AMO - Advanced Modeling and Optimization[1], *Volume 9, Number 1, 2007*

A Quorum Queueing System with Bernoulli Vacation Schedule and Restricted Admission

Khalid Alnowibet Lotfi Tadj
Department of Statistics and Operations Research
College of Science, King Saud University
P.O. Box 2455, Riyadh 11451, Saudi Arabia.
email: [knowibet,lotftadj]@ksu.edu.sa

May 25, 2007

Abstract

In this paper, we investigate the bulk service queue with Bernoulli vacation schedule and restricted admissibility introduced and studied by Madan and Abu-Dayyeh, [Revista Investigacion Operacional, Vol.24, No.2, pp.113-123, 2003]. Madan and Abu-Dayyeh assume exponentially distributed service times while we allow these times to be generally distributed. We obtain various performance measures that we use to derive the optimal values of the quorum size and maximum server capacity. The effect of the admission control parameter is further shown in a numerical example.

Keywords: Queue, quorum, Bernoulli schedule, vacation, restricted admission, optimal policy.

AMS Subject Classification: Primary 60K10, 60K25, secondary 90B22, 90B25.

1 Introduction

An interesting control problem for queues is that of admission control. In typical admission control problems, queue lengths are controlled by rejecting some of the incoming arrivals. Recently, Madan and Abu-Dayyeh [8] considered the following $M^x(RA)/M^{(b,n)}(V_S)/1(BS)$ queueing system:

- Customers arrive in batches of variable size according to a compound Poisson process, but not all arriving batches are allowed to join the system (RA: restricted admission).

[1]AMO - Advanced Modeling and Optimization. ISSN: 1841-4311

- Service is batch of fixed size b following a $\min(b, n)$ rule, which means that a fixed number b of customers or the entire queue length, which ever is less, is taken up for service. Service times are exponentially distributed.

- Server vacations are according to a Bernoulli schedule, which means that the server may take a single vacation at a service completion epoch (BS: Bernoulli schedule, V_S: single vacation). Vacation times are generally distributed.

The restricted admissibility policy has also been considered in other papers by Choudhury and Madan [3], Madan and Choudhury [5, 6, 7], and Madan and Abu-Dayyeh, [9]. None of these papers however assumes a batch service. Madan and Abu-Dayyeh [8] is the only one that assumes a batch service, however it is a Markovian system. Our aim is to generalize their model by considering the following $\mathrm{M(RA)/G}^{(r,R)}\mathrm{(V}_S\mathrm{)/1(BS)}$ queueing system:

- Service is batch of variable size according to the following bilevel policy: Given two fixed integers r (quorum size) and R (maximum server capacity), with $1 \leq r \leq R$,

  - if the queue size is less than r at a service completion epoch, then the server waits for the queue to accumulate r units before starting serving the batch of r units;
  - if the queue size at a service completion epoch is less than R but larger than r, then the server serves immediately the entire queue in a single batch;
  - if the queue size at a service completion epoch is larger than R, then the server serves immediately a group of R customers in a single batch.

  In all cases, service times are generally distributed.

- An optimal policy is obtained which prescribes the values of r and R that should be implemented in order to minimize the system expected total cost per unit of time.

It is true that the arrival process in our system is orderly rather than compound Poisson, which is a specialization rather than a generalization of Madan and Abu-Dayyeh's model. Nevertheless, we believe that it is worth investigating a system with generally instead of exponentially distributed service times, since both systems require quite different analysis approaches. Also, it is our plan to generalize in another paper the results obtained in this paper to the $\mathrm{M}^x\mathrm{(RA)/G}^{(r,R)}\mathrm{(V}_S\mathrm{)/1(BS)}$ queueing system.

In the next section we describe the model formally by introducing the notation used throughout the paper. In Section 3, we use the embedded Markov chain approach to study the discrete time parameter queueing process. In Section 4, we use semi-regenerative techniques to study the continuous time parameter queueing process. In Section 5 we derive the optimal policy and end the paper in Section 6 with a numerical example.

