International Journal of Operations Research

International Journal of Operations Research Vol. 3, No. 1, 47−55 (2006)

Adjusting the Workload of an Under-utilized Server by Scheduling Supplementary Work

Zhe George Zhang[1,*], Naishuo Tian[2], and Ernie Love[3]

[1]Department of Decision Sciences, College of Business and Economics, Western Washington University, Bellingham, WA 98225-9077, U.S.A.

[2] Department of Mathematics and Physics, Yanshan University, Qinhuangdao, 066004, China.

[3]Faculty of Business Administration, Simon Fraser University, Burnaby, B.C. Canada.

Received November 2005; *Revised January 2006*; *Accepted March 2006*

Abstract——This paper addresses an important question of how to achieve an appropriate and nearly uniform work-load for an under-utilized staff in a waiting line situation by scheduling an appropriate amount of supplementary work during the idle time. We present an M/G/1 queue with a special server's vacation policy to model this situation. In this system, after serving all arriving customers, the server can perform a random maximum number H of possible supplementary jobs before staying idle. The distribution of H can be determined by a supplementary job assignment policy to reduce the idle time proportion. Major performance measures are obtained to evaluate this class of policies. Numerical examples are presented to illustrate the application of this study.

Keywords——Scheduling, M/G/1 queue, Server vacations, Idle-time utilization, Stochastic decomposition property

1. INTRODUCTION

This study is motivated by a real situation as follows: In a department store or a supermarket, a customer-service -desk (CSD) worker's primary job is to serve after-sales customers with special needs. However, sometimes when the CSD worker is less busy, the store manager may ask him or her to do some optional jobs. For example, when the CSD worker becomes idle, he or she may check if other jobs such as restocking empty shelves, cleaning floors, or helping customers are available. Of course, there are also other employees in charge of these jobs, so it is not compulsory for the CSD worker to do them. In other words, the CSD worker's top priority or "must-do" jobs are serving CSD customers and his or her secondary or optional jobs are these non-CSD jobs. The store manager is interested in knowing the appropriate amount of optional idle-time duty scheduled to the CSD worker. In many practical waiting line systems, managers face the same question of how to achieve a uniform work-load by appropriately scheduling some supplementary jobs for an under-utilized server.

In this paper, we develop a special queueing model with a flexible server's vacation policy to study the issue of work-load adjustment by scheduling supplementary jobs. In this model, a queue is formed by randomly arriving customers representing the primary jobs and the idle server can take *vacations* representing the durations of performing supplementary jobs. We present a set of formulas to quantify the performance of a supplementary work scheduling policy. In fact, the first study on vacation models was motivated by the question of effectively utilizing server's idle time (see Levy and Yachiali, 1975).

Over the past two decades, queueing systems with vacations (or simply called vacation models) have been studied by many researchers due to their wide applications in manufacturing and telecommunication systems. Several excellent and comprehensive surveys on the recent results and references for a variety of vacation models can be found in Doshi (1986), Doshi (1990), Fuhrmann and Cooper (1985), Shanthikumar (1988), Takagi (1991) and Takagi (1993). The vacation policy in this paper is more general and flexible than most classical vacation policies to model the situation of our interest and is called a *multiple adaptive vacation policy* (MAV-policy) which was first studied by Tian (1992). The discrete-time system with MAV policy was treated by Zhang and Tian (2001). However, in Tian (1992), the issues of modeling the supplementary work and controlling the server's utilization level were not addressed. In this paper, we focus on these issues and investigate the impact of assigning supplementary jobs on the server's utilization level.

This paper is organized as follows. An MAV vacation model has been formulated and the major performance measures are developed in section 2. In section 3, scheduling the supplementary work-load is modeled as a

[*] Corresponding author's email: george.zhang@wwu.edu

random maximum number of vacations that the server can take before a primary job arrives. Several practical situations are discussed. In section 4, numerical examples are presented to discuss the idle time utilization via scheduling secondary jobs in these "not-so-busy" server situations. Finally, the paper concludes with a summary in section 5.

2. THE MAV MODEL

In this section, a simple vacation model is presented to represent the system of our interest. Consider an M/G/1 system with arrival rate λ and an MAV policy in which the server will take a random maximum number, denoted by H, of vacations after emptying the system. The probability mass function (p.m.f.) of H is $P(H = j) = c_j, j = 1, 2, \ldots,$ and its p.g.f is $H(z) = \sum_{j=1}^{\infty} c_j z^j$. The random variable H may represent the maximum number of tasks or jobs available for the server to work on during his or her idle time. The vacations are general independent and identically distributed (i.i.d.) random variables, denoted by V. At each vacation completion instant, the server checks the system state to decide an action to take. There are three cases at this instant. Case 1: if there are some waiting customers, the server will resume serving the queue immediately; Case 2: if there is no waiting customer and the total number of vacations taken is still smaller than H, the server will take another vacation; Case 3: if there is no waiting customer and the number of vacations taken is equal to H, the server will stay idle and wait for the next arrival. Note that this MAV policy is appropriate for modeling two main characteristics of a situation in which the server can perform some supplementary work during his or her idle time. These are (1) any arriving customer to an empty system will not wait for more than a residual vacation time; and (2) the proportion of the server's idle time is controllable by adjusting the parameters of the distribution of H. We assume that the random variables - time between arrivals, T, service times, S, maximum number of vacations, H, and vacation duration, V, are mutually independent and $\rho = \lambda E(S) < 1$.

