J. Appl. Math. & Informatics Vol. **28***(2010), No. 1 - 2, pp. 531 - 537*
Website: http://www.kcam.biz

SOME WAITING TIME AND BOTTLENECK ANALYSIS

JONG SEUL LIM

ABSTRACT. In this paper, some vacation policies are considered, which can be related to the past behavior of the system. The server, after serving all customers, stays idle or to wait for some time before a vacation is taken. General formulas for the waiting time and the amount of work in the system are derived for a vacation policy. Using the analysis on the vacation system, we derived the waiting time in the sequential bottleneck station.

AMS Mathematics Subject Classification : 68M20
Key words and phrases : Waiting time, queue length, bottleneck, unfinished work

1. Introduction

Models of single server queues with vacations have been widely used to study the performance of many computer, communication, and production systems. In this paper, some vacation policies are considered, which can be related to the past behavior of the system. The server, after serving all customers, stays idle or waits for some time before a vacation is taken. Literature [2,3,7] have dealt with the unfinished work, waiting time, and queue length for many vacation models. The analysis of the delay due to vacations are particularly difficult for models with complicated policies. The basic structure of a vacation policy is investigated and some general results for the waiting time are obtained to facilitate easy analysis.

2. Analysis for the delay

We derive the equations for the delay and begin with stating some existing results. For an arbitrary customer C_i, let $C_{f(i)}$ be the first customer of the busy period to which C_i belongs. Let U_i denote the amount of work in the system found by C_i upon arrival.

Received June 5, 2009. Accepted October 8, 2009.

This decomposition property is shown in Lemma 2 of Doshi[1] for the GI/G/1 queue. Let D denote the amount of work found in system by the first customer of an arbitrary busy period. Let the Laplace Transform for D and U are $D^*(s)$ and $U^*(s)$, respectively. From the well-known waiting-time decomposition in Proposition 4 in Fuhrmann and Cooper[3], we obtain

$$U^*(s) = D^*(s)\frac{(1-\rho)s}{s-\lambda+\lambda X^*(s)}.$$

Now let $Y_v(t)$ and $Y_c(t)$ be the amount of vacation work and customer work in the system at time t. Let S_k and E_k be the time at which the k^{th} busy period starts and ends, assuming $S_1 = 0$ and using S_k^- to be the time instant immediately prior to S_k. To find $D^*(s)$, since the server does not stay idle as long as work exits in the system, we assume without loss of generality that customer work has a preemptive resume priority over vacation work. Under this preemptive resume priority assumption, $Y_c(S_k^-) = 0$ for all $k = 1, 2, \cdots, \infty$. Hence, the amount of work D must be either vacation work or zero. Under the priority assumption, the amount of customer work in system B becomes zero at the end of a busy period. Thus, $Y_c(E_k) = 0$ for all $k = 1, 2, \cdots, \infty$.

By the definition of an idle period (i.e., time at which the server in system A is idle), we have $Y_c(t) = 0$ for $t \in (E_k, S_{k+1})$ for any k. Therefore, $Y_c(S_k^-)$=0, for all $k = 1, 2, \cdots, \infty$. Since there are only two types of work in system B and since $Y_c(S^-)$ =0, D must be either vacation work or zero. $Y_v(t)$ is a constant for $t \in (S_k, E_k]$ for all $k = 1, 2, \cdots, \infty$. Based on the Markovian property, the constant behavior of $Y_v(t)$ during all busy periods is irrelevant to the characterization of D. To find $D^*(s)$, one can construct a new process, $Z_v(\hat{t})$, of a virtual time $\hat{t}$ by contracting the time intervals, $(S_k, E_k]$ for $k = 1, 2, \cdots, \infty$, of $Y_v(t)$ into a single point in time. Formally, let

$$a(t) = \max\left\{k : t \geq S_k\right\},$$
$$b(t) = \max\left\{k : t \geq E_k\right\},$$

then

$$Z_v(\hat{t}) = Y_v(t) \tag{1}$$

where

$$\hat{t} = \begin{cases} t - \sum_{k=1}^{b(t)}(E_k - S_K) & if\ a(t) - b(t) = 0 \\ S_{a(t)} - \sum_{k=1}^{b(t)}(D_k - S_k) & if\ a(t) - b(t) = 1. \end{cases}$$