2 Model Description

Consider an infinite capacity queueing facility where customers arrive at a service facility according to an orderly Poisson process. According to the bilevel control policy assumed, an idle period begins when the queue drops below level r and a busy period starts as soon as the queue accumulates the same number r. However, after each service completion, the server takes a vacation with probability p and starts a new service (if r customers are present) with probability $(1-p)$. The decisions about taking a vacation after each service completion or vacation completion are independent. Also, the vacations are iid random variables whose length is independent of the length of the service times. The service times are iid random variables independent of the input process. In order to fully describe the model, we use the following notation:

Parameters:

$$\begin{cases} \theta & : \text{ probability that a customer is allowed admission upon arrival;} \\ \lambda & : \text{ arrival rate during idle period;} \\ \lambda_1 & : \text{ arrival rate during busy period, with } \lambda_1 = \theta\lambda; \\ p & : \text{ probability that the server takes a vacation at a service completion epoch;} \\ b, b_2 & : \text{ first and second moments of the service time of a batch;} \\ v, v_2 & : \text{ first and second moments of the vacation time of the server.} \end{cases}$$

Random variables:

$$\begin{cases} B & : \text{ service time of a batch;} \\ V & : \text{ vacation time of the server;} \\ G & : \text{ required service time of a batch;} \\ C_n & : \text{ number of customer arrivals during } n\text{th required service time;} \end{cases}$$

and we note that the time required by a batch of customers to complete the service cycle is such that $G = \begin{cases} B+V, & p, \\ B, & q = 1-p. \end{cases}$

Probability distribution functions (PDF):

$$\begin{cases} B(t) & : \text{ PDF of the service time } B \text{ of a batch;} \\ V(t) & : \text{ PDF of the vacation time } V \text{ of the server;} \\ G(t) & : \text{ PDF of the required service time } G \text{ of a batch.} \end{cases}$$

Laplace-Stieltjes transforms (LST):

$$\begin{cases} B^*(\theta) & : & \text{LST of } B(t) & \text{with} & \beta(z) = B^*(\lambda - \lambda z); \\ V^*(\theta) & : & \text{LST of } V(t) & \text{with} & \nu(z) = V^*(\lambda - \lambda z); \\ G^*(\theta) & : & \text{LST of } G(t) & \text{with} & \gamma(z) = G^*(\lambda - \lambda z) = [q + p\nu(z)]\beta(z). \end{cases}$$

Stochastic processes:

$$\begin{cases} Q(t) & : & \text{number of customers in the system at an arbitrary instant of time } t; \\ Q_n & : & \text{number of customers in the system at the } n\text{th service completion epoch.} \end{cases}$$

3 Discrete Time Process

The queueing process $\{Q_n; n = 0, 1, \cdots\}$ is a Markov chain since it is such that

$$Q_{n+1} = (Q_n - R)^+ + C_{n+1}, \tag{3.1}$$

where $f^+ = \max\{f, 0\}$. Denote by A the transition probability matrix (TPM) of $\{Q_n\}$ and by $A_i(z) = E\left[z^{Q_{n+1}} \,|\, Q_n = i\right]$ the probability generating function (pgf) of the ith row of A. Then, from the recursive expression (3.1), we have

$$A_i(z) = z^{(i-R)^+}\gamma(z). \tag{3.2}$$

Using results from Abolnikov and Dukhovny [1], it can be shown that the Markov chain $\{Q_n\}$ is ergodic if and only if

$$\rho < 1, \tag{3.3}$$

where

$$\rho = \frac{\lambda_1(b + pv)}{R}. \tag{3.4}$$

Introduce the pgf $P(z) = \sum_{i=0}^{\infty} p_i z^i$ where, when it exists (i.e., when $\rho < 1$), $p_i = \lim_{n\to\infty} P\{Q_n = i\}$ is the steady-state probability of state i. Since $P(z) = \sum_{i=0}^{\infty} A_i(z)p_i$, then, using expression (3.2), we have

$$P(z) = \frac{\gamma(z)\sum_{i<R}(z^R - z^i)p_i}{z^R - \gamma(z)}. \tag{3.5}$$

Using a variant of Rouché's theorem, see Abolnikov and Dukhovny [1], the R unknown probabilities $p_0, \cdots, p_{R-1}$ in (3.5) are the solution of the following system of linear equations:

$$\begin{cases} \displaystyle\sum_{i<R} \frac{d^k}{dz^k}\left[\gamma(z) - z^i\right]_{z=z_s} p_i &= 0, k = 0, 1, \cdots, k_s - 1; s = 1, 2, \cdots, S, \\ \displaystyle\sum_{i<R}(R - i)p_i &= R - \rho R, \end{cases} \tag{3.6}$$

where z_s are the roots of the characteristic equation

$$z^R - \gamma(z) = 0 \tag{3.7}$$

in the region $\bar{B}(0,1)\backslash\{1\}$ with their multiplicities k_s such that $\sum_{s=1}^{S} k_s = R - 1$.

