprised of an unlimited (infinite) number of servers. Customers of class i arrive independently to the system in accordance to a Poisson process with rate λ_i, $i = 1, 2$. Class 1 customers require *deterministic* service, each of duration D (as in the previous section), while class 2 customers require *random* (i.i.d) service with c.d.f. $F(\cdot)$ (as in Section 1). The system can only process jobs of the same class at any point in time. Customers of class 1 who arrive to find jobs of class 2 in service queue up until the system is free of all class 2 jobs, at which time they proceed into service (simultaneously) and vice versa, while all customers who arrive to find their own class currently being serviced proceed directly into service (this is called the "exhaustive" service discipline). We will assume that there is no switchover time incurred by the system when it switches from serving one class to the other. (This model was studied for single server systems under the name of an *alternating priority queue* (see Avi-Itzhak, Maxwell and Miller 1965), and is the genesis of much of the literature on exhaustive polling systems (e.g., Cooper 1970, Takagi 1986).) While there are many aspects of interest that warrant study for this model, here we will find the distribution of the *system busy period*, by which we mean the time from the first instance that a customer arrives to an empty system until the next instance that a customer leaves behind him an empty system. (A model of a polling system with infinite servers who utilize a "gating" service discipline is discussed in Browne et al. (1992a), and polling systems with a finite number of exponential servers moving together are analyzed in Browne and Weiss (1992).)

In the following, let B denote a random variable with the distribution of the system busy period, let $p = \lambda_1/(\lambda_1 + \lambda_2)$, and let θ_1 denote an ordinary busy period in an $M/G/\infty$ queue with arrival rate λ_2 with $\rho(t) = \lambda_2 \int_0^t (1 - F(u))\,du$. The LST of B is our main result in this section.

Theorem 5

$$\begin{aligned}\tilde{B}(\alpha) = {} & \frac{\varphi(\alpha + \lambda_1)}{1 - \varphi(\alpha) + \varphi(\alpha + \lambda_1)} \\ & \cdot (p + (1-p)[\tilde{\theta}_1(\alpha) - \tilde{\theta}_1(\alpha + \lambda_1)]) \\ & + (1-p)\tilde{\theta}_1(\alpha + \lambda_1)\end{aligned} \tag{59}$$

where

$$\varphi(s) = \frac{1}{A_0(s)}\int_0^{\infty} e^{-st-\rho(t)}\tilde{B}_D(\alpha + \rho'(t))\,dt \tag{60}$$

and where B_D is a random variable that has the distribution of a busy period in an $M/D/\infty$ queue with arrival rate λ_j, i.e.,

$$\tilde{B}_D(\alpha) = \frac{(\lambda_1 + \alpha)e^{-(\lambda_1+\alpha)D}}{\alpha + \lambda_1 e^{-(\lambda_1+\alpha)D}}. \tag{61}$$

Remark. Note that λ_2 is implicit in $\varphi(\cdot)$, where it enters through $\rho(t)$, and $\rho'(t)$.

Proof. To begin, let $\mathbf{Q}(t) = (Q_1(t), Q_2(t))$ where $Q_i(t)$ is the number of customers of type i in the system at time t. Let $B_{1,0}$ denote a system busy period that is initiated from the state (1, 0), i.e., by an arrival of a class 1 customer. This occurs with probability p. Note that since class 1 customers have deterministic service, obviously we have $B_{j,0} \stackrel{d}{=} B_{1,0}$ for all $j \geq 1$. Next, let $B_{0,k}$ denote a system busy period that starts from the state $(0, k)$. From total probability, we get the *distributional* equalities:

$$B_{1,0} \stackrel{d}{=} B_D + \sum_{k=0}^{\infty} B_{0,k} I(N_2(B_D) = k), \tag{62}$$

and

$$B_{0,k} \overset{d}{=} \theta_k + B_{1,0} I(N_1(\theta_k) > 0), \tag{63}$$

where $I(A)$ is the indicator of A, $N_i(\cdot)$ is Poisson at rate λ_i, and θ_k is a random variable that has the same distribution as a busy period that starts with k customers in an $M/G/\infty$ queue with arrival rate λ_2 and service d.f. F.

From (63), we get directly that

$$\tilde{B}_{0,k}(\alpha) = \tilde{\theta}_k(\alpha + \lambda_1) + \tilde{B}_{1,0}(\alpha)[\tilde{\theta}_k(\alpha) - \tilde{\theta}_k(\alpha + \lambda_1)], \tag{64}$$

while (62) gives us

$$\tilde{B}_{1,0}(\alpha) = \tilde{B}_D(\alpha + \lambda_2) + E\left(e^{-(\alpha+\lambda_2)B_D} \sum_{k=1}^{\infty} \tilde{B}_{0,k}(\alpha) \frac{(\lambda_2 B_D)^k}{k!}\right). \tag{65}$$

Substitute (64) into (65), take the sum using the explicit forms of (4–5), then take the expectation over B_D in the last part of (65) and simplify to find

$$\tilde{B}_{1,0}(\alpha) = \varphi(\alpha + \lambda_1) + \tilde{B}_{1,0}(\alpha)[\varphi(\alpha) - \varphi(\alpha + \lambda_1)],$$

which gives us

$$\tilde{B}_{1,0}(\alpha) = \frac{\varphi(\alpha + \lambda_1)}{1 - \varphi(\alpha) + \varphi(\alpha + \lambda_1)}. \tag{66}$$

The proof of the theorem is completed by recognizing that

$$\tilde{B}(\alpha) = p\tilde{B}_{1,0}(\alpha) + (1 - p)\tilde{B}_{0,1}(\alpha), \tag{67}$$

and then using (64) (with $k = 1$) and (66) accordingly.