We notice that this is the contraction operation discussed in Section 4.2 of Levy and Kleinrock[7]. Each busy period is contracted into a single point. Let these points be referred to as sampling points. D is distributed as $Z_v(\hat{t})$ at an

arbitrary sampling point. Before presenting the formulas for $D^*(s)$, referring that results of the excess residual life distribution in Wolff[8], we let V_j for $j = 1, 2, \cdots, \infty$ be the sequence of vacation lengths. For every $t \geq 0$, let $I_j(t) = 1$ if $V_j > t$, $I_j(t) = 0$ otherwise, where $\int_0^\infty I_j(t)dt = V_j$. The vacation lengths have

$$V(t) = 1 - \lim_{n\to\infty} \sum_{j=1}^{n} I_j(t)/n.$$

Let $V^*(s)$ denote the Laplace transform of the derivative of $V(t)$. The average vacation length is given by

$$\overline{v} = \lim_{n\to\infty} \sum_{j=1}^{n} V_j/n$$

where $0 < \overline{v} < \infty$. It is also assumed that $\overline{v} = \int_0^\infty \Big[1 - V(t)\Big]dt$.

If P_0 is the probability that first customer of an arbitrary busy period finds the system empty at the arrival instant, then

$$D^*(s) = P_0 + (1 - P_0)\left[\frac{1 - V^*(s)}{s\overline{v}}\right].$$

Because of the memoryless property of Poisson arrivals, the length of any idle period, $[\, E_k, S_{k+1}\,)$ in time t, is exponentially distributed with rate λ. Therefore, the time interval in virtual time $\hat{t}$ between two consecutive sampling points has an exponential distribution with the parameter λ. In other words, the sampling points arrive according to a Poisson process in virtual time $\hat{t}$. Now let P_i for i $= 1, 2, \cdots, N$ be the probability that the first customer of an arbitrary busy period finds the system with non-zero type-i vacation work. Further, use P_0 to denote the probability that a busy period starts when the system is empty. Then, we have $\sum_{i=0}^{N} P_i = 1$ and

$$D^*(s) = P_0 + \sum_{i=1}^{N} P_i \frac{1 - V_i^*(s)}{s\overline{v}_i}\ . \tag{2}$$

The vacation-work process, $Z_v(\hat{t})$, can be obtained from (1). Further, D is statistically equal to $Z_v(\hat{t})$ sampled by a Poisson process. Thus, if a sampling point finds $Z_v(\hat{t}) = 0$, then $D^*(s) = 1$. Since each sampling point represents a busy period, P_i defined above is also the probability that the sampling point finds $Z_v(\hat{t})$ having type-t vacation work.

Given that an arbitrary sampling point finds that $Z_v(\hat{t})$ has non-zero type-i vacation work, by the same excess distribution results used, $Z_v(\hat{t})$ found by the

sampling point upon arrival is distributed as the residual life of a type-i vacation, which is characterized by the Laplace Transform, $\left[1-V_i^*(s)\right]/s\overline{v}_i$. Unconditioning this with P_i for all i $= 1, 2, \cdots, N$ and combining with the case of $Z_v(\hat{t}) = 0$, we obtain (2). Since each sampling point finds $Z_v(\hat{t})$ either zero or representing non-zero type-i vacation work for i $= 1, 2, \cdots, N$, we have $\sum_{i=0}^{N} P_i = 1$. After serving all customers, the server decides whether to start a single vacation or remain idle in the system. The vacation decision and the subsequent vacation length may depend on the past system history, but are mutually independent to each other. Once the server chooses to stay idle, the server remains idle until a busy period has started and ended.

