Analysis of M/M/S Queues with 1-Limited Service

Anshuman Tyagi
Nortel Networks
PO Box 3511, Station C
Ottawa, ON K1Y 4H7
Canada

Janelle Harms
Department. of Computing Science
University of Alberta
Edmonton, AB T6G 2H1
Canada

Ahmed Kamal
Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011
U.S.A.

Abstract

In this paper, a multiple server queue, in which each server takes a vacation after serving one customer is studied. The arrival process is Poisson, service times are exponentially distributed and the duration of a vacation follows a phase distribution of order 2. Servers returning from vacation immediately take another vacation if no customers are waiting. A Matrix Geometric Method is used to find the steady state joint probability of number of customers in the system and busy servers, and the mean and the second moment of number of customers and mean waiting time for this model. This queueing model can be used for the analysis of different kinds of communication networks, such as multislotted networks, multiple token rings, multiple server polling systems and mobile communication systems.

I Introduction

Queueing systems that allow servers to be on vacation arise in many computer and communication systems. Server vacations may be due to lack of work, server failure or another task being assigned to the server. In these systems the server is not always available to serve its primary customers. Some of the applications which can be modeled using these systems are computer maintenance and testing[1], CPU scheduling, TDMA networks[2], and priority queues[3]. Another important application of vacation systems are polling systems or cyclic queues[4,5]. Cyclic queues arise in multiprocessor systems[6], multiple token ring and multislotted networks[5,7,8]. Single server queues with vacations have been used to study single server cyclic queues[4]; similarly multiple server queues with vacation can be used to study multiple server cyclic queues.

There are several variations of vacation models that can be defined. There may be a single server or multiple servers. The server may take a vacation at a random time (non-exhaustive), after serving at most k customers (k-limited) or after all the customers in the queue are served (exhaustive). Also, depending on the application, when the server finishes a vacation and there is no customer to be served, the server may take another vacation (multiple vacation model V_M) or it may wait, ready to serve, until a customer arrives (single vacation model V_S).

In this paper a 1-limited M/M/S/V_M vacation queueing model is studied. The customer interarrival and service times are assumed to be exponential; while the vacation times obey a phase distribution of order 2. This gives us more flexibility in adapting the queueing model to analyze different applications. Applications of this model are discussed in detail in Section 6, they include cyclic queue applications such as multiple token ring networks and mobile communication systems. The transition rate matrix of this model is infinite in size, and is of the quasi-birth-death process type [9]. The steady state probability vector can be solved by using one of the following five methods:

1. truncation of the state space,
2. the matrix geometric method for matrices of the G/M/1 type[9],
3. methods for the structured stochastic matrices of the M/G/1 form[10].
4. methods based on the generating functions[11], or
5. methods based on obtaining the eigenvalue and eigenvectors of difference equations involving the steady state probabilities[12].

In this paper we use the matrix geometric method to obtain the joint probability of number of customers in the system and the number of servers, mean number of customers in the system and mean waiting time for any value of S, the number of servers. It should be noted that the 2-phase vacation distribution has been chosen for the sake of mathematical tractability. Yet, with two phases, it is still possible to match the first three moments of the phase type distribution to measured moments, which should result in a reasonably accurate system model.

In the next section we survey the M/M/S type Vacation queues which have already been studied in the literature. The basic queueing techniques which have been used by the different researchers to analyze these queues are presented. In Section 3, we briefly explain the phase distribution and matrix geometric solution. In Section 4, the model presented in this paper is developed. Numerical results are presented and analyzed in Section 5. In Section 6 we describe two applications of our model and in Section 7 we present a brief conclusion.

We should note that as an easy reference, and for the benefit of the reader, we placed a list of all the symbols used in Appendix B.

II A Survey of Analytical Models for Queues with Vacations

Single Server Vacation models have been studied for different arrival, service and vacation characteristics. Some of the techniques that have been used to analyze this model are an embedded Markov chain approach[4,13,14,2], a decomposition method[15,16], and a level crossing argument[17]. In this section, analyses of M/M/S queues in which the servers take exponentially distributed vacations are discussed.

An M/M/S/V_M queue with exhaustive service has been analyzed by Levy and Yechiali[18] and Kao and Kumar[19]. In both of these papers, the service follows an exponential distribution, arrivals are Poisson distributed and vacation times are exponential. Levy and Yechiali, use a balance

equation method based on generating functions. Kao and Kumar[19] use a matrix-geometric approach for modeling the system. They derive the stationary, joint probability distribution of queue length and the number of busy servers, the distributions of waiting time and the length of busy period. The balance equation method is also used to study a single vacation model, M/M/S/V_S with exhaustive service[18].

There are few other M/M/S vacation models that have been studied. A steady state M/M/S/V_M queueing system where each server is subject to random breakdown (non-exhaustive vacation model) of exponentially distributed duration has been studied by Mitrani and Avi-Itzhak[20] and Neuts and Lucantoni[21]. Mitrany and Avi-Itzhak have used the balance equation method similar to the one used by Levy and Yechiali[18] to obtain the generating function of the queue size. For $S \leq 2$, they derive the explicit form but for large S a numerical method is suggested. Neuts and Lucantoni used the Matrix Geometric Method to solve this model and they obtained an algorithm to solve the waiting time distribution and steady state probability.

In[22] the balance equation method is used to analyze a similar M/M/S/V_M model with 1-Limited service (the model studied in this paper). However it is only possible to derive the distribution of the number of customers in the system and the mean number of customers in the system for $S \leq 3$. For higher values of S the method fails to give results due to the lack of sufficient equations. In this paper we study this model using a matrix geometric method.

III Background

In this section we provide information on the Phase type distribution which has been assumed for the servers' vacation. We also describe briefly the matrix geometric solutions.

III.1 The Phase Type Distribution

The phase distribution is a generalization of Erlang's method of stages[23] and is well-suited for numerical computation[9]. Since it is more general but still numerically solvable, it is preferred to the exponential distribution. The advantage of using this distribution is the ability to more accurately model practical and complex distributions. Moreover, a large number of distributions are special types of phase distributions. For example an n-stage Erlangian distribution can be treated as an n order phase distribution, similarly an exponential distribution can be treated as a phase type distribution of order 1, by properly choosing the values of parameters required to describe a phase distribution. Thus from the results of phase type distribution we can derive results for other distributions as well.

A phase type distribution of order m is described by an $(m+1)$ state Markov process, with infinitesimal generator Q defined as follows[9]:

$$Q = \begin{bmatrix} \mathbf{T} & T^0 \\ \mathbf{0} & 0 \end{bmatrix}$$

where the $m \times m$ matrix $\mathbf{T}$ satisfies $T_{kk} < 0$, for $k \leq m$ and $T_{kl} \geq 0$, for $k \neq l$. The elements of $\mathbf{T}$, T_{kl}, give the rate of transition from phase k to phase l. The column vector T^0 gives the rate of entering the absorption phase from the different phases. Also $\mathbf{T}\vec{e}+T^0 = \mathbf{0}$ since Q is an infinitesimal generator. The initial probability vector of Q is given by $(\nu\ ,\ \nu_{m+1})$ with $\nu\vec{e} + \nu_{m+1} = 1$. All the states $1, \ldots, m$ are transient, so that the absorption into state $m+1$ from any initial state is certain. In our model, this implies that the vacation time is finite. If the vacation time is phase distributed,

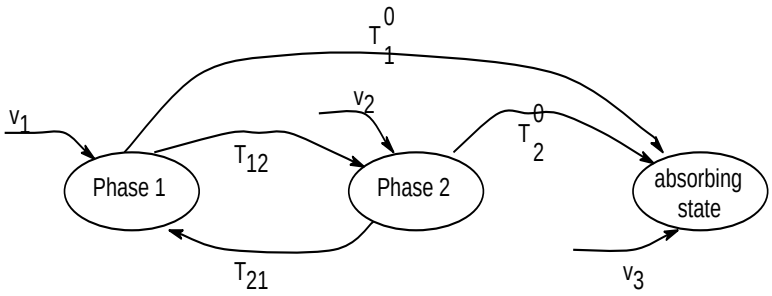

Figure 1: Phase Distribution of order 2

then its mean value is given by $\bar{v} = -\nu \mathbf{T^{-1}} \vec{e}$. A phase distribution of order 2 is shown in figure 1 and describes the vacations of the servers in our model. A server on vacation can move between phase 1 or 2. Once it enters phase 3 (the absorbing phase), it returns and, if no unserved customer is present, it will immediately take another vacation. When a server leaves (either upon service completion or if it arrives at an empty line) it will enter either phase 1 or 2 with probability ν_1 or ν_2 respectively. In our case the value of $\nu_3 = 0$, since a server must take a vacation when it leaves a queue.

III.2 The Matrix Geometric Solution Approach

A generalized quasi-birth-death transition matrix $\tilde{Q}$ is as follows:

$$\tilde{Q} = \begin{bmatrix} A_1^0 & A_2^0 & & & & & & \\ A_2^1 & A_1^1 & A_0^1 & & & & & \\ & A_2^2 & A_1^2 & A_0^2 & & & & \\ & & & \ddots & & & & \\ & & & A_2^{S-1} & A_1^{S-1} & A_0^{S-1} & & \\ & & & & A_2^S & A_1^S & A_0 & \\ & & & & & A_2 & A_1 & A_0 \\ & & & & & & & \ddots \end{bmatrix}$$

The matrix $\tilde{Q}$ is an infinitesimal generator and hence $A_2^i \vec{e} + A_1^i \vec{e} + A_0^i \vec{e} = 0$ and $A_2 \vec{e} + A_1 \vec{e} + A_0 \vec{e} = 0$ are satisfied.

The stationary probability of $\tilde{Q}$ is denoted by $\vec{\mathbf{x}} = (\vec{x_0}, \vec{x_1}, ...)$. The normal procedure to solve for the joint probability vector $\vec{\mathbf{x}}$ is to use $\vec{\mathbf{x}}\tilde{Q} = 0$ equations and the normalizing equation $\sum_{i=0}^{\infty} \vec{x_i} \vec{e_i} = 1$, where $\vec{e_i}$ is a column vector with all elements equal to 1. The solution using the above procedure seems quite impossible due to the infinite number of unknown variables. However, the stationary joint probability $\vec{\mathbf{x}}$ is "modified matrix-geometric", that is, it is of the form $\vec{x_k} = \vec{x_S} R^{k-S}$, $k \geq S$, where R is the nonnegative solution to a matrix quadratic equation. Therefore, it is only required to obtain the values of $(\vec{x_0}, \vec{x_1}, \ldots, \vec{x_S})$ by using the normal procedure and the rest can be obtained by using the "matrix-geometric" property.