Let J represent the actual number of vacations taken by the server in a cycle, then

$$J = \min\left\{H, k: \ V^{(k-1)} < \mathrm{T} < V^{(k)}\right\},$$

where $V(k)$ stands for the sum of k vacations, with $V^{(0)} \equiv 0$. To reflect the fact that the primary jobs receive higher service priority, the server continues serving the queue until the queue is empty (also called an exhaustive service type).

Let A_I and A_V denote the event that the first customer arrival to an empty system occurs in a server's idle state and in a server's vacation state, respectively. We have

$$P(A_I) = \sum_{v=0}^{\infty} P(H=v)P(T \ge V^{(v)}) = \sum_{v=0}^{\infty} c_v \int_0^{\infty} e^{-\lambda t} dV^{(v)}(t)$$
$$= H(\tilde{V}(\lambda)),$$
$$P(A_V) = 1 - H(\tilde{V}(\lambda)),$$

where $\tilde{V}(s)$ is the Laplace Transform (LST) of V.

Let L_n be the number of customers left behind by n th customer departure. Then the queue-length process $\{L_n, n \ge 1\}$ is an embedded Markov chain and

$$L_{n+1} = \begin{cases} L_n - 1 + A & if \ \ L_n \ge 1 \\ Q_b - 1 + A & if \ \ L_n = 0 \end{cases} \tag{1}$$

where A denotes the number of customers arriving during the service of a customer and Q_b the number of customers present when the busy period begins. Because A and Q_b are i.i.d random variables, we omit the subscript n. In addition, we introduce the following probabilities -

$$a_j = \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} dS(t),$$
$$v_j = \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} dV(t), \ j \ge 0,$$

where a_j (or v_j) is the probability that j customers arrive during a service time (or a vacation time).

We first determine the probability distribution of Q_b which is needed in developing the performance measures of a vacation model. Event $\{Q_b = 1\}$occurs in either of the two mutually exclusive cases: (a) the first arrival to an empty system occurs in an idle period because the maximum number of vacations have been finished; and (b) the first arrival to an empty system occurs during a server vacation period and only one arrival happens until the completion of this vacation. Therefore,

$$P\{Q_b = 1\} = H(\tilde{V}(\lambda)) + \frac{1 - H(\tilde{V}(\lambda))}{1 - \tilde{V}(\lambda)} v_1$$

Similarly, Event $\{Q_b = j\}(j \ \ge \ 2)$ represents

$$P\{Q_b = j\} = \frac{1 - H(\tilde{V}(\lambda))}{1 - \tilde{V}(\lambda)} v_j, \ j \ge 2$$

For this type of vacation model, the elements in the first row of the transition probability matrix of the M/G/1 type embedded Markov chain

$$P = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ & a_0 & a_1 & a_3 & \cdots \\ & & a_0 & a_1 & \cdots \\ & & & \vdots & \vdots \end{bmatrix}$$

should be

$$b_j = H(\tilde{V}(\lambda))a_j + \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}\sum_{i=1}^{j+1} v_i a_{j+1-i}, \quad j \geq 0. \tag{2}$$

It is easy to find the expected number of customers present when the server resumes serving the queue as

$$\Theta = E(Q_b) = H(\tilde{V}(\lambda)) + \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}\lambda E(V),$$

and it is used in the following development of the major performance measures. We present the stochastic decomposition property on the stationary performance measures. This property shows the net effects of using the server's idle time for supplementary work.

The stationary queue length and the waiting time for an M/G/1 queue with the MAV policy, denoted by L_v, and W_v, respectively, can be decomposed into the sum of two independent random variables as

$$\begin{aligned} L_v &= L + L_d \\ W_v &= W + W_d \end{aligned} \tag{3}$$

where L and W are the queue length and the waiting time (with a FIFO discipline), respectively, of a classical M/G/1 queue and L_d and W_d are the additional queue length and the additional delay, respectively, due to the MAV policy with the following LST and the z-transform as

$$\begin{aligned} L_d = \{&1-H(\tilde{V}(\lambda))z - \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)} \\ &\times\left[\tilde{V}(\lambda(1-z)) - \tilde{V}(\lambda)\right]\}/\Theta(1-z), \end{aligned} \tag{4}$$

$$\begin{aligned} \tilde{W}_d(s) = \{&\lambda-(\lambda-s)H(\tilde{V}(\lambda)) - \lambda\frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)} \\ &\times(\tilde{V}(s)-\tilde{V}(\lambda))\}/\Theta_s. \end{aligned} \tag{5}$$

Proof. See APPENDIX.