At this stage, various performance measures related to the Markov chain $\{Q_n\}$ can be obtained. Namely, the mean system size at a service completion epoch is given by

$$L_d = \frac{N''(1) - D''(1)}{2D'(1)}, \tag{3.8}$$

where

$$\begin{aligned} N''(1) &= 2\rho R(R - \rho R) + \sum_{i<R} [R(R-1) - i(i-1)]p_i, \\ D'(1) &= R - \rho R, \\ D''(1) &= R(R-1) - \lambda_1^2 \left[b_2 + 2pbv + pv_2\right]. \end{aligned}$$

Also, let $P = (p_0, p_1, \cdots)$ and for $i = 0, 1, \cdots$, let $\beta_i = E\left[T_{n+1} - T_n \,|\, Q_n = i\right]$. Then the mean busy cycle is given by

$$P\beta = \sum_{i<r} \left(\frac{r-i}{\lambda}\right) p_i + b + pv. \tag{3.9}$$

Using the the mean busy cycle (3.9), the system intensity defined by $\mathcal{I} = \lambda \sum_{i<r} p_i\beta_i + \lambda_1 \sum_{i\geq r} p_i\beta_i$ is given by

$$\mathcal{I} = \sum_{i<r} (r-i)p_i + \rho R. \tag{3.10}$$

Finally, we note that the server load is

$$\Gamma_{n+1}(Q_n) = \begin{cases} r, & Q_n < r, \\ \min\{Q_n, R\}, & Q_n \geq r, \end{cases}$$

so that the mean server load defined by $\ell = \lim_{n\to\infty} E\left[\Gamma_{n+1}(Q_n) \,|\, Q_n = i\right]$ is given by

$$\ell = (r - R) \sum_{i<r} p_i + \sum_{r\leq i<R} (i - R)p_i + R. \tag{3.11}$$

Using expression (3.10) for the system intensity, expression (3.11) for the mean server load, and the last equation (which is due to the fact that $P(1) = 1$) in the system of equations (3.6), we can show that $\mathcal{I} = \ell$.

4 Continuous Time Process

The queueing process $\{Q(t); t \geq 0\}$ is readily seen to be semi-regenerative relative to the point process $\{T_n; n = 0, 1, \cdots\}$ and $\{(Q_n, T_n); n = 0, 1, \cdots\}$ is a Markov renewal process. Introduce the pgf $\pi(z) = \sum_{i=0}^{\infty} \pi_i z^i$ where $\pi_i = \lim_{n\to\infty} P\{Q(t) = i\}$ are the steady-state system size probabilities. This stationary distribution exists under the condition $\rho < 1$. It can be determined using the main convergence theorem for semi-regenerative processes which, provided the probability distribution of the embedded Markov chain is known, gives much quicker result than the more popular method of supplementary variables. The main convergence theorem for semi-regenerative processes, see Çinlar [4], yields

$$\pi(z) = \frac{1}{P\beta} \sum_{i=0}^{\infty} p_i \sum_{j=0}^{\infty} \int_0^{\infty} P\left(Q(t) = j, T_1 > t \,|\, Q_n = i\right) dt\, z^j, \tag{4.1}$$

where the mean busy cycle $P\beta$ is given by expression (3.9). Specializing the main convergence theorem for semi-regenerative processes to our case yields, after some lengthy computations

$$\pi(z) = \frac{1}{\lambda_1 P\beta} \left\{ \frac{1-\gamma(z)}{1-z} P(z) + [\gamma(z) + \theta - 1] \sum_{i<r} \frac{\left(z^i - z^r\right) p_i}{1-z} \right\}. \tag{4.2}$$