IV The Model and Its Solution

IV.1 Model Description

The model can be formulated as a continuous time Markov chain (CTMC). The possible states are defined by $(i, j; k)$ where i denotes the number of customers in the system, j denotes the number of servers at the queue, and k denotes the number of servers in phase 1 of the vacation distribution. The remaining servers $S - j - k$ are in phase 2 of the vacation distribution. The variables can take on the following values: $0 \leq i \leq \infty$ and $0 \leq j, k \leq S$. Being a multiple vacation model, the $i \geq j$ condition holds.

Next, we define the submatrices of $\tilde{Q}$. The entries within the submatrices, are subsubmatrices (B_i's and E's), which are defined in detail below. We define I_j as the identity matrix of dimension $(S - j + 1)$. In this section, the matrix 0_j is a square matrix whose elements are all zeros, and has the same dimensions as I_j. We define $0_j^{'}$, as a $(S - j + 1) \times (S - j)$ matrix whose elements are all zero. The matrices I_j, 0_j and $0_j^{'}$ are used to clearly define the matrices A's, B's and E and to make their dimensions consistent.

A_2^i gives the rate at which the server leaves and hence a customer leaves after service completion when the system has i customers. The value of A_2^i, where $i = 1, \ldots S$ is as follows:

$$A_2^i = \begin{bmatrix} 0_0 & & & & \\ B_0^1 & 0_1 & & & \\ & B_0^2 & \ddots & & \\ & & \ddots & 0_{i-1} & \\ & & & B_0^i & \end{bmatrix}_{\frac{(i+1)(2S+2-i)}{2} \times \frac{i(2S+3-i)}{2}}$$

B_0^j denotes the rate of entering one of the phases of vacation after service completion by a server when there are j servers at the queue.

The definition of A_1^i is as follows. The off-diagonal matrices of A_1^i give the rate of server arrival to the queue. The diagonal matrices give the rate at which the number of customers and the servers at the queue remains the same though the number of servers in the two phases may change. The diagonal matrices are obtained using the relationship $A_2^i\vec{e} + A_1^i\vec{e} + A_0^i\vec{e} = 0$ and $A_1^0 = B_1^0 + E^0$, where E^i denotes the rate of transition when there are servers on vacation but there are no more customers to be served at the queue, thus $(i, j; k) \rightarrow (i, j; k^{'})$ for $i = j = 0, \ldots, S$. The value of A_1^i, where $i = 1, \ldots S$ is as follows:

$$A_1^i = \begin{bmatrix} B_1^0 & B_2^0 & & & \\ & B_1^1 & B_2^1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & B_2^{i-1} \\ & & & & B_1^i + E^i \end{bmatrix}_{\frac{(i+1)(2S+2-i)}{2} \times \frac{(i+1)(2S+2-i)}{2}}$$

B_1^j denotes the rate of transition of servers among the two phases and the rate at which the number of servers and customers remains the same in the system when there are j servers at the queue. B_2^j denotes the rate of vacation termination of a server when there are j servers at the queue.

The matrix A_0^i gives the rate of arrival of a customer. $A_0^0 = \begin{bmatrix} B_3^0 & 0_0^{'} \end{bmatrix}$, while A_0^i for $i =$

$1, \ldots S-1$ is as follows:

$$A_0^i = \begin{bmatrix} B_3^0 & & & & \\ & B_3^1 & & & \\ & & \ddots & & \\ & & & B_3^i & 0_i \end{bmatrix}_{\frac{(i+1)(2S+2-i)}{2} \times \frac{(i+2)(2S+1-i)}{2}}$$

When the number of customers in the system exceeds S, the matrices A_2^i, A_1^i and A_0^i correspond to A_2, A_1 and A_0 and are defined as follows:

$$\begin{aligned} A_2 &= \begin{bmatrix} 0_0 & & & & \\ B_0^1 & 0_1 & & & \\ & B_0^2 & 0_2 & & \\ & & \ddots & \ddots & \\ & & & B_0^S & 0_S \end{bmatrix} \\ A_1 &= \begin{bmatrix} B_1^0 & B_2^0 & & & \\ & B_1^1 & B_2^1 & & \\ & & \ddots & \ddots & \\ & & & B_1^{S-1} & B_2^{S-1} \\ & & & & B_1^S \end{bmatrix} \\ A_0 &= \lambda I \end{aligned}$$

The dimensions of A_2, A_1 and A_0 are $\frac{(S+1)(S+2)}{2} \times \frac{(S+1)(S+2)}{2}$.

Now we define the subsubmatrices (B^j and E^j) that make the A submatrices. The superscript of these subsubmatrices give the number of servers at the queue before transition. The positions of elements within the subsubmatrices correspond to the number of servers in phase 1 and 2 of vacation. For example, if there are j servers serving the queue then the rest of the $S-j$ servers should be in one of the two phases.

B_0^j denotes the rate of entering one of the phases of vacation after service completion by a server when there are j servers at the queue. Thus, the transition occurs from $(i,j;k) \rightarrow (i-1, j-1; k^{'})$ for $j = 1, \ldots S$.

$$B_0^j = \begin{bmatrix} j\mu\nu_2 & j\mu\nu_1 & & \\ & \ddots & \ddots & \\ & & j\mu\nu_2 & j\mu\nu_1 \end{bmatrix}_{(S-j+1)\times(S-j+2)}$$

B_1^j denotes the rate of transition of servers among the two phases and the rate at which the number of servers and customer remains the same, where $(i,j;k) \rightarrow (i,j;k^{'})$ for $j = 0 \ldots S$. Its contents are (Let $f_i = iT_{11} + (S-j-i)T_{22}$.)

$$B_1^j = \begin{bmatrix} f_0 & (S-j)T_{21} & & \\ T_{12} & f_1 & \ddots & \\ & \ddots & \ddots & T_{21} \\ & & (S-j)T_{12} & f_{S-j} \end{bmatrix} - (\lambda + j\mu)I_j$$

B_2^j denotes the rate of vacation termination of a server, thus $(i,j;k) \rightarrow (i,j+1;k')$, $j = 0,\ldots,S-1$.

$$B_2^j = \begin{bmatrix} (S-j)T_2^0 & & & & & \\ T_1^0 & (S-j-1)T_2^0 & & & & \\ & 2T_1^0 & (S-j-2)T_2^0 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & T_2^0 & \\ & & & & (S-j)T_1^0 & \end{bmatrix}$$

where T_1^0 and T_2^0 are the rates at which a server comes back to the queue after vacation from phase 1 and phase 2 respectively. This is shown in figure 1. $B_3^j = \lambda I_j$, denotes the rate of customer arrival.

In order to compute E^i, note that a server, after completing its vacation, checks the queue, finds that there are no customers to be served then goes for another vacation instantaneously. If $h_j = (S-i-j)T_2^0\nu_2 + jT_1^0\nu_1$, then

$$E^i = \begin{bmatrix} h_0 & (S-i)T_2^0\nu_1 & & & \\ T_1^0\nu_2 & h_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & T_2^0\nu_1 \\ & & & (S-i)T_1^0\nu_2 & h_{S-i} \end{bmatrix}$$

h_j gives the rate of service completion and instantaneous return to the same phase of vacation by a server.

IV.2 Stability Condition and State Probabilities

As the infinitesimal generator $\tilde{Q}$ has been defined completely, we now derive the condition for the system to reach steady state. To do this we define matrix $A = A_2 + A_1 + A_0$. A can be written as

$$A = \begin{bmatrix} B_1^0 + \lambda I_0 & B_2^0 & & & \\ B_0^1 & B_1^1 + \lambda I_1 & B_2^1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & B_2^{S_1} \\ & & & B_0^S & B_1^S + \lambda I_S \end{bmatrix}$$

The stationary probability vector $\mathbf{\Pi} = (\vec{\pi}_0, \vec{\pi}_1, \ldots \vec{\pi}_S)$ of A can be obtained using the $\mathbf{\Pi}A = 0$ relationship and the normalizing relationship, $\sum_j \vec{\pi_j}\vec{e_j} = 1$, where the vector $\vec{e_j}$ is a column vector of 1's of size $(S-j+1)$.

Using the stability condition, that is $\mathbf{\Pi}A_0\vec{e} < \mathbf{\Pi}A_2\vec{e}$ (see[9]), we can verify that for the equilibrium and also for $\tilde{Q}$ to be positive recurrent, the following equation should be satisfied.

$$\bar{v} + \frac{1}{\mu} < \frac{S}{\lambda} \tag{1}$$

The above relation means that the sum of mean vacation time and service time divided by S (equivalent to mean server availability) should be less than the mean interarrival time for the system to achieve steady state.

To study the performance of this system we require the steady state probability. Let $\vec{\mathbf{x}} = [\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_{S-1}, \vec{x}_S, \vec{x}_{S+1}, \ldots]$ be the steady state probability vector. The probability that there are i customers in the system is $\vec{x}_i = (\vec{x}_{i,0}, \vec{x}_{i,1}, \ldots, \vec{x}_{i,min(S,i)})$, where $\vec{x}_{i,j}$ is the joint probability that there are i customers in the system, and j servers at the queue and is equal to $(x_{(i,j;0)}, x_{(i,j;1)}, \ldots, x_{(i,j;S-j)})$. Finally $x_{(i,j;k)}$ denotes the probability of i customers in the system, j servers present at the queue and k servers present in phase 1 and $S-j-k$ servers present in phase 2 of the vacation distribution. The value of $\vec{x}_k$, where $k \geq S$ can be obtained by the relation $\vec{x}_k = \vec{x}_S R^{k-S}$. R is an $\frac{(S+1)(S+2)}{2} \times \frac{(S+1)(S+2)}{2}$ matrix and is the minimal non-negative solution of the quadratic equation(refer to[9])

$$R^2 A_2 + R A_1 + A_0 = 0 \tag{2}$$

The value of R can be obtained from the above quadratic equation and the following relation:

$$R A_2 \vec{e} = A_0 \vec{e} \tag{3}$$

The above equation implies that the rate of transition from a state where there are $\mathbf{i}$ customers, to a state with $\mathbf{i+1}$ matches the transition rate from $\mathbf{i}$ to $\mathbf{i-1}$.