Let D_v be the busy period of the M/G/1 queue with MAV policy and note that the only difference between D_v and the busy period of a classical M/G/1 queue, denoted by D is that D_v starts with Q_b customers. Thus $\tilde{D}_v(s) = Q_b(\tilde{D}(s))$. Using the z-transform of Q_b which is derived in APPENDIX, we obtain the LST and the mean of D_v as

$$\begin{aligned} \tilde{D}_v(s) &= H(\tilde{V}(\lambda))\tilde{D}(s) + \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)} \\ &\quad\times\left\{\tilde{V}(\lambda(1-\tilde{D}(s))) - \tilde{V}(\lambda)\right\}, \\ E(D_v) &= \frac{\rho}{1-\rho}\left(\frac{\Theta}{\lambda}\right). \end{aligned}$$

The z-transform of the number of vacations taken at a time, J, can be obtained as follows:

$$J(z) = 1 - \frac{1-z}{1-\tilde{V}(\lambda)z}(1-H(\tilde{V}(\lambda)z)). \tag{6}$$

The derivation of (6) can be found in Tian (1992).

Based on (6), the total length of vacations taken at a time, denoted by V_G, has the LST and the mean as

$$\begin{aligned} \tilde{V}_G(s) &= J(\tilde{V}(s)) = 1 - \frac{1-\tilde{V}(s)}{1-\tilde{V}(\lambda)\tilde{V}(s)}(1-H\left[\tilde{V}(\lambda)\tilde{V}(s)\right]), \\ E(V_G) &= \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}E(V). \end{aligned}$$

Other useful performance measures include the expected server idle time, $E(I_v)$, the expected cycle time, $E(B_c)$, and probabilities of the server being busy, on vacation, or idle, denoted by P_B, P_V, P_I, respectively, and they are obtained easily as follows.

$$\begin{aligned} E(I_v) &= \frac{1}{\lambda}H(\tilde{V}(\lambda)), \\ E(B_c) &= E(D_v) + E(V_G) + E(I_v) = \frac{\Theta}{\lambda(1-\rho)}, \\ P_B &= \frac{E(D_v)}{E(B_c)} = \rho, \\ P_V &= \frac{E(V_G)}{E(B_c)} = \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}\lambda E(V)\frac{(1-\rho)}{\Theta}, \\ P_I &= \frac{E(I_v)}{E(B_c)} = H(\tilde{V}(\lambda))\frac{(1-\rho)}{\Theta}. \end{aligned}$$

Note that the special cases of $H = 1$ and ∞ correspond to a single vacation model and a multiple vacation model, respectively, (see Fuhrmann and Cooper (1985) for the details about these models).

3. MODELING THE LEVEL OF SCHEDULING SUPPLEMENTARY WORK

In this study, we consider two types of supplementary work with random availability. Type 1 is the sum of Geometric random variables and of type 2 is the sum of Bernoulli random variables.

These forms of H are flexible to model the level of supplementary work for different situations. As an example of type 1 supplementary work, consider an employee whose primary task is to answer "inbound phone" calls from customers. If the volume of such calls is low, then the employee could be assigned the secondary task of making "outbound calls" to potential customers to attempt to sell them a product or service. Usually, the employee will need to make several calls to reach a successful sale. Suppose that each call has an exponentially distributed duration and a success probability p of making a sale. Then the total time until a sale is made will be a sum of geometrically distributed number of exponential random variables. Assume that the employee follows a policy where after each call (regardless of the cutcome - success or failure), he or she checks the queue of holding inbound calls, if there are no waiting calls, continue making outbound calls to potential customers until he or she has made n_g sales - the maximum number of sales made when no calls are waiting. If no inbound calls arrive until n_g sales are made, the employee will become idle and wait for the next inbound call. It is also worth noting that multiserver call centers with both inbound and outbound calls have been studied in the past (see Deslauriers et al., 2005; Gans et al., 2003; Koole and Mandelbaum, 2002).

For type 2 supplementary work, consider a customer service desk (CSD) of a supermarket. In such a situation, the idle CSD employee may check n_b possible supplementary jobs (e.g. checking n_b shelves) and each job has a probability p to be available (empty shelf found). He will do these jobs (restock the empty shelf) as long as no customer arrives at the service desk. In this case the maximum number of supplementary jobs is a binomial random variable.

In fact, the advantage of this model is that the supplementary work level and the server's work-load can be controlled by the decision variables n_b and n_g as described above. These two cases are presented below for the numerical illustrations

Case 1: H = the sum of i.i.d. Geometric random variables.

Let G_i be the i.i.d. Geometric random variables. In this case, we have

$$H = G_1 + G_2 + \ \dots \ + G_{n_g} .$$

$$G_1(z) = G_2(z) = \ \dots \ = G_{n_g}(z) = \frac{pz}{1-(1-p)z},$$

where p is the parameter of the geometric distribution. This type of H can be used to model the situation that the server must search with certain success probability for the secondary jobs to do (like the example described above).

That is, G_i represents the number of attemps made to complete the ith job successfully. In other words, completing ith job requires G_i activities and the duration of each activity is modeled as a vacation in the MAV model with an exponential distribution of rate θ. Immediately, we get

$$H(z) = \left(\frac{pz}{1-(1-p)z}\right)^{n_g}, \ E(H) = \frac{n_g}{p},$$

$$H(\tilde{V}(\lambda)) \ = \ \frac{1}{\left(\dfrac{\lambda}{p\theta}+1\right)^{n_g}} .$$

Case 2: H = the sum of i.i.d. Bernoulli random variables.