At this stage, various performance measures related to the semi-regenerative process $\{Q(t)\}$ can be obtained. Namely, the mean system size at an arbitrary instant of time is given by

$$L_c = \frac{1}{\lambda_1 P\beta} \left\{ \rho R L_d + \frac{\lambda_1^2}{2} \left(b_2 + 2pbv + v_2\right) + \rho R \sum_{i<r} (r-i) p_i + \frac{\theta}{2} \sum_{i<r} \left[i(i-1) - r(r-1)\right] p_i \right\}. \tag{4.3}$$

Also, the mean idle period is given by

$$I = \frac{1}{\lambda} \frac{\sum_{i<r} (r-i) p_i}{\sum_{i<r} p_i}, \tag{4.4}$$

the mean busy period by

$$B = \frac{1 - \sum_{i<r} \pi_i}{\sum_{i<r} \pi_i} I. \tag{4.5}$$

and the mean busy cycle by

$$C = I + B. \tag{4.6}$$

5 Optimal Policy

The performance measures derived in the previous section are now be used to optimize the performance of the system. The design of an optimal management policy for a queueing

system has received a lot of attention, as shown by the survey conducted by Tadj and Choudhury [10]. This is known in queueing theory as the optimal control of the system. The aim is to find the the best values that the decision maker would implement in order to minimize the total expected cost per unit of time. Using a linear cost structure, this cost is given by

$$TC(r, R) = c_h L_c + c_o \frac{B}{C} + c_s \frac{1}{C} + c_a \frac{I}{C}, \tag{5.1}$$

where

- c_h: holding cost per unit time for each customer present in the system;
- c_o: cost per unit time for keeping the server on and in operation;
- c_s: setup cost per busy cycle;
- c_a: startup cost per unit time for the preparatory work of the server before starting the service.

This expression of the total expected cost per unit of time is highly nonlinear and it is not possible to show that it is a convex function of r and R. Denoting by

$$R^*(r) = \min\{R \geq 1 \,|\, TC(r, R+1) - TC(r, R) > 0\}, \tag{5.2}$$

the following quick search procedure is reproduced from Tadj and Choudhury [10] to obtain the optimal values of r and R:

Step 1. Set $r = 1$. Determine $R^*(r)$ using (5.2) and compute $TC(r, R^*(r))$ using (5.1).

Step 2. Compute $R^*(r+1)$ using (5.2) and $TC(r+1, R^*(r+1))$ using (5.1).

Step 3. If $TC(r+1, R^*(r+1)) > TC(r, R^*(r))$, STOP. The optimal values are $(r^*, R^*) = (r, R^*(r))$. Otherwise, set $r = r + 1$, GOTO Step 2.

6 Numerical Illustration

To illustrate the application of the optimal policy derived in the previous section we present a numerical example in which we assume that the service and vacation times are exponentially distributed. Then

$$\beta(z) = \frac{1}{1 + \lambda b(1-z)}, \quad \nu(z) = \frac{1}{1 + \lambda v(1-z)}, \quad \text{and} \quad \gamma(z) = \frac{1 + q\lambda v(1-z)}{[1 + \lambda v(1-z)][1 + \lambda b(1-z)]}.$$

Also, the second moments of the service and vacation times are $b_2 = 2b^2$ and $v_2 = 2v^2$, respectively. The characteristic equation (3.7) becomes in this case

$$\lambda^2 bv z^{R+1} - [\lambda^2 bv + \lambda(b+v)]z^R + z^{R-1} + z^{R-2} + \cdots + z + 1 + q\lambda v = 0. \qquad (6.1)$$

It is well known, see for example Chaudhry and Templeton [2], that this equation has $R-1$ simple roots $z_s, s = 1, \cdots, R-1$ that belong to the unit ball $\bar{B}(0,1)$ in $\mathbb{C}$. The system of equations (3.6) becomes:

$$\begin{cases} \sum_{i<R} \left[\frac{1+q\lambda v(1-z)}{[1+\lambda v(1-z)][1+\lambda b(1-z)]} - z^i \right]_{z=z_s} p_i &= 0, \quad s = 1, \cdots, R-1, \\ \sum_{i<R} (R-i)p_i &= R - \rho R. \end{cases} \qquad (6.2)$$

To form the expected total cost per unit time (5.1), we successively compute $\pi_0, \cdots, \pi_{r-1}$ from (4.2), L_d (3.8), L_c (4.3), $\bar{I}$ (4.4), $\bar{B}$ (4.5), and $\bar{C}$ (4.6). We implemented the search procedure described above by taking the following system parameters $\lambda = 1, b = 0.6, v = 0.2, p = 0.75, \theta = 0.9$ and units costs $c_h = 5, c_o = 100, c_s = 10000, c_a = 250$. The curve representing the total expected cost per unit of time is shown in Figure 1.