Using the simultaneous equations obtained from $[\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_{S-1}, \vec{x}_S, \vec{x}_{S+1}, \ldots]\tilde{Q} = 0$

$$\vec{x}_0 A_1^0 + \vec{x}_1 A_2^1 = 0 \tag{4}$$

$$\vec{x}_{r-1} A_0^{r-1} + \vec{x}_r A_1^r + \vec{x}_{r+1} A_2^{r+1} = \mathbf{0} \quad \text{for } 1 \leq r \leq S-1 \tag{5}$$

$$\vec{x}_{S-1} A_0^{S-1} + \vec{x}_S (A_1^S + R A_2) = \mathbf{0} \tag{6}$$

and the normalizing equation

$$\vec{x}_0 + \vec{x}_1 \vec{e} + \ldots + \vec{x}_{S-1} \vec{e} + \vec{x}_S (I - R)^{-1} \vec{e} = 1 \tag{7}$$

we can solve for $[\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_{S-1}, \vec{x}_S, \vec{x}_{S+1}, \ldots]$.

To find R, we know that

$$R^2 A_2 + R A_1 + A_0 = 0$$
$$\Rightarrow R = -A_0 A_1^{-1} - R^2 A_2 A_1^{-1}$$

Taking the initial value of $R = \mathbf{0}$ we can iteratively solve for R and can check the accuracy of this approximation by using equation 3. The value of R will converge since $-A_1^{-1}$ and $(A_0 + R^2 A_2)$ are positive. Hence, in each iteration, the value of R will increase monotonically.

Now, to solve for $\vec{\mathbf{x}}$, we represent each $\vec{x}_k$, where $k \leq S$ in terms of $\vec{x}_S$ and then obtain the value of $\vec{x}_S$. We can write $\vec{x}_S = \vec{x}_S I$.

Using equation 6 we can write

$$\vec{x}_{S-1} = -\vec{x}_S (A_1^S + R A_2) \Delta_{S-1}(\lambda^{-1})$$

where $\Delta_{S-1}(\lambda^{-1})$ is of dimension $\frac{(S+1)(S+2)}{2} \times \frac{S(S+3)}{2}$ and all its elements are 0 except, where the indices are equal, the value of that element is λ^{-1}.

From equation 5 we get

$$\begin{aligned} \vec{x}_{S-r} &= -(\vec{x}_{S-r+1}A_1^{S-r+1} + \vec{x}_{S-r+2}A_2^{S-r+2})\Delta_{S-r}(\lambda^{-1}) \\ & \quad \text{for r = 2 to S} \end{aligned}$$

where $\Delta_{S-r}(\lambda^{-1})$ is of dimension $\frac{(S+r+1)(S-r+2)}{2} \times \frac{(S-r+1)(S+r+2)}{2}$.

To represent $\vec{x}_k$ for $k \leq S$ in terms of $\vec{x}_S$, we assume

$$\vec{x}_{S-r} = \vec{x}_S C_{S-r} \qquad \text{r = 0 to } S \tag{8}$$

The value that C_i will take at different i is as follows:

$$\begin{aligned} C_S &= I \\ C_{S-1} &= -(A_1^S + RA_2)\Delta_{S-1}(\lambda^{-1}) \\ C_{S-r} &= -(C_{S-r+1}A_1^{S-r+1} + C_{S-r+2}A_2^{S-r+2})\Delta_{S-r}(\lambda^{-1}) \\ & \quad \text{for r = 2 to S} \end{aligned} \tag{9}$$

The dimension of C_{S-r} is $\frac{(S+1)((S+2)}{2} \times \frac{(S-r+1)(S+2+r)}{2}$. From the above set of equations we can recursively solve for $C_i (i = 0, 1, \ldots, S-1)$.

Using equation 8, and equations 4 - 6, we can solve for $\vec{x}_S$.

$$\vec{x}_S[\sum_{r=0}^{S-1} C_r\vec{e} + (I-R)^{-1}\vec{e}] = 1 \tag{10}$$

$$\begin{aligned} \vec{x}_S[C_{r-1}A_0^{r-1} + C_rA_1^r + C_{r+1}A_2^{r+1}] &= \mathbf{0} \\ & \quad \text{for r=1 to } S-1 \end{aligned} \tag{11}$$

$$\vec{x}_S[C_0A_1^0 + C_1A_2^1] = 0 \tag{12}$$

Each value of r will give $S - r + 1$ equations. Using the normalizing equation, that is, equation 10 we will have the required number of equations to obtain the value of $\vec{x}_S$.

From $\vec{x}_S$, we can find the values of $\vec{x}_i$, for $i = 0, 1, \ldots S$, by using equation 8 and $\vec{x}_k$, for $k > S$, by using the relation $\vec{x}_k = \vec{x}_S R^{k-S}$. In this way, we can obtain the steady state joint probability vector $\vec{\mathbf{x}}$, for any value of S. The boundary probabilities, that is $(\vec{x}_0, \vec{x}_1, \ldots, \vec{x}_S)$ can be obtained from equations 11 and 12.

These steady state joint probabilities are then used to find the mean and the second moment of number of customers in the system and finally to derive the mean waiting time.

IV.3 Analysis of the Number of Customers

The mean and second moment of the number of customers in the system can be obtained exactly as in [19,9,21]. Using the following relations:

$$E[Q] = \sum_{i=0}^{S} i\vec{x}_i\vec{e} + \vec{x}_S R[(I-R)^{-2} + S(I-R)^{-1}]\vec{e} \tag{13}$$

$$\begin{aligned} E[Q^2] &= \sum_{i=0}^{S-1} i^2\vec{x}_i\vec{e} + \vec{x}_S[(S^2-2S+1)(I-R)^{-1} + (2S-3)(I-R)^{-2} \\ & \quad +2(I-R)^{-3} - S^2I]\vec{e} - (E[Q])^2 \end{aligned} \tag{14}$$

IV.4 Waiting Time Analysis

In order to derive the waiting time distribution, we consider an arriving (tagged) customer. States corresponding to the number of customers present in the system {0,1,2,...} and an absorbing state { *} form the state space of the CTMC. Thus the state space of the CTMC is $\{*\} \cup \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots, \mathbf{S-1}, \mathbf{S}, \mathbf{S+1}, \ldots\}$. On entering the absorbing state denoted by $*$, a tagged customer starts receiving service. This happens at the arrival of a server from vacation when the customer is at the head of the queue.

We define the transition rate matrix for this CTMC as Q_1, which is given by

$$Q_1 = \begin{array}{c} \\ * \\ 0 \\ 1 \\ \vdots \\ S-1 \\ S \\ S+1 \end{array} \begin{bmatrix} & & & & & & & & \\ 0 & 0 & & & & & & & \\ g_0 & D_0 & & & & & & & \\ g_1 & F_1 & D_1 & & & & & & \\ \vdots & & F_2 & \ddots & & & & & \\ g_{S-1} & & & & D_{S-1} & & & & \\ g_S & & & & F_S & D & & & \\ & & & & & A_2 & D & & \\ & & & & & & & \ddots & \end{bmatrix}$$

g_i is a column vector of size $\frac{(i+1)(2S+2-i)}{2}$ and is given by

$$g_i = \begin{bmatrix} 0_{S+1} \\ 0_S \\ \vdots \\ 0_{S-i+1} \\ (S-i)T_2^0 \\ T_1^0 + (S-i-1)T_2^0 \\ \vdots \\ (S-i)T_1^0 \end{bmatrix}$$

where 0_i is a column vector of size i with all elements equal to 0. g_i gives the rate at which the tagged customer enters the absorbing state. This only happens when the number of customers ahead of the tagged customer are equal to the number of servers at the queue and then a server arrives after completing its vacation.

The matrix D_i gives the rate at which servers arrive at the queue and hence the customers ahead of the tagged customer start receiving service, and is given by A_1^i but without the λ's and E^i. In matrix Q_1, λ, the customer arrival rate, is not required since we are considering a FCFS queue and hence customers that arrive after the tagged customer do not have any impact on the analysis of waiting time. E^i is also not included because the tagged customer is present at the server arrival instant and therefore it will not take another vacation instantaneously.

The matrix $F_i = A_2^i$ gives the rate at which customers depart from the system. When the number of customers ahead of a tagged customer exceed S then D_i corresponds to D, which is given by $A_1 + \lambda I$. Here I is an identity matrix of dimensions $\frac{(S+1)(S+2)}{2}$.

To derive the mean waiting time, we apply a method used in[19] and [21] with modifications as required by our model. The basic intent is to find the time it takes for the tagged customer to reach the absorbing state $(*)$. At steady state, the tagged customer on its arrival will see the system in state $(i, j; k)$ with probability $x_{i,j;k}$. The tagged customer will not receive service immediately

upon arrival since it must wait for the customers which are ahead of it to receive service (handled by D's and A_2's). If all customers ahead of it have received or are receiving service it must wait for a server to arrive from vacation (handled by g_i's). Thus the tagged customer receives service only when the number of servers in the queue becomes equal to the number of customers present and then a server arrives at the queue after vacation.

Define $\mathbf{y(t)} = (\mathbf{y}_*(\mathbf{t}), \mathbf{y_0(t)}, \mathbf{y_1(t)}, \ldots)$, where $\mathbf{y_i(t)} = \{y_{i,j;k}(t)\}$ is of size $\frac{(i+1)(2S-i+2)}{2}$ when $i < S$ and $\frac{(S+1)(S+2)}{2}$ when $i \geq S$ and denotes the probability of i customers in the system, j servers at the queue and k servers in phase 1 of the vacation present at time t. $\mathbf{y}_*(\mathbf{t})$ is the probability that the tagged customer is in the absorbing state at time t. Because of the memoryless property of the Poisson arrival process, at time 0, $\mathbf{y(0)} = \{0, \vec{x}_0, \vec{x}_1, \vec{x}_2, \ldots\}$, where $\vec{x}_i$s are the steady state probabilities obtained earlier. Let $w(t)$ denote the pdf of waiting time. Then $w(t) = \mathbf{y}_*(\mathbf{t})$.