In this case, H is a Binomial random variable and is more appropriate for the customer service desk (CSD) employee in a supermarket. Each available job is treated as an exponentially distributed vacation with rate of θ. Thus, we have

$$H(z) = \{pz + (1-p)\}^{n_b}, \ E(H) = n_b p,$$

$$H(\tilde{V}(\lambda)) \ = \ \left\{1 \text{-} \left(\frac{\lambda}{\theta+\lambda}\right)p\right\}^{n_b} .$$

Note that this case is also appropriate for modeling the machine maintenance problem in which H represents the number of machines inspected during the idle time and the customer arrivals represent the failure machines requesting repairs.

Based on the exponential vacations (supplementary jobs) with the mean of $1/\theta$ and the LST of $\tilde{V}(s) = \dfrac{\theta}{\theta+s}$, we can also obtain the major performance measures below:

$$\Theta = 1 + \frac{\lambda}{\theta}(1 - H(\tilde{V}(\lambda)))$$

$$P_I = (1-\rho)\frac{H(\tilde{V}(\lambda))}{1 + \dfrac{\lambda}{\theta}(1-(H(\tilde{V}(\lambda)))},$$

$$\mathrm{E}(\mathrm{W}_d) = \frac{1}{\theta}(1+\frac{\lambda}{\theta})\frac{1-H(\tilde{V}(\lambda))}{1 + \dfrac{\lambda}{\theta}(1-(H(\tilde{V}(\lambda)))},$$

$$\mathrm{E}(V_G) = (1+\frac{\lambda}{\theta})(1-(H(\tilde{V}(\lambda)))\frac{1}{\theta}.$$

With these formulas and the expression for $H(\tilde{V}(\lambda))$, we can also obtain the variations of the performance measures with respect to the decision variable n_g or n_b. For example, we can obtain the rate of change of P_I with respect to n_g for case 1 as

$$\frac{dP_I}{dn_g} = -\frac{\ln(\dfrac{\lambda}{p\theta}+1)}{(\dfrac{\lambda}{p\theta}+1)^{n_g}} < 0 .$$

Note that other vacation distributions such as deterministic or phase type can be treated in the MAV model.

4. CONTROL OF IDLE TIME UTILIZATION-COMPUTATIONAL RESULTS

Using these two special cases, we numerically demonstrate that the server's average utilization level can be effectively controlled by choosing an appropriate parameter n_b or n_g for a given environment. In the geometric H case, n_g is the required number of successful supplementary jobs completed during the idle time before the server can stay idle and in the binomial H case, n_b is the number of potential jobs with random availability to be inspected during the idle time. Some numerical examples are presented below to illustrate the impact of changing n_g or n_b on the server's idle time proportion. The system parameters of the four cases are in Table 1.

The main reason for choosing this system parameter dataset is because these cases are all low traffic load of ρ=1/4. Therefore, scheduling some supplementary work help improve the server's utilization.

Some Observations from the Numerical Examples:

(1) In the Geometric type H cases, the idle time proportion reduction decreases more rapidly with n when p is small or "more-difficult-to-be-successful" type supplementary jobs are scheduled. For example, in Case 1 shown in Figure 1, requesting the idle server to complete one successful search before becoming idle (n = 1) will reduce the idle time proportion from 75% to 31% for p = 0.1 case compared to 75% to 65% for p = 0.9 case. It is not surprising that smaller p values result in the idle time reduction being more sensitive to increase in n, because as p decreases each geometric random variable in the sum is becoming stochastically larger. This also means that the server checks to see if a primary job has arrived less frequently, and so primary jobs may have to wait longer. In contrast, for the Binomial type H cases, the idle time proportion drops more significantly as n increases for bigger p values or "easier-to-occur" type secondary jobs. For example, in Case 3 of Figure 3, an n = 5 policy cuts the idle time proportion from 75% to 24% for p = 0.9 compared to 75% to 66% for p = 0.1.

(2) Comparing Figure 1 to Figure 2 or Figure 3 to Figure 4 indicates that scheduling larger supplementary jobs (or smaller θ) to idle servers is more effective in reducing the idle time proportion than scheduling smaller supplementary jobs (or larger θ).