$Z_v(\hat{t})$ for this vacation policy is constructed. Let β_S be the long-term fraction of decision epochs at which the server starts a single vacation. $1-\beta_S$ is the long-term fraction of decision epochs for the server to stay idle. Applying the same definition and arguments for a cycle, the average cycle length is

$$\overline{c} = \beta_S \left[\overline{v} + V^*(\lambda)\frac{1}{\lambda}\right] + (1-\beta_S)\frac{1}{\lambda}. \tag{3}$$

We recognize that the sum of $\beta_S\left[V^*(\lambda)/\lambda\right]$ and $(1-\beta_S)/\lambda$ in (3) represent the average duration of virtual time $\hat{t}$ where the server is idle. Thus, P_0 is given by

$$P_0 = \frac{\beta_S V^*(\lambda)\frac{1}{\lambda} + (1-\beta_S)\frac{1}{\lambda}}{\overline{c}} = \frac{1+\beta_S\left[V^*(\lambda)-1\right]}{\beta_S\left[\lambda\overline{v}+V^*(\lambda)\right]+1-\beta_S}.$$

And the Laplace Transform for the waiting time which is the amount of work found in the system by an arbitrary customer upon arrival is

$$U^*(s) = \frac{1-\rho}{s-\lambda+\lambda X^*(s)}\left\{\frac{\left[1+\beta_S(V^*(\lambda)-1)\right]s+\lambda\beta_S\left[1-V^*(s)\right]}{1+\beta_S[\lambda\overline{v}+V^*(\lambda)-1]}\right\}.$$

3. Bottleneck Of the system

Now we consider the bottleneck in the system. We assume that $\rho_1 < \rho_2 < \cdots < \rho_K$. Using the equations in Harrison[4], we can assume that each station has a distinct traffic intensity. This assumption will be relaxed below. With $\rho_1 < \rho_2 < \cdots < \rho_K$, station K is the unique bottleneck. Thus we can have waiting time at the K_{th} station as follows.

$$\hat{W}_K = \tau_K + \frac{\rho_K\tau_K}{2(1-\rho_K)}\left\{\left(\gamma_K\left[1-\hat{p}_K\right]\right)^{-1}\sum_{i=1}^{K}\lambda_i q_{iK}(q_{iK}c_{a,i}^2+1-q_{iK}) + c_{s,K}^2(1-\hat{p}_K)+\hat{p}_K\right\}.$$

We also assume that all stations with indices larger than k are supersaturated, while all stations with indices smaller than k are instantaneous switches. Some of this analysis is shown in [5]. A supersaturated station has two main characteristics, which follow from it having an infinite queue length. First, customers which are routed there never return, and second, arrivals from these form a renewal process since the server is always busy. To look at the network from the point of view of station k, we let ${}_kP = \{p_{ji},\ 1 \le j, l \le k\}$ be the $k \times k$ submatrix of P obtained by removing all elements with an index larger than k. Since a customer which is routed to a station with index larger than k is considered to have left the network, P is the new routing matrix. Let ${}_kR = (I - {}_kP)^{-1}$, and let

$${}_kq_{ji} = \begin{cases} {}_kR_{jl}/\ {}_kR_{il}, & 1 \le j \le k \\ & 1 \le l \le k. \\ \displaystyle\sum_{i=1}^{K} {}_kq_{il}p_{ji}, & k < j \le K \end{cases}$$