The tagged customer sees the system in state $(i, j; k)$ with probability $y_{i,j;k}(0)$ for $i \geq S$, the LST of the first passage time to a state $(S, j^{'}; k)$ in $\mathbf{S}$ is given by the $(\frac{j^{'}(2S+3-j^{'})}{2} + k + 1)$th element of the row vector $\Psi(s)$.

$$\Psi(s) = \sum_{i=S}^{\infty} \mathbf{y_i(0)}[(sI - D)^{-1}A_2]^{i-S} \tag{15}$$

Let $\phi_j(i, s; k)$ be the LST of the absorption time to state $*$ given that the process starts from state $(i, j; k)$, for $0 \leq i \leq S$, $0 \leq j \leq i$ and $0 \leq k \leq S - j$. Let $\Phi(i, s)$ denote the column vector of dimension $\frac{(i+1)(2S-i+2)}{2}$ containing the elements $\phi_j(i, s; k)$. On the basis of Q_1 we can write the following relations:

$$\Phi(0, s) = (sI - D_0)^{-1}g_0 \tag{16}$$

$$\Phi(i+1, s) = (sI - D_{i+1})^{-1}F_{i+1}\Phi(i, s) + (sI - D_{i+1})^{-1}g_{i+1} \tag{17}$$

$$0 \leq i \leq S - 1$$

The LST for the waiting time distribution is given by

$$W^*(s) = \sum_{i=0}^{S-1} \mathbf{y_i(0)}\Phi(i, s) + \Psi(s)\Phi(S, s) \tag{18}$$

Mean Waiting Time

The mean waiting time can be obtained from $W^*(s)$:

$$E[W] = -\sum_{i=0}^{S-1} \mathbf{y_i(0)}\Phi^{'}(i, 0) - \Psi^{'}(0)\vec{\mathbf{e}} - \Psi(0)\Phi^{'}(S, 0) \tag{19}$$

The first term gives the mean time to reach an absorbing state by the tagged customer if the system is in a state $\leq (\mathbf{S} - \mathbf{1})$ on its arrival; the second and third terms give the time to reach the absorbing state if the system is in state $\geq \mathbf{S}$ on the arrival of a tagged customer.

To solve for the mean waiting time we must calculate the value of each term in equation 19. Differentiating and substituting $s = 0$ in equation 16 will give

$$\Phi^{'}(0, 0) = -(-D_0)^{-1}I(-D_0)^{-1}g_0 \tag{20}$$

Similarly, differentiating $\Phi(i+1, s)$ and substituting $s = 0$ in equation 17 and using the relation $F_i\vec{e} + D_i\vec{e} + g_i = 0$ gives

$$\Phi^{'}(i+1, 0) = D_{i+1}^{-1}[\vec{e} - F_{i+1}\Phi^{'}(i, 0)] \tag{21}$$

Thus we can find $\Phi^{'}(i,0)$ recursively. Since $\mathbf{y_i(0)} = \vec{x_i}$, and using equations 20 and 21 we can solve the first term of equation 19

The value of $\Psi(0) = \sum_{i=S}^{\infty} \mathbf{y_i(0)} U^{i-S}$, where $U = (-D)^{-1} A_2$, is obtained by substituting $s = 0$ in equation 15. The value of $\Psi(0)\vec{e} = 1 - \sum_{i=0}^{S-1} \vec{x_i}\vec{e}$, since $U\vec{e} = \vec{e}$ due to the relation $A_2\vec{e} + D\vec{e} = 0$. The value of $\Psi(0)\vec{e}$ can also be used, as mentioned in [19] to obtain an approximate value of $\Psi(0)$ by finite summation. Using equations 20 and 21 we can get the value of $\Phi^{'}(S,0)$, thus we can solve the third term of equation 19.

$$\Psi^{'}(0) = -\sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} \sum_{j=0}^{k-1} U^j (-D)^{-1} U^{k-j}$$

where $U = (-D)^{-1} A_2$. Using the $U\vec{e} = \vec{e}$ relationship we obtain

$$-\Psi^{'}(0)\vec{e} = \sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} \sum_{j=0}^{k-1} U^j (-D)^{-1} \vec{e} \tag{22}$$

To obtain the value of $-\Psi^{'}(0)\vec{e}$ (second term of equation 19) from equation 22 we modify the method used in [19] and [21]. We define a stochastic matrix U^0 by deleting the last row and column of our U matrix. We can obtain the values of vector $\vec{\mathbf{u^0}}$ by using the relations that it should satisfy $\vec{\mathbf{u^0}} U^0 = \vec{\mathbf{u^0}}$ and $\vec{\mathbf{u^0}}\vec{e} = 1$. A square matrix U_2 can be constructed in the following way

$$\begin{aligned} (U_2)_{kk'} &= u^0_{k'}, \qquad \text{for } 0 \le k \le \tfrac{(S+1)(S+2)}{2} - 1,\ 0 \le k' \le \tfrac{(S+1)(S+2)}{2} - 2 \\ &= 0, \qquad \text{for } 0 \le k \le \tfrac{(S+1)(S+2)}{2} - 1,\ k' = \tfrac{(S+1)(S+2)}{2} - 1. \end{aligned}$$

The following relation is satisfied owing to the property $UU_2 = U_2U = U_2$

$$\sum_{r=0}^{k-1} U^r (I - U + U_2) = I - U^k + kU_2 \tag{23}$$

Using this relation and the fact that $(I - U + U_2)^{-1}$ exists[22], equation 22 can be simplified to

$$\begin{aligned} -\Psi^{'}(0)\vec{e} = \{&\sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} - \sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} U^k \\ &+ \sum_{k=1}^{\infty} k\mathbf{y_{k+S}(0)} U_2\} (I - U + U_2)^{-1} (-D)^{-1} \vec{e} \end{aligned} \tag{24}$$

The value of $-\Psi^{'}(0)\vec{e}$ can be calculated by substituting the following values:

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} &= \vec{x}_S((I-R)^{-1} - I) \\ \sum_{k=1}^{\infty} \mathbf{y_{k+S}(0)} U^k &= \Psi(0) - \vec{x}_S \\ \sum_{k=1}^{\infty} k\mathbf{y_{k+S}(0)} &= \vec{x}_S((I-R)^{-2} R) \end{aligned}$$

The second relation is obtained from equation 15 by putting $s = 0$. Thus we can solve the second term of equation 19. Now all three terms of equation 19 can be solved and we can obtain the mean waiting time.

We are able to calculate the value of steady state joint probability and the mean waiting time using algorithms for this model. In Appendix A we consider the analysis of a special case of the phase distribution, viz., the exponential distribution

IV.5 Vacations with Higher Order Phase Distribution

In this paper we have discussed the 2-phase vacation distribution for the sake of mathematical tractability. However, the model is extendable to higher phase distributions. For a 3-phase distribution, the states can be defined as $(i, j; k, l)$, where l denotes the servers in phase 2. The remaining $S - j - k - l$ servers are in phase 3.

In the 3-phase distribution the entries within the matrices B_j will be submatrices B_{jk} for possible value of k, number of servers in phase 1. This extension will make the dimensions of matrices A_2, A_1, A_0 and R an order higher (that is, $O(n^3)$ instead of $O(n^2)$). The dimensions will be $\sum_{i=0}^{S} \frac{(i+1)(i+2)}{2} = \frac{1}{6}S^3 + S^2 + \frac{11}{6}S$. The same technique used for the 2-phase vacation distributions can then be applied to this 3-phase vacation distribution.

V Numerical Results and Their Analysis

In this section we first present verification of our model implementation and then present numerical results obtained for our model. The effect of various stopping criteria on the mean customer number in the system and mean waiting time are discussed. We discuss the effect of arrival, service and vacation rates on the mean number of customers in the system and the mean waiting time for the special case of exponentially distributed vacations. We perform similar experiments for the more general phase distribution. Special cases of the order 2 phase distribution: Erlangian distribution and hyperexponential distribution results are also presented. Using the stability condition in (1) we define the stability factor for the system as $\alpha = \frac{\lambda}{S} * (\frac{1}{\mu} + \frac{1}{\theta})$. The load per server is defined as $\rho = \frac{\lambda}{S\mu}$.

V.1 Model Verification

The model implementation is verified using exponentially distributed vacations. This is done in several ways: first using an alternative analytical model that provides exact results for a small number of servers, then using simulation for a larger number of servers and finally looking at the extreme case of very high vacation rate. We also do some experiments on the phase model to see the effect of the stopping criteria on the results.

In [22] an alternative method was used to derive the mean number of customers in the system for the case of exponentially distributed vacations and small values of S. The state was defined as $\{(i, j) : i \geq j; j = 0, 1, \ldots, S\}$, where i denotes the number of customers in the system, and j the number of servers at the queue. Based on this model, balance equations are found. Unfortunately the number of variables in the balance equations is larger than the number of equations obtained so the system could not be solved fully. By applying a technique used in [14,18,20] results were found for $S < 4$. In Table 1, we give the mean number in the system obtained from the Balance Equation Method[22] and the Matrix Geometric Method for various values of λ, μ and θ. As is shown in the table, the results for $S = 2$ and $S = 3$ are almost identical. What differences occur is due to the stopping criteria used in the Matrix Geometric solution. For higher values of S, results for the exponential case were verified using a simulation. In figures 2 and 4 simulation results for the curves with $\mu = 1$ and $\theta = 2$ are shown. The analytical results are very close to the simulation results.

S	λ	μ	θ	L (Balance)	L (Matrix Geometric)
2	1	1	1.2	8.9091	8.9085
2	.5	.5	.55	16.4127	16.4084
2	.3	.3	.35	10.4103	10.4075
3	.5	.25	1	5.2992	5.2989
3	.5	.25	.6	14.6692	14.6664
3	1	1	1	2.2630	2.2630
3	2	1	2.5	12.3337	12.3321

Table 1: Comparison of Mean Number in System obtained from Balance Equation and Matrix Geometric Methods

To further check the implementation we compared the results of our model at high vacation rate ($\theta = 10^4$) with the $M/M/S$ queue without vacation. The results in Table 2 show that the models are very close. The values of the mean waiting time are higher for the vacation model since the servers still take a (brief) vacation after serving each customer. Further note that the results satisfy Little's result (that is, $L = \lambda * (W + 1/\mu)$).

S	λ	μ	L (M/M/S)	L (Matrix Geometric)	W (M/M/S)	W (Matrix Geometric)
2	.95	.5	19.4872	19.5014	18.5128	18.5280
3	7	3	4.4733	4.4773	.3057	.3063
5	9	2	11.3624	11.3790	.7625	.7643

Table 2: Comparison of Mean Number in the System and Mean Waiting Time obtained from M/M/S queue and $M/M/S/V_M$ queue at $\theta = 10^4$

Although the matrix geometric method is exact when summations are taken to infinity, there are 2 terms R and $\Psi(0)$ where a stopping criteria is used to limit the number of iterations. This introduces some approximation error. In the next two tables we consider the effect of the stopping criteria on these values. In Table 3 we present the mean waiting time for different stopping criteria used to calculate $\Psi(0)$. For the approximation, the number of terms used to calculate $\Psi(0)$ are 10S, 20S and 30S. The other stopping criteria used is to calculate $\Psi(0)$ until error in $\Psi(0)\vec{e}$ is less than 10^{-3}, 10^{-5} and 10^{-7}. The stopping criteria for R is kept the same. Based on these results, it is evident that with a number of iterations on the order of $30S$, one can achieve an accuracy within 10^{-3}, for the mean waiting time. In Table 4 we present the mean number of customers in the system and mean waiting time for the stopping criteria used to calculate R. The stopping criteria used is to have the difference between each term of $A_2\vec{e}$ and $RA_0\vec{e}$ less than 10^{-3}, 10^{-5} and 10^{-7}. The stopping criteria for $\Psi(0)$ is kept the same. It is evident that the results converge quickly. For the results in this paper we use a stopping criteria of 10^{-5} for both $\Psi(0)$ and R.