Table 1. The system parameters of the four cases

	λ (arrival rate)	μ (service rate)	θ (vacation rate)	H type
Case1	1	4	8	Geometric
Case2	1	4	4	Geometric
Case3	1	4	8	Binomial
Case4	1	4	4	Binomial

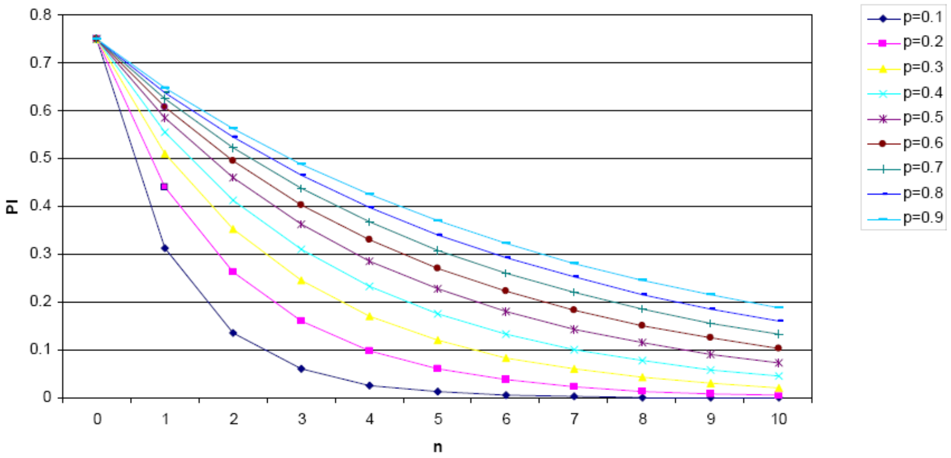

Figure 1. The proportion of idle time v.s. policy parameter n_g − θ = 4 case.

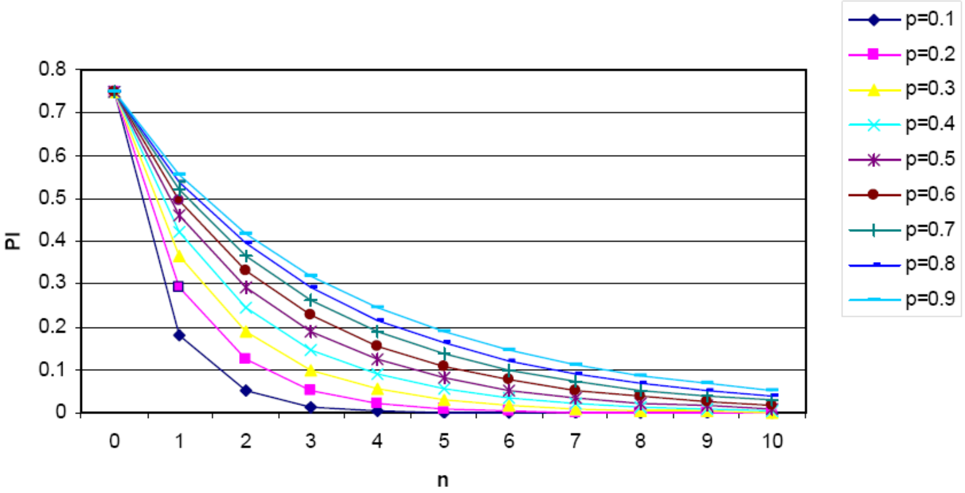

Figure 2. The proportion of idle time v.s. policy parameter n_g − θ = 4 case.

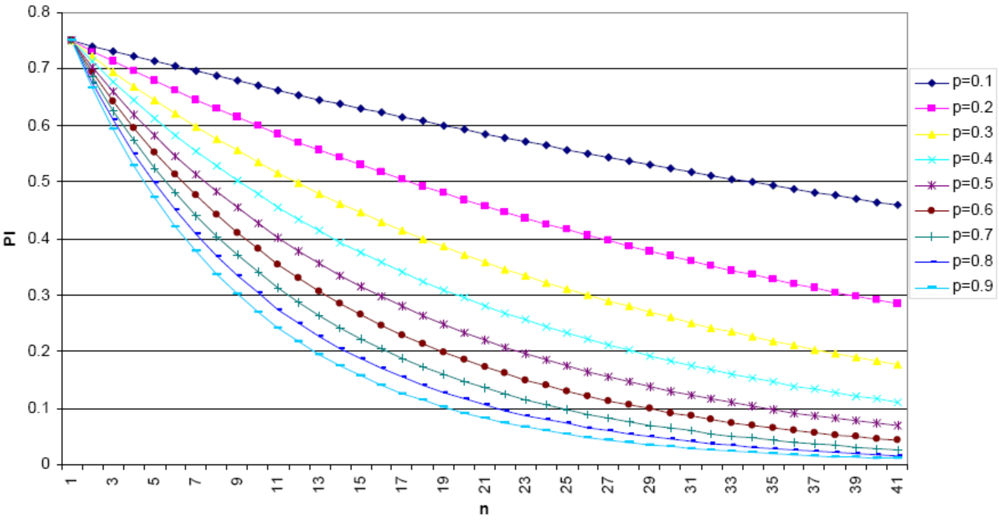

Figure 3. The proportion of idle time v.s. policy parameter n_b – θ = 8 case.