The moments of the system busy period can be established accordingly, but since no new insights are to be gained, we leave the derivation for the reader. The reader should note that what allowed a direct analysis of this model is the fact that class 1 (sub) busy periods are invariant with respect to the initial number, hence there is no dependency structure between successive (sub) busy periods of the same class. This would not be the case if class 1 customers required random services as well.

Extensions of this polling model are being developed, and will be reported upon elsewhere.

REFERENCES

AVI-ITZHAK, B., W. L. MAXWELL AND L. W. MILLER. 1965. Queues With Alternating Priorities. *Opns. Res.* **13,** 306–308.

ABRAMOWITZ, M., AND I. E. STEGUN. 1972. *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, D. C.

BELL, C. E. 1971. Characterization and Computation of Optimal Policies for Operating an $M/G/\infty$ Queueing System With Removable Server. *Opns. Res.* **19,** 208–219.

BROWNE, S., AND G. WEISS. 1992. Dynamic Priority Rules When Polling With Multiple Servers. *O. R. Letts.* **12,** 129–138.

BROWNE, S., AND J. M. STEELE. 1992. Transient Behavior of Coverage Processes. *J. Appl. Prob.* **30,** Sept. 1993.

BROWNE, S., E. G. COFFMAN JR., E. N. GILBERT, AND P. E. W. WRIGHT. 1992a. Gated, Exhaustive, Parallel Service. *Prob. Engin. and Infor. Sci.* **6,** 217–239.

BROWNE, S., E. G. COFFMAN JR., E. N. GILBERT, AND P. E. W. WRIGHT. 1992b. The Gated Infinite Server Queue: Uniform Service Times. *SIAM J. Appl. Math.* **52,** 1751–1762.

COOPER, R. B. 1970. Queues Served in Cyclic Order: Waiting Times. *Bell Sys. Tech. J.* **49,** 399.

DESMIT, J. H. A. 1973. Some Results for the Many Server Queue. *Adv. Appl. Prob.* **5,** 153–169.

DOSHI, B. 1986. Queueing Systems With Vacations—A Survey. *Queue. Syst.* **1,** 29–66.

FUHRMANN, S., AND B. COOPER. 1985. Stochastic Decompositions in an $M/G/1$ Queue With Generalized Vacations. *Opns. Res.* **33,** 1117–1129.

HARRIS, C. M., AND W. G. MARCHAL. 1988. State Dependence in $M/G/1$ Server—Vacation Models. *Opns. Res.* **36,** 560–565.

HEYMAN, D. P. 1968. Optimal Operating Policies for $M/G/1$ Queueing Systems. *Opns. Res.* **16,** 362–382.

HEYMAN, D. P., AND M. J. SOBEL. 1982. *Stochastic Models in Operations Research, Vol. 1.* McGraw-Hill, New York.

KELLA, O. 1989. The Threshold Policy in the $M/G/1$ Queue With Server Vacations. *Naval Res. Logist.* **36,** 111–123.

KELLA, O., AND W. WHITT. 1991. Queues With Server Vacations and Lévy Processes With Secondary Jump Input. *Anns. Appl. Prob.* **1,** 104–117.

LEVY, Y., AND U. YECHIALI. 1976. $M/M/S$ Queues With Server Vacations. *INFO* **14,** 153.

PRABHU, N. U. 1980. *Stochastic Storage Processes.* Springer-Verlag, New York.

SHANBAG, D. N. 1966. On Infinite Server Queues With Batch Arrivals. *J. Appl. Prob.* **3,** 274–279.

SHANTHIKUMAR, J. G. 1986. On Stochastic Decomposition in $M/G/1$ Type Queues With Generalized Server Vacations. *Opns. Res.* **36,** 566–569.

SOBEL, M. J. 1969. Optimal Average Cost Policy for a Queue With Startup and Shutdown Costs. *Opns. Res.* **17,** 145–162.

STADJE, W. 1985. The Busy Period of the Queueing System $M/G/\infty$. *J. Appl. Prob.* **22,** 697–704.

TAKÁCS, L. 1956. On a Probability Problem Arising in the Theory of Counters. *Proc. Cambridge Phil. Soc.* **52,** 488–498.

TAKAGI, H. 1986. *Analysis of Polling Systems.* MIT Press, Cambridge, Mass.

TIJMS, H. C. 1986. *Stochastic Modelling and Analysis.* John Wiley, New York.

WEINS, D. P. 1989. On the Busy Period Distribution for the $M/G/2$ System. *J. Appl. Prob.* **26,** 858–865.

YADIN, M., AND P. NAOR. 1963. Queueing Systems With a Removable Server. *Opns. Res. Quart.* **14,** 393–405.