V.2 Results for Exponentially Distributed Vacations

To gain an understanding of the performance of this queue, we first study the simpler model with exponentially distributed vacations. We study the effect of the parameters of λ, μ and θ on the

S	λ	μ	$\frac{1}{\bar{v}}$	W($10S$)	W($20S$)	W($30S$)	W(10^{-3})	W(10^{-5})	W(10^{-7})
1	3	4	20	2.37604935	2.38441752	2.38690536	2.38791298	2.38795746	2.38795790
2	4	2.5	20	1.77347542	1.77682250	1.77714980	1.77714051	1.77718479	1.77718527
5	3	4	.8	2.60800158	2.60966106	2.60966519	2.60850426	2.60965302	2.60966509
8	3	4	.465	2.77454574	2.77464388	2.77464389	2.77255236	2.77462380	2.77464370

Table 3: Effect of approximation on the Mean waiting time

S	λ	μ	$\frac{1}{\bar{v}}$	L(10^{-3})	L(10^{-5})	L(10^{-7})	W(10^{-3})	W(10^{-5})	W(10^{-7})
1	3	4	20	7.893848	7.913852	7.914060	2.38092346	2.38690536	2.38696757
2	4	2.5	20	8.693833	8.708725	8.708877	1.77382056	1.77714980	1.77718388
5	3	4	.8	8.568927	8.578985	8.579086	2.60666492	2.60966519	2.60969544
8	3	4	.465	9.060869	9.073919	9.074043	2.77074467	2.77464389	2.77468100

Table 4: Effect of approximation of R on the Mean Number in System and Mean Waiting Time

mean number in the system and mean waiting time. Figures 2 and 3 show the mean number of customers in the system as λ is increased for the cases of 5 and 8 servers respectively. The value of λ is chosen such that $\alpha \leq .98$. As expected, increasing the arrival rate of customers causes an increase in the mean number in the system. When the stability factor, that is α, approaches .98, the mean number in the system increases in an unbounded fashion since at this high load the servers are unable to cope with the arrival rate of customers.

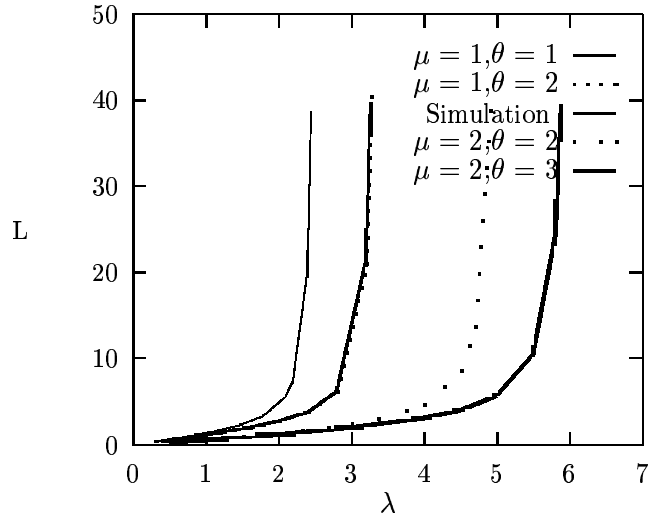

Figure 2: Mean Number in the System vs λ ($S = 5$)

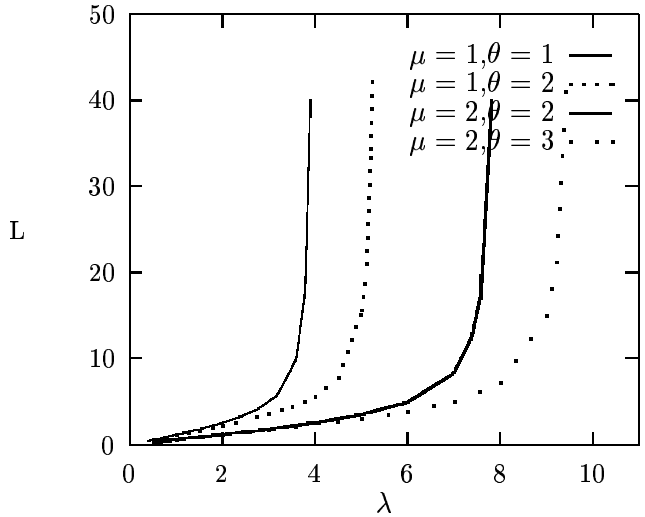

Figure 3: Mean Number in the System vs λ ($S = 8$)

On these graphs, results for different values of vacation and service rate are shown as well. From the figures, it is clear that systems with larger values of vacation rate θ can support greater values of λ. This can be understood mathematically from the definition of the stability factor. For higher values of θ, that is for low mean vacation time, the arrival rate of the system can be higher since the mean time between server availability is lower. We observe the same effect for different values of μ. When the service rate, μ, increases, the mean service time decreases and hence, the

mean number in the system decreases. With the increase in μ the customers are served faster and therefore leave the queue faster.

The mean waiting times for these experiments were alse calculated and the results are given in figures 4 (for $S = 5$) and 5 (for $S = 8$). We see that as λ increases the mean waiting time also increases. As $\alpha \to .98$, the waiting time increases in an unbounded fashion. Higher λ means the rate of customer arrival is greater and hence an arriving customer sees on average more customers in the system and thus must wait longer before receiving service. As shown in the figure, for higher

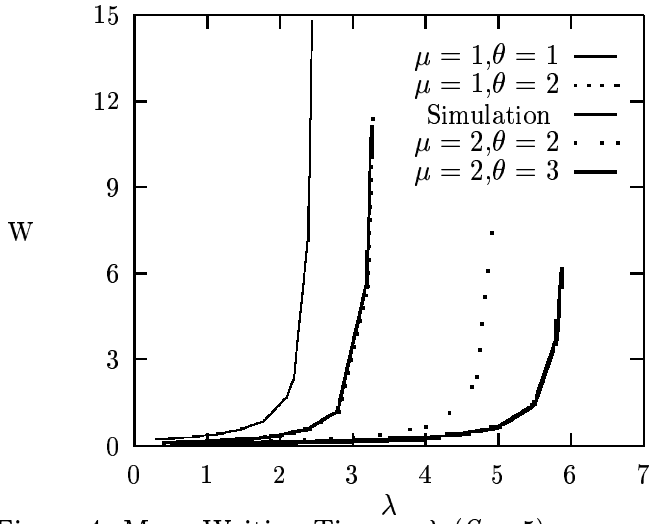

Figure 4: Mean Waiting Time vs λ ($S = 5$)

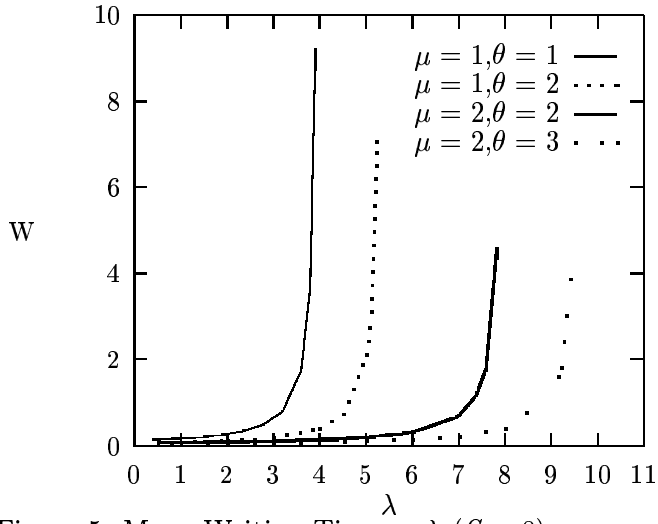

Figure 5: Mean Waiting Time vs λ ($S = 8$)

values of θ we can accommodate a higher arrival rate without an increase in waiting time. The same behavior occurs for higher values of μ. For higher θ or μ values, the mean server availability increases and thus reduces the time for which the customers must wait. Note that the mean waiting time in figures 4 and 5 is lower at $\alpha = .98$ for higher values of θ and μ. As seen in figures 2 and 3, this is not true for mean number in the system.

V.3 Constant Stability factor: Exponential Distribution

In the last section, the increasing load strongly effected the values for the metrics. In this section, in order to gain more insight into the effect of the parameters, we study the effect of increasing the number of servers when the stability factor is kept constant. In figure 6 we plot mean number in the system against the number of servers for stability factor .95 and two different loads (.75 and .9). For load .75, two curves are shown: one for the case where $\lambda = S$, $\mu = 1.333$ and $\theta = 5$ and the other for the case where $\lambda = 0.5S$, $\mu = 0.667$ and $\theta = 2.5$. Similarly, for load 0.9, two curves are shown; one with long service times and the other with more arrivals and longer vacation durations. To keep the load constant for a particular curve, we vary λ with S. As seen in the figure, for the same load but for different μ and θ, the mean number in the system is the same. As λ and S increases, there is a slight increase in the mean number in the system. This is due to the increase in the arrival rate of customers which increases the number of customers. Though the number of servers is increased proportionally, the servers still must take vacations and cannot serve the queue all the time. Therefore the arrival rate has more effect than the increase of servers on the mean number of customers in the system. In figure 7 we plot the mean waiting time against S where λ is varied to keep the load constant at .75 and .9. We find that with the increase in the number of servers the mean waiting time reduces considerably. This change is due to the increase in the number of servers which are able to serve the queue better even though there is a slight increase