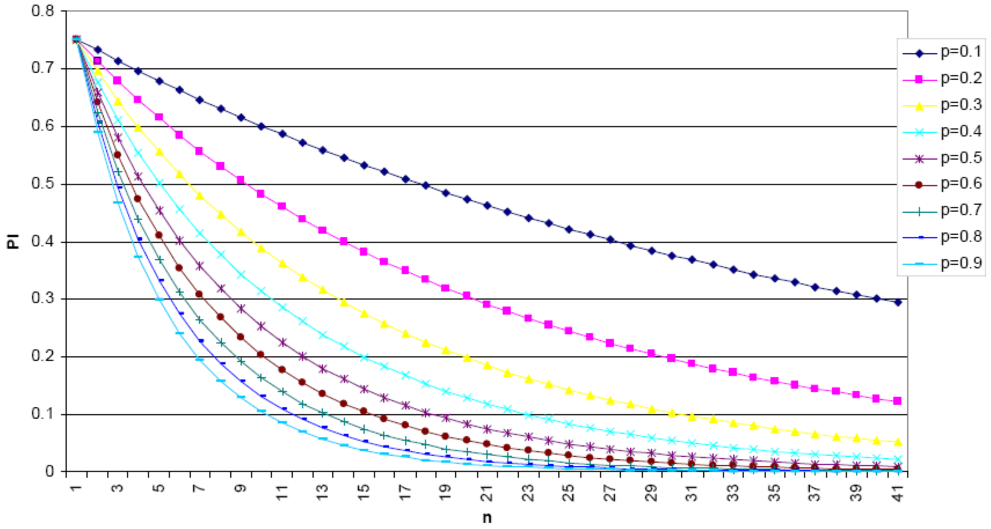

Figure 4. The proportion of idle time v.s. policy parameter n_b – θ = 4 case.

(3) There are other important performance measures that might interest practitioners implementing MAV type policies. One of them is the additional expected waiting time of arriving customers due to performing the supplementary jobs, $E(W_d)$, and the other is the expected number of vacations taken (or supplementary jobs done) each time- $E(N_V)$. Both of them are increasing in the policy parameter n as shown in Figures 5 and 6 for Case 1. Because randomly arriving customers are top priority jobs which trigger an immediate server's returning at the completion of the next supplementary job and the queue service is exhaustive, both $E(W_d)$ and $E(N_V)$ are upper bounded by the finite limits. The upper limit for $E(W_d)$ is the well-known additional expected delay due to a classical multiple vacation policy (that is an $n = \infty$ case of this MAV model) – $V^{(2)}/2E(V)$ and the upper limit for $E(N_V)$ is $1/(1 - \tilde{V}(\lambda))$. In the exponential supplementary job case with θ as the processing supplementary job rate, these two limits are $1/\theta$ and $1+\theta/\lambda$, respectively. These values are useful in assessing the performance effects of utilizing server's idle time.

(4) Using Case 3 with p = 0.6, we investigate the issue of achieving a target idle time proportion, P_I, at different traffic load $\rho = \lambda/\mu$. For example, in Figure 7, to achieve a target P_I = 15%, we adjust n value from 0 to 22 when the arrival rate λ changes from 3.5 to 1.0. Specifically, n = 0, 2, 5, 8, 13, and 22 for λ = 3.5, 3.0, 2.5, 2.0, 1.5, and 1.0, respectively, to approximately achieve this target idle time proportion which might be the organization-wide standard or average level for employees.

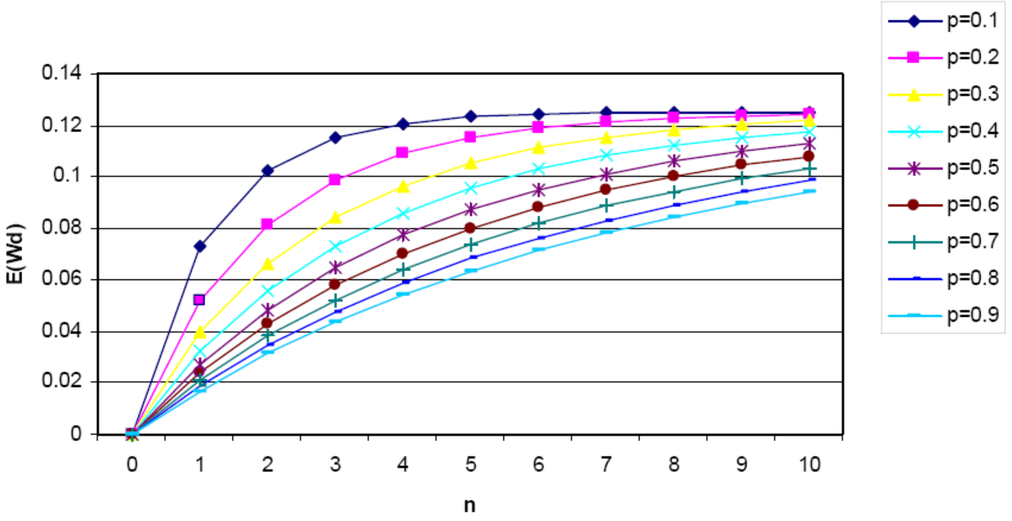

Figure 5. The expected waiting time due to performing the secondary jobs v.s. policy parameter n_g for θ= 8.