Thus ${}_kq_{jl}$ is the probability that a customer, starting from station j reaches station l before leaving the network, when all stations with indices larger than k are supersaturated. When $1 \le j \le K$, the definition of ${}_kq_{jl}$ follows as before. When $k < j \le K$, since there are no rows or columns of ${}_kR$ corresponding to station j, we calculate ${}_kq_{jl}$ by considering paths which enter the non-supersaturated network and then get to station l. Finally, let ${}_k\hat{p}_j = 1 - {}_kR_{jj}^{-1}$, $1 \le j \le k$, so that ${}_k\hat{p}_j$ is the probability that a customer at station j returns to station j before leaving the network when all stations with indices larger than k are supersaturated. Combining all of the above, we have, for $1 \le k \le K$

$$\hat{W}_k = \tau_k + \frac{\rho_k\tau_k}{2(1-\rho_k)}\left[\gamma_k(1 - {}_k\hat{p}_k)\right]^{-1}\left\{\sum_{j=1}^{k}\lambda_j\left(1 + {}_kq_{jk}[c_{a,j}^2 - 1]\right){}_kq_{jk}\right.$$

$$\left. + \sum_{j=k+1}^{K}\gamma_j\left(1 + {}_kq_{jk}[c_{s,j}^2 - 1]\right){}_kq_{jk} + \gamma_k\left(c_{s,k}^2[1 - {}_k\hat{p}_k]^2 + {}_k\hat{p}_k[1 - {}_k\hat{p}_k]\right)\right\}, \tag{4}$$

where an empty sum is taken to be zero. From (4), we note that for $1 \le j \le k$, which correspond to the under-saturated and bottleneck station, they are the same. For $k < j \le K$, two changes occur. First, the arrival rate, λ_j, is replaced by the throughput rate, γ_j. Second, $c_{a,j}^2$ is replaced by $c_{s,j}^2$. We have $c_{a,i}^2 = c_{s,i}^2$, $1 \le k \le K$, so that

$$\hat{W}_k = \tau_k + \frac{\rho_k\tau_k}{2(1-\rho_k)}[\gamma_k(1 - {}_k\hat{p}_k]^{-1}\left\{\sum_{j=1}^{k}\lambda_j\ {}_kq_{jk} + \sum_{j=k+1}^{K}\gamma_j\ {}_kq_{jk} + \gamma_k(1 - {}_k\hat{p}_k)\right\}$$

$$= \tau_k + \frac{\rho_k\tau_k}{2(1-\rho_k)}\left\{1 + \gamma_k^{-1}\left[\sum_{j=1}^{k}\left(\lambda_j + \sum_{l=k+1}^{K}\gamma_l p_{lj}\right)\ {}_kR_{jk}\right]\right\}. \tag{5}$$

When station $k+1, \cdots, K$ are supersaturated, the adjusted external arrival rates are

$$\hat{\lambda}_i = \lambda_i + \sum_{l=k+1}^{K} \gamma_l p_{li}, \quad 1 \le i \le k. \tag{6}$$

Consider the reduced traffic equation

$$\hat{\gamma}_i = \hat{\lambda}_i + \sum_{l=1}^{k} \hat{\gamma}_l p_{li}, \quad 1 \le i \le k, \tag{7}$$

which has the unique solution

$$\hat{\gamma}_i = \sum_{j=1}^{k} \hat{\lambda}_j \, {}_kR_{ji}. \tag{8}$$

Substituting (6) into (7), we obtain

$$\hat{\gamma}_i = \lambda_i + \sum_{l=1}^{k} \hat{\gamma}_l p_{li} + \sum_{l=k+1}^{K} \gamma_l p_{li}. \tag{9}$$

Combining (6), (8), and (9), we find

$$\gamma_k = \sum_{j=1}^{k} \left(\lambda_j + \sum_{l=k+1}^{K} \gamma_l p_{lj} \right) {}_kR_{jk},$$