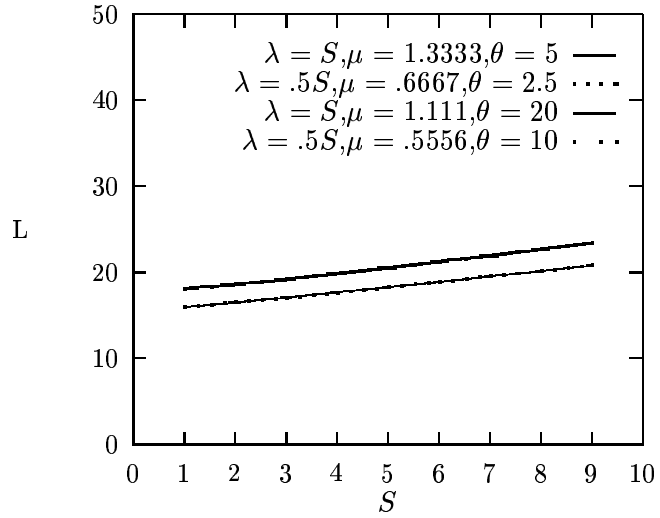

Figure 6: Mean Number in the System vs S ($\alpha = .95$)

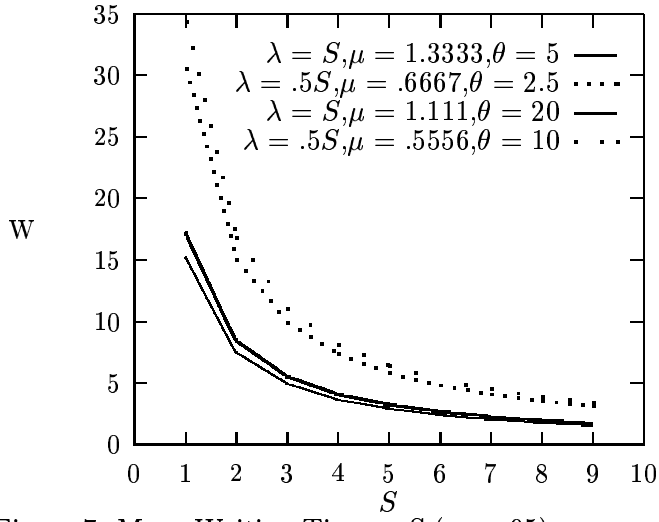

Figure 7: Mean Waiting Time vs S ($\alpha = .95$)

in number in the system due to increased arrival rate. The waiting time also becomes relatively constant after S increases to a certain value. In contrast to the number of customers metric, mean waiting time is significantly larger when mean service time is larger for the same load per server.

In figure 8 and figure 9 we plot the number in the system and the waiting time against S for $\alpha = 0.9$. Two sets of curves are shown corresponding to cases of $\lambda = 3$ and $\lambda = 4$. The stability factor is kept the same by changing the vacation rate or service rate. We would expect that the waiting time will be proportional to the mean service and vacation times (divided by the number of servers) and thus, the curves of figure 9 are relatively flat when the stability factor is kept constant. In figure 8, we see that the increase in service duration causes the mean number in the system to rise for the same stability factor at a much faster rate than the increase in vacation time. This is because when the service duration increases, the customers are in the system for a longer period of time. For the case where θ is varied, the service rate remains the same so the number in the system does not rise as quickly. The number in the system does rise slightly though due to the effect of taking longer vacations when multiple vacations occur. This results in the small increase in waiting times as well. These results follow Little's law.

In figures 8 and 9, the results exhibit a decrease, followed by an increase which results in minima. This is not an anomaly, and can be explained as follows. Consider the case where μ is constant and $\alpha = 0.9$, where $\alpha = \frac{\lambda}{S\mu} + \frac{\lambda}{S\theta}$. As S increases, two things happen:

1. the first term decreases, and since its contribution to the mean number of customers in the system is $(1 - \frac{\lambda}{S\mu})^{-1}$, it decreases quickly first, and then more slowly (negative exponential).

2. the second term increases linearly, and therefore the mean vacation time also increases linearly.

The mean waiting time (and proportionally the mean number of customers in the queue, since λ is constant) is a linear combination of 1 and 2, which when added together exhibit such minima. The minima in case of constant θ can be explained similarly.

V.4 Phase Distribution Results

After gaining an understanding of the effect of the parameters (particularly vacation rate) for the special case of exponentially distributed vacation times, we now consider the case where vacation times follow a phase distribution. The experiments that were run for the exponential case were

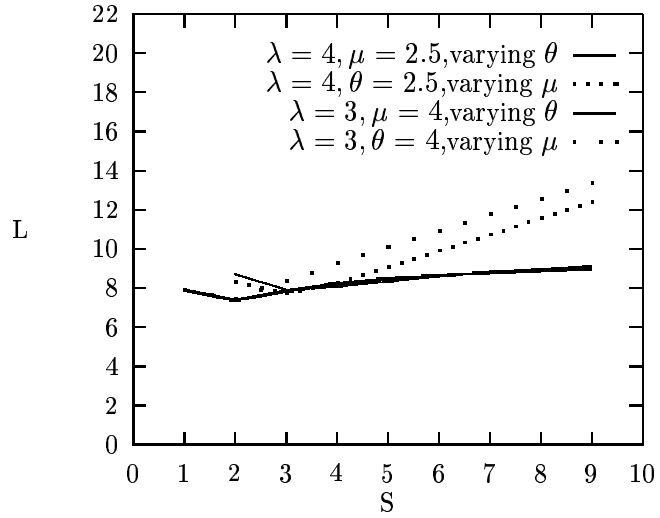

Figure 8: Mean Number in the System vs S ($\alpha = .90$)

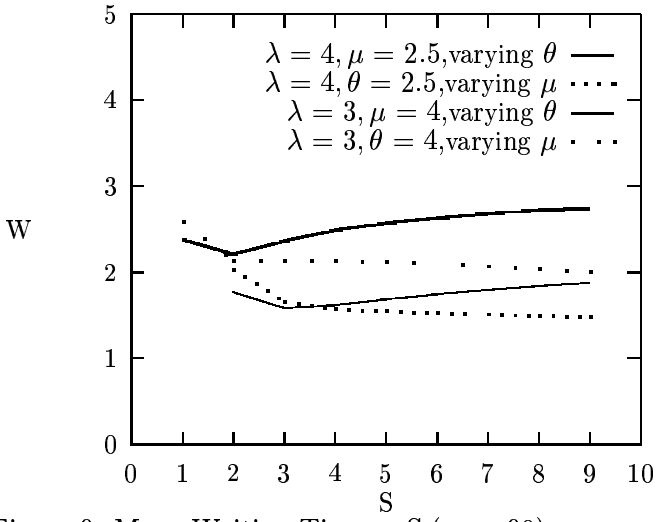

Figure 9: Mean Waiting Time vs S ($\alpha = .90$)

also run for the phase distributed vacation case. The results are generally similar so not all of the graphs are shown here. Two special cases of phase distribution of order 2 are also considered, namely: Erlangian and Hyperexponential[23]. To define an Erlangian distribution, the matrix $\mathbf{T}$ and vector T^0 should take the following values

$$\mathbf{T} = \begin{bmatrix} -T_{12} & T_{12} \\ 0 & -T_2^0 \end{bmatrix}$$

and

$$T^0 = \begin{bmatrix} 0 \\ T_2^0 \end{bmatrix}$$

The value of $\nu_1 = 1$ and hence $\nu_2 = 0$. For the Hyperexponential distribution the matrix $\mathbf{T}$ and vector T^0 should take the following values

$$\mathbf{T} = \begin{bmatrix} -T_1^0 & 0 \\ 0 & -T_2^0 \end{bmatrix}$$

and

$$T^0 = \begin{bmatrix} T_1^0 \\ T_2^0 \end{bmatrix}$$

The effect of λ, μ and $\bar{v}$ for all three cases is similar to the case where vacation follows an exponential distribution. The stability condition for phase distribution is the same as that for exponential vacation time and hence the definition of stability factor remains the same.

We present the effect on mean waiting time of increasing λ for Erlangian, Hyperexponential and Phase vacation distributions in figures 10, 11 and 12 respectively. In each graph, different combinations of μ and $\frac{1}{\bar{v}}$ are considered. The values of λ are chosen such that the value of $\alpha \leq .98$. As in the exponential case (see Figure 4), we notice that generally the mean waiting time decreases for higher values of μ and $\frac{1}{\bar{v}}$. However in certain cases, such as with the hyperexponential distribution the mean waiting time when $\mu = 1$ and $\frac{1}{\bar{v}} = 2$ is higher than when $\mu = 1$, $\frac{1}{\bar{v}} = 1$ for $\alpha = .98$. The reason for this is that, although the parameters chosen to describe the different vacation distributions have the same mean value, they have different higher moments.

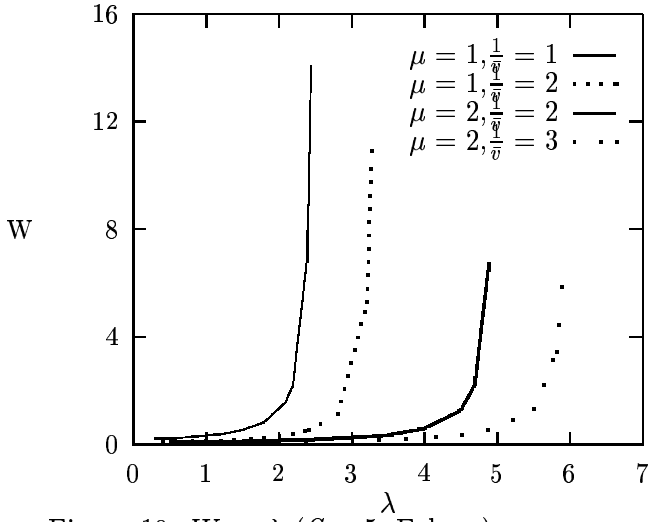

Figure 10: W vs λ ($S = 5$, Erlang)

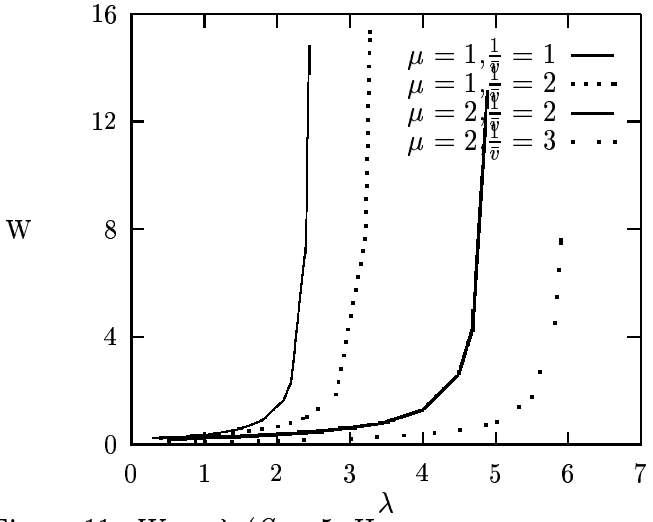

Figure 11: W vs λ ($S = 5$, Hyperexponential)