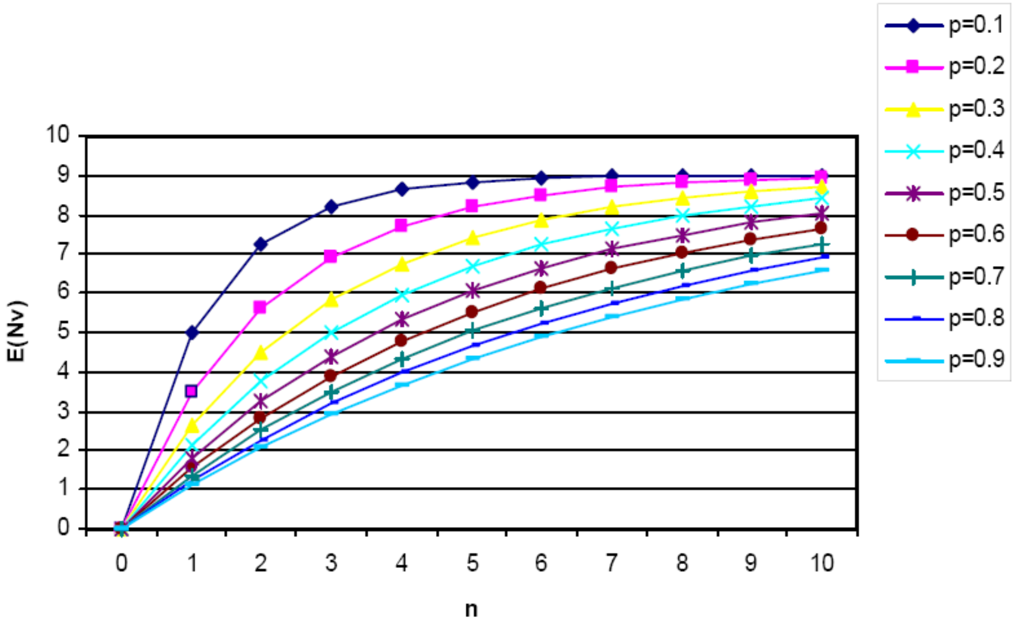

Figure 6. The expected waiting time due to performing the secondary jobs v.s. policy parameter n_g for θ = 8.

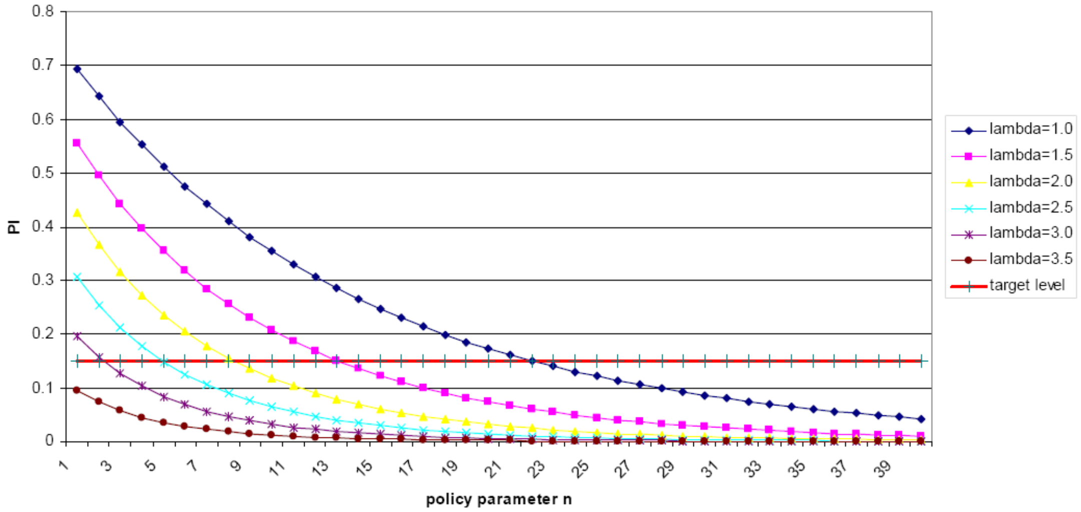

Figure 7. The proportion of idle time v.s. policy parameter n_b for different arrival rates in Case 3.

5. CONCLUDING REMARKS

This paper presents a useful quantitative model for practitioners to adjust the server's work-load level and other performance measures of a queueing system by scheduling an appropriate amount of supplementary work during the server's idle time. Because all random variables are generally distributed except for the Poisson arrivals, this model provides a general analytical framework for answering the important and practical question of how to improve the server's time utilization in a queueing situation with a low traffcintensity. With this model, queueing managers are able to develop an appropriate "idle-time-work" assignment policy for effectively utilizing server's idle time and maintaining a fast service response to primary customers. For example, a desired constant average work-load or time utilization level can be achieved in the situation where $\rho = \lambda/\mu$ varies significantly. The uniform work-load is measured in terms of either the proportion of idle time or the proportion of busy time and is usually the average employee work-load level for the whole organization. In classical multiple vacation models and single vacation models (see Shanthikumar, 1988) which are two special cases of our model, the server's idle time proportion is fixed and is not controllable. In our model, the server's idle time proportion is completely controllable.