We now remove the restriction that each station must have a different traffic intensity, and assume only that $\rho_1 \le \rho_2 \le \cdots \le \rho_K$, which is without loss of generality. Although we develop an approximation for the case $\rho_{K-1} = \rho_K$, we should point out that it is not guaranteed to be asymptotically exact in heavy traffic. The idea behind this extension is simple, so in order to avoid obscuring it we consider an example where $K = 3$, and $\rho_1 = \rho_2 < \rho_3$. Suppose we consider two related, the networks where service rates are changed so that $\rho_1 = \rho_2 + \epsilon$ in one, and $\rho_1 = \rho_2 - \epsilon$ in the other. Now suppose that there are L station in a balanced subnetwork. Although there are $L!$ orderings of station in the subnetwork, not all of them give different approximations for a given station. In fact, the only thing that matters to a station is which other stations have larger traffic intensities. There are 2^{L-1} different ways to choose which of the other $L-1$ stations have larger traffic intensities, and this is the number of different combinations which must be properly weighted by the number of permutations that it represents. With l stations having larger traffic intensities, $0 \le l \le L-1$, the correct weight is $l!(L-l-1)!$. Although we have defined sequential bottleneck for all possible traffic intensity vectors, $\rho = (\rho_1, \cdots, \rho_k)$ it should be noted that $\hat{W}_k$ in (5) is not necessarily a continuous function of ρ.

4. Conclusions

A general formula for the waiting time in the system has been obtained in this paper. We analyzed and the amount of work with a vacation policy, which

would otherwise be difficult to analyze. The analysis approach may well be applicable to other related queueing models like variants of priority queues, if they conform with the basic model considered in this paper.These results can serve as a basis for the formulation and solution of certain optimization problems involved with vacation models. Using the analysis on the vacation system, we derived the waiting time in the sequential bottleneck station. Assuming that no two stations have identical traffic intensities, all stations with larger traffic intensities are treated as if they are supersaturated, which turns them into sinks for customers routed to them, and sources for customers routed from them. Stations with smaller traffic intensities are again treated as instantaneous switches. When several stations have the same traffic intensity, an average was taken over all possible orderings which could be induced by a perturbation of the traffic intensities. The results shows that both of these decomposition approximations share some appealing features. When applied to a network with a sequential bottleneck station they turned out to be asymptotically exact in heavy traffic.

REFERENCES

1. Doshi, B.T., *A Note on Stochastic Decomposition in a GI/G/1 Queue with Vacations or Set-Up Times,* J.Appl.Prob., 22, 419-428, 1985a.
2. Doshi, B.T., *An M/G/1 Queue with Variable Vacation,* Proc. Int. Conf. on Performance Modeling, Sophia Antipolis, France, 1985b.
3. Fuhrmann, S.W. and Cooper, R.B., *Stochastic Decompositions in the M/G/1 Queue with Generalized Vacations,* Opns.Res., 33, 1117-1129, 1985.
4. Harrison, J. M., *Brownian models of queueing network with heterogeneous customer populations. Stochastic Differential Systems, Stochastic Control Theory and Applications,* W. Fleming and P. L. Lions Eds., Springer-Verlag, 147-186, 1998.
5. Johnson, D. P., *Diffusion approximations for optimal filtering of jump processes and for queueing network,* Ph. D. Thesis, University of Wisconsin, 1983.
6. Kleinrock, L., *Queueing Systems Vol.II: Computer Application,* John Wiley & Sons, Inc., New York, 1976.
7. Levy,H. and Kleinrock, L., *A Queue with Starter and A Queue with Vacations: Delay Analysis by Decomposition,* Opns.Res., 34, 426-436, 1986.
8. Wolff ,R.W., *Sample-Path Derivations of the Excess, Age, and Spread Distributions,* J.Appl.Prob. 25, 432-436, 1988.

Jong Seul Lim received the BS degree from Seoul National University and the Ph.D degree in communications and operations engineering from Polytechnic University, New York, in 1988. He worked with AT&T Bell Laboratories and developed the computer network and cellular communication systems. After then, he joined SK Telecom Corp. Since 1993, he has worked with Sunmoon University, Korea as a professor.
e-mail: jslsky7@hotmail.com