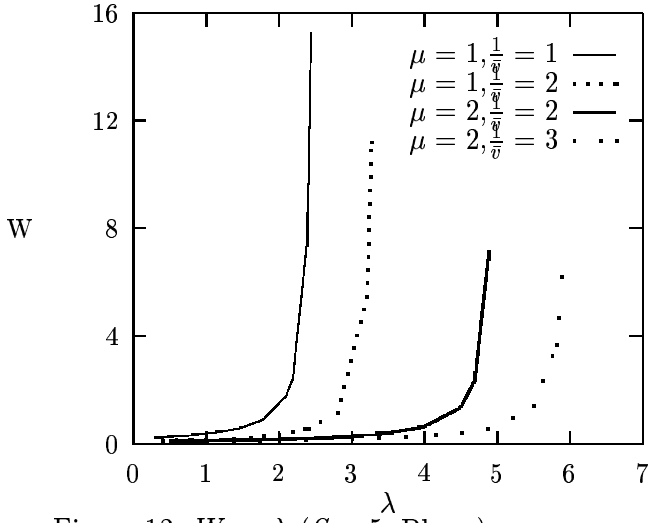

Figure 12: W vs λ ($S = 5$, Phase)

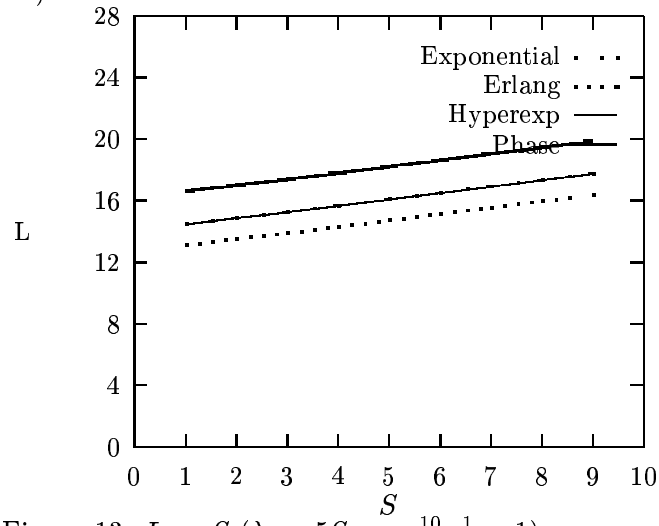

Figure 13: L vs S ($\lambda = .5S, \mu = \frac{10}{9}, \frac{1}{\bar{v}} = 1$)

These observations are further supported in figures 13 and 14 which show how the mean number in the system and mean waiting time depend on S for different types of Phase distributions with the same parameters of λ, μ and $\frac{1}{\bar{v}} = 1$. The variance for the Phase, exponential, hyperexponential and Erlangian are 1.7777, 1, 1, and .5 respectively. For higher variance, as expected, the mean number of customers in the system and the mean waiting time are both higher.

Finally, figure 15 shows how the mean waiting time varies with S when the stability factor is kept at 0.9. The graph is similar to the exponential case (See Figure 9). Note, however, that the curves for the case of varying mean vacation time are more variable than in the exponential case. This is because, by varying the first moment (to keep constant stability factor), the higher moments of the vacation distribution change as well.

V.5 Summary

The steady state analysis of the M/M/S/V_M queue with 1-limited service has been considered in this paper. The Matrix Geometric solution technique, gives us the mean and second moment of number of customers in the system and the mean waiting time for phase distributed vacation times.

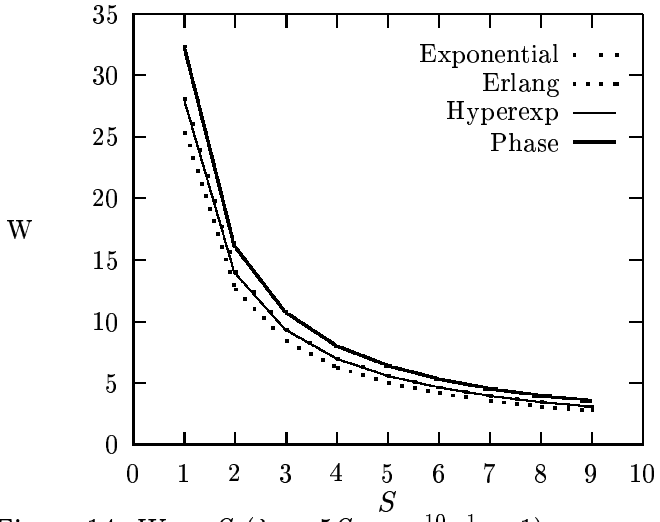

Figure 14: W vs S ($\lambda = .5S, \mu = \frac{10}{9}, \frac{1}{\theta} = 1$)

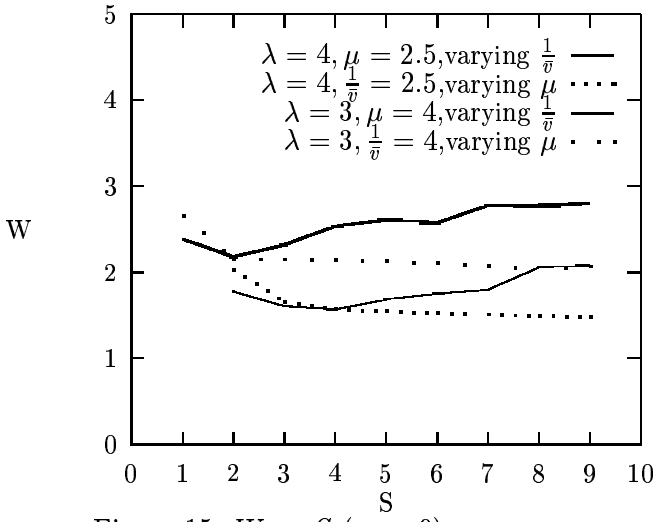

Figure 15: W vs S ($\alpha = .9$)

These performance measures are obtained algorithmically.

From the study, we make the following important conclusions about this type of queue:

1. The mean number of customers in the system is affected more by decreasing service rate than the vacation rate since with larger mean service time, the customers will remain in the system for a longer time.

2. The vacation rate may affect the mean waiting time more than the service rate does since, in a 1-limited case, the customer waiting for service must wait for a new server to arrive since the servers at the queue have to leave after serving their respective customers. Also when vacation durations are larger, the effect of multiple vacations will be stronger.

3. Increasing the number of servers and the arrival rate in order to keep the load the same, yields a slight increase in the mean number in the system and a decrease in the waiting time.

4. The results for the more complex phase-distributed vacation model, are similar in nature to the exponential case, however, there are differences due to the higher moments of the vacation distribution. This has ramifications toward the ability to more closely model vacations in real systems.

VI Applications

In this section we present two applications where the model analyzed in the this paper can be used. First we discuss the multiqueue mutliserver polling system and then we present a mobile communication systems application.

VI.1 The multiqueue mutliserver polling system

In this case, we concentrate on just one queue, and a server's vacation corresponds to the time spent by the server on serving other queues as well as on walking between them. Since the vacation time distribution is dependent on the time spent at queues, an approximate technique can be used

in which the vacation time of server i can be expressed as

$$v_i = \sum_{j=1}^{N} w_{i,j} + \sum_{\substack{j=1 \\ j \neq k}}^{N} s_{i,j} \tag{25}$$

where $w_{i,j}$ is the time spent by server i in walking between queues j and $j+1$, while $s_{i,j}$ is the time spent by server i on serving customers, if any, at queue j. Since servers are identical, we can drop the i subscript in the above equation. If we let p_A define the probability that the queue is not empty when a server arrives at the queue, then, assuming independence between servers and queues,

$$p_A = \frac{\rho}{1-\rho}\frac{\bar{w}}{\bar{s}}$$

where $\rho = \frac{N\lambda\bar{s}}{S}$, $\bar{s}$ is the mean service time and $\bar{w}$ is the mean walking time from queue j to $j+1$. The derivation of p_A is as follows. At steady state for our model the ratio of server serving the customers and walking is equal to $\frac{\rho}{1-\rho}$, where ρ is the load per server. The ratio of the service time and walking time can also be given by $\frac{Np_A\bar{s}}{N\bar{w}}$. Equating the two relations gives the value of p_A. s_j is equal to the customer service time with probability p_A; otherwise, it is equal to zero.

Equation 25 can be used to match moments of the right hand side to moments of v, therefore determining the parameters of the phase distribution. That is, after solving the model, assuming any reasonable parameters for the distribution of v, one can feed the values obtained from the model in equation 25 to obtain new values for the parameters of v. This procedure can be iteratively repeated in order to improve the accuracy of the results.

VI.2 Mobile communication systems

In cellular mobile communication systems, calls can be of two types: newly arriving calls; and handover (or hand-off) calls which cross cell boundaries, and demand service from a new cell base station. The two-phase vacation system can be used to model the performance of new arriving calls, in which handover calls have a higher non-preemptive priority over new calls[24]. This is particularly important since the GSM standard for mobile communication systems allows queueing of calls[25]. The S servers are the S channels in a cell, with the service time being the holding time of a call that is admitted to the cell. The two phases of the vacation correspond to a service phase for a handed over call, and the phase in which the queue of handover calls are inspected, respectively. Therefore, the vector ν is given by $[0, 1]$, and $\nu_3 = 0$, i.e., the process starts by inspecting the queue of handover calls. The matrices $\mathbf{T}$ and T^0 are given by

$$\mathbf{T} = \begin{bmatrix} -\beta & \beta \\ p\gamma & -\gamma \end{bmatrix}$$

and

$$T^0 = \begin{bmatrix} 0 \\ (1-p)\gamma \end{bmatrix}$$

where β the rate of the exponentially distributed service time of the handover call, while γ is a very large rate that corresponds to the small time required for the inspection of the queue. p is the probability that a handover call is found waiting in the queue, while $1-p$ is the probability that the handover call queue is empty, which causes the server to go back to serve waiting new calls, if available.

In the above model, unless the cell size is very small, i.e., microcells, then the rate of arrivals of handover calls is very small, and the above two-phase distribution should be sufficiently accurate. The system can then be used to evaluate the performance of new calls in mobile communication systems.