Although in this paper we only present two types of *H*'s, other types of *H* distributions fitting different real-life situations can be studied similarly. In this model, different values of *p'* or *p* may represent different success probabilities in the Geometric *H* case or different random availability in the Binomial *H* case for different types of supplementary jobs. Hence, this model can help managers schedule different amounts of supplementary work for different types to the idle employee to achieve an appropriate utilization level.

In this study, we assume that the arriving customers have higher service priority than the supplementary jobs during the idle time for the server. Note that this type of queue service resumption is a special case of the threshold policy (or the N-policy). A direction of future study is to combine the MAV policy with a general N-threshold policy for resuming the queue service. For some past work on the threshold policy vacation models, see Levy and Yachiali (1975), Takagi (1993), and Zhang et al. (1997). Another future research topic is to extend this MAV policy model to a multi-server queueing system by using different methods such as matrix geometric solution approach.

REFERENCES

1. Deslauriers, A., Pichitlamken, J., Ingolfsson, A., and Avramidis (2005). Markov chain models of a telephone call center with call blending. *Computers and Operations Research*, forthcoming.
2. Doshi, B.T. (1986). Queueing systems with vacations - a survey. *Queueing Systems*, 1: 29-66.
3. Doshi, B.T. (1990). Single server queues with vacations, in: H. Takagi (Ed.) *Stochastic Analysis of Computer and Communication Systems.* North-Holland, Amsterdam 217-265.
4. Fuhrmann, S.W. and Cooper, R.B. (1985). Stochastic decomposition in the M/G/1 queue with generalized vacations. *Operations Research*, 33: 1117-1129.
5. Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: tuto-rial, review, and research prospects. *Manufacturing and Service Operations Management*, 5: 79-141.
6. Kella, O. (1989). The threshold policy in M/G/1 queue with vacations. *Naval Research Logistics*, 36: 111-123.
7. Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: an introduction. *Annals of Operations Research*, 113: 41-59.
8. Levy, Y. and Yachiali, U. (1975). Utilization of idle time in an M/G/1 queueing system. *Management Science*, 22: 202-211.
9. Shanthikumar, J.G. (1988). On stochastic decomposition in an M/G/1 type queue with generalized server vacation. *Operations Research*, 36: 566-569.
10. Takagi, H. (1991). *Queueing Analysis - A Foundation of Performance Evaluation Vol. 1 Vacation and Priority Systems.* North-Holland, Amsterdam.
11. Takagi, H. (1993). *Queueing Analysis - A Foundation of Performance Evaluation Vol. 3 Discrete-Time Systems.* North-Holland, Amsterdam.
12. Tian, N. (1992). Multi-stage adaptive vacation policies in an M/G/1 queueing system. *Appl. Math.*, 4: 12-18.
13. Zhang, Z.G. and Love, C.E. (1998). The threshold policy in M/G/1 queue with an exceptional first vacation, *INFOR* 36(4): 193-204.
14. Zhang, Z.G., Vickson, R.G., and van Eenige, M. (1997). Optimal two threshold policies in an *M*/*G*/1 queue with two vacation types. *Performance Evaluation* 4(2): 131-149.
15. Zhang, Z.G. and Tian, N. (2001). Discrete time Geo/*G*/1 queue with multiple adaptive vacations. *Queueing Systems*, 38: 419-429.

APPENDIX: A BRIEF PROOF OF THE THEOREM

Based on the stochastic decomposition theorem for the M/G/1 queue with general vacations (see Doshi, 1986 ; Doshi, 1990; Fuhrmann and Stochastic, 1985; Levy and Yachiali, 1975; Shanthikumar et al., 1988), we have

$$L_v(z) = L(z)L_d(z)$$
$$=\frac{(1\text{-}\rho)(1\text{-z})\tilde{\text{S}}(\lambda(1\text{-z}))}{\tilde{S}(\lambda(1-z))-z}\frac{1-Q_b(z)}{E(Q_b)(1-z)}$$

Furthermore, we find the z-transform of $Q_b(z)$ and $E(Q_b)$

$$Q_b(z) = H(\tilde{V}(\lambda))z + \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}\sum_{j=1}^{\infty} z^j v_j$$
$$=H(\tilde{V}(\lambda))z + \frac{1-H(\tilde{V}(\lambda))}{1-\tilde{V}(\lambda)}\left[\tilde{V}(\lambda(1-z))-\tilde{V}(\lambda)\right],$$

$$E(Q_b) = \Theta = H(\tilde{V}(\lambda)) + \frac{1 - H(V(\lambda))}{1 - \tilde{V}(\lambda)} \lambda E(V).$$

Substituting these expressions into the expression of $L_v(z)$ gives (4). Similarly, for the waiting time, we have

$$\tilde{W}_v(s) = \tilde{W}(s)\tilde{W}_d(s) = \frac{(1 - \rho)s}{s - \lambda(1 - \tilde{S}(s))} \frac{\lambda\left[1 - Q_b(1 - \frac{s}{\lambda})\right]}{E(Q_b)s}$$

Substituting $Q_b(z)$ and $E(Q_b)$ into $\tilde{W}_v(s)$ gives (5). It is easy to prove that L_v is also the stationary queue length distribution at any time due to PASTA.