VII Conclusions

We have presented a method of finding the steady state joint probability of number in the system and busy servers, the mean and the second moment of number in the system and mean waiting time for the M/M/S/V_M 1-limited model. The vacation follows phase distribution of order 2. From these results we have found that for applications which can be modeled as 1-limited service models, the service rate should be kept high if it is required to have a smaller mean number of customers in the system, but, if the emphasis is on having a smaller mean waiting time, then the vacation rate should be high.

The model analyzed in this paper has a number of important applications. Two such applications are multiqueue multiserver polling systems and mobile communication systems are discussed in the paper.

Appendix A. The Exponential Model

The phase distribution of order 1 is equivalent to the exponential distribution. For the exponential case the states can be defined as (i, j) where i denotes the number of customers and j denotes the number of servers at the queue. These variables can take on the following values $0 \leq i \leq \infty$ and $0 \leq j \leq S$. The value of j must be $\leq i$, since this is a multiple vacation model as described earlier.

We can define this infinitesimal generator $\tilde{Q}$ as described for phase distribution. The row and column position of each submatrix in $\tilde{Q}$ indicate the number of customers present in the system before and after the transition, respectively. The matrices A_2's, A_1's and A_0's can be obtained directly from the phase distribution matrices by assuming a phase distribution of order 1. To make the distribution exponential the value of $\nu_2 = 1$, $T_{22} = -\theta$ and $T_2^0 = \theta$ and the values of other phase parameters are 0. The submatrices B_i's and E's are all of dimension 1×1 since the phase distribution of vacation is order 1.

On the basis of our phase model we can obtain the values of the submatrices. The matrices A_1^0 and A_0^0 are $-\lambda$ and $\begin{bmatrix} \lambda & 0 \end{bmatrix}$, respectively.

The matrix A_2^i is an $(i+1) \times i$ matrix for i= 1 to S. Its contents are

$$A_2^i = \begin{bmatrix} 0 & & & & \\ \mu & 0 & & & \\ & 2\mu & 0 & & \\ & & & \ddots & 0 \\ & & & & i\mu \end{bmatrix}$$

and it gives the rate of departure of a customer and therefore also the server.

The matrix A_0^i is of size $(i+1) \times (i+2)$, for i=1 to $S-1$, and its contents are

$$A_0^i = \begin{bmatrix} \lambda & & & & \\ & \lambda & & & \\ & & \ddots & & \\ & & & \lambda & 0 \end{bmatrix}$$

and it denotes the rate of arrival of a customer at the queue.

Let $f_i = \lambda + i\mu + (S-i)\theta$. We can write the contents of A_1^i, where i=1 to S as

$$A_1^i = \begin{bmatrix} -f_0 & S\theta & & & & \\ & -f_1 & (S-1)\theta & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & (S-i+1)\theta & \\ & & & & -(\lambda + i\mu) & \end{bmatrix}$$

This matrix is of dimension $(i+1) \times (i+1)$ and the off-diagonal elements give the rate of server arrival to the queue; while the diagonal elements give the rate at which the state remains unchanged and are obtained using the relationship $A_2^i\vec{e} + A_1^i\vec{e} + A_0^i\vec{e} = 0$.

When the number of customers at a queue exceeds S, the matrices A_2^i, A_1^i and A_0^i correspond to A_2, A_1 and A_0, respectively. Their dimensions are $(S+1) \times (S+1)$.

The matrix $A_0 = \lambda I$, A_2 is

$$A_2 = \begin{bmatrix} 0 & & & & & \\ \mu & 0 & & & & \\ & 2\mu & 0 & & & \\ & & & \ddots & 0 & \\ & & & & S\mu & 0 \end{bmatrix}$$

The matrix A_1 is as follows

$$A_1 = \begin{bmatrix} -f_0 & S\theta & & & & \\ & -f_1 & (S-1)\theta & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \theta & \\ & & & & -f_S & \end{bmatrix}$$

The steady state probability $\vec{\mathbf{x}}$ can be obtained as described for phase distribution.

The method of solving the mean waiting time is the same as that used in the previous model. The infinitesimal rate matrix Q_1 is similar to that described for phase distribution except that the values of the submatrices are different. Each element of the state space except $*$ represents $min(i,S)+1$ state pairs, (i,j), corresponding to $(i,0),(i,1)\ldots(i,min(i,S))$, where (i,j) represents i customers and j servers at the queue.

g_i is a column vector of size $i+1$, it gives the rate at which the tagged customer enters the absorbing state. The value is 0 for all states except where the number of customers present equals

the number of servers serving the queue, in which case the value is $(S-i)\theta$. F_i gives the rate at which the customers ahead of the tagged customer leave the queue and hence is identical to A_2^i. D_i is equal to $A_1^i + \lambda I - diag\{0, \ldots, (S-i)\theta\}$. It gives the rate at which the number of servers increases and the rate at which the customers ahead of a tagged customer remain same.

D has the same significance as D_i and is identical to $A_1 + \lambda I$. In matrix Q_1, λ, the customer arrival rate is not required since we are considering a FCFS queue and hence customers that arrive after the tagged customer do not have any impact on the analysis of waiting time. The transition rate matrix Q_1 is infinitesimal generator. Hence $g_i + F_i\vec{e} + D_i\vec{e} = 0$ and $A_2\vec{e} + D\vec{e} = 0$.

A technique and argument similar to that used for phase distribution of order 2 give us the value of mean waiting time.

Appendix B. List of Symbols

Term	Explanation
$(i, j; k)$	Denotes the state of the system, where i: number of customers in the system j: number of servers at the queue k: number of servers in Phase 1
$\vec{\mathbf{x}}$	Steady state probability of the system
$x_{(i,j;k)}$	denotes the probability of i customers, j servers present at the queue and k servers present in phase 1 of the vacation distribution
λ	Arrival rate
μ	Service rate
θ	Vacation rate
$\bar{v}$	Mean Vacation time
α	Stability Factor $= \frac{\lambda}{S}(\frac{1}{\mu} + \bar{v})$
ρ	Load per server $= \frac{\lambda}{\mu S}$
E[Q] or L	Mean Number of Customers in the System
E[W] or W	Mean Waiting Time
A_0 or A_0^i	gives the rate of arrival of a customer
A_1 or A_1^i	The offdiagonal matrices gives the rate of server arrival to the queue and diagonal matrices give the rate at which the server and customers queue remain the same
A_2 or A_2^i	gives the rate of service completion
B_0^j	denotes the rate of entering one of the phase of vacation after service completion by the server
B_1^j	denotes the rate of transition of servers among the two phases and the rate at which the number of servers and customers remain the same at the queue
B_2^j	denotes the rate of vacation termination of a server
B_3^j	denotes the rate of customer arrival
E^i	denotes the rate of transition of servers among the phases when the server on vacation completion finds no customer to serve at the queue

Table 5: Description of symbols

Acknowledgment

The authors acknowledge the constructive comments of the reviewers which helped improve the quality of the paper presentation.

This research was supported by the NSERC grant number OGP-121332 and OGP-9187.

References

[1] Doshi BT (1986). Queueing systems with vacations - a survey. *Queueing Systems*, 1:29–66.

[2] Scholl M and Kleinrock L (1983). On the M/G/1 queue with rest periods and certain service-independent queueing disciplines. *Operations Research*, 31:705–719.

[3] Miller LW (1964). Alternating priorities in multi-class queues, Ph.D. dissertation. Cornell University, Ithaca, New York.

[4] Cooper RB (1970) Queues served in cyclic order: Waiting times. *Bell System Technical Journal*, 49:399–413.

[5] Kamal AE and Hamacher VC (1989). Approximate analysis of non-exhaustive multiserver polling system with application to local area networks. *Computer Network and ISDN Systems*, 17:15–27.

[6] Morris RJT and Wang YT (1984). Some results for multi-queue systems with multiple cyclic servers. In: Rudin H and Bux W (*eds*). *Performance of Computer-Communication Systems*, North-Holland, pp. 245-258.

[7] Ajmone Marsan M et al (1992). Cycles and Waiting Times in Symmetric Exhaustive and Gated Multiserver Systems. *proceedings of IEEE INFOCOM.*

[8] Watson KS (1994). *Performance Evaluation of Cyclic Service Strategies - A Survey.* E. Gelenbe (ed), Elsevier-Science B. V. (North Holland).

[9] Neuts MF (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* The John Hopkins University Press, Baltimore and London.

[10] Neuts MF (1989). *Structured Stochastic Matrices of the M/G/1 Type and Their Applications*, Marcel Dekker, New York.

[11] Gail HR, Hantler SL and Taylor BA (1996). *Spectral Analysis, Advances in Applied Probability* 114-165.

[12] Mitrani IL and Chakka R (1995). Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Performance Evaluation 23:241–260*

[13] Heyman DP (1977). The t-policy for the M/G/1 queue. *Management Science*, 23:775–778.

[14] Levy Y and Yechiali U (1975). Utilization of idle time in an M/G/1 queueing system. *Management Science*, 22:202–211.

[15] Fuhrmann SW (1984). A note on the M/G/1 queue with server vacations. *Operations Research*, 32:1368–1373.

[16] Fuhrmann SW and Cooper RB (1985). Stochastic decomposition in the M/G/1 queue with generalized vacations. *Operations Research*, 33:1117–1129.

[17] Shanthikumar J (1984). Analysis of priority queues with server control. *OPSEARCH*, 27:183–192.

[18] Levy Y and Yechiali U (1976). An M/M/S queue with servers' vacations. *INFOR*, 14:153–163.

[19] Kao EPC and Narayanan KS (1991). Analysis of an M/M/N queue with servers' vacations. *European Journal of Operational Research*, 54:256–266.

[20] Mitrany IL and Avi-Itzhak B (1967). A many-server queue with service interruptions. *Operations Research*, 16:628–638.

[21] Neuts MF and Lucantoni DM (1979). A Markovian queue with N servers subjected to breakdowns and repairs. *Management Science*, 25:849–861.

[22] Tyagi A (1994). Analysis of M/M/S/V_M Queue, M.Sc. Thesis. Department of Computing Science, The University of Alberta, Edmonton, Alberta, Canada.

[23] Kleinrock L (1975). *Queueing Systems, Vol. I: Theory.* Wiley Interscience, New York.

[24] McMillan D (1995). Delay Analysis of a Cellular Mobile Priority Queueing System. *IEEE/ACM Trans. on Networking*, 3:310–319.

[25] Mouly M and Pautet MB (1994). The Evolution of GSM. Int. Zurich Seminar on Digital Comm., 1994.