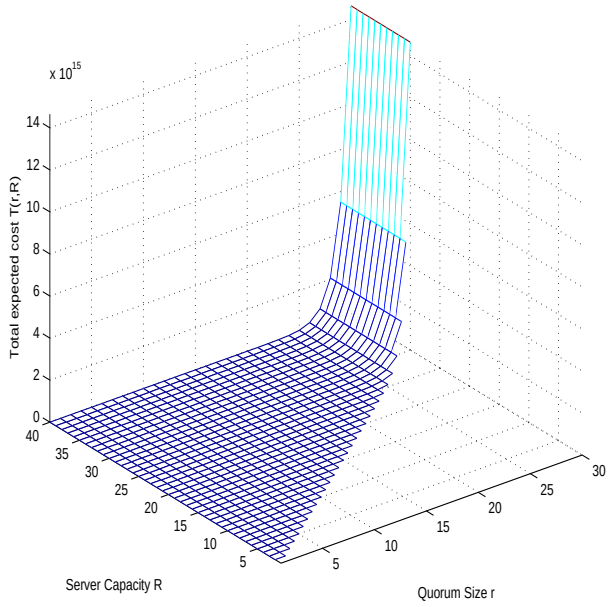

Figure 1. Variations of the total expected cost per unit of time.

The optimal values are $r^* = 1$ and $N^* = 5$ for a minimum total cost $TC(r^*, N^*) = 8691.60$. Since the restricted admissibility policy is characterized by the admission probability θ, we wanted to show the effect of this parameter on the optimal solution. To this end, we kept all the parameters at the same values, except for the probability θ that we varied from 0.1

to 0.9 by increments of 0.1, and computed the value of the total cost for each value of θ. The variations of the the total cost as a function of θ is depicted in Figure 2.

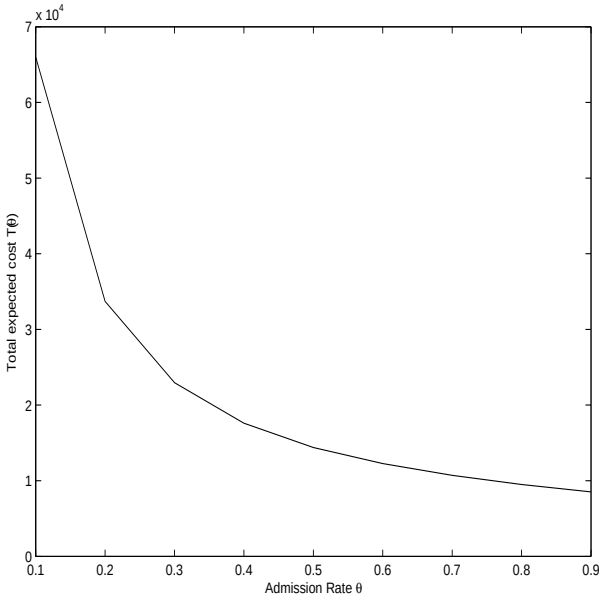

Figure 2. Variations of the total expected cost per unit of time.

As can be seen, the higher the probability of admission, the lower the total cost.

Acknowledgement This research was supported by the College of Science Research Center at King Saud University under project No. Stat/2005/30.

References

[1] L. Abolnikov and A. Dukhovny, Markov chains with transition delta matrix: Ergodicity conditions, invariant probability measures and applications, *Journal of Applied Mathematics and Stochastic Analysis* **4**:4 (1991) 333-356.

[2] M.L. Chaudhry, M.L. and J.G.C. Templeton, *A First Course in Bulk Queues*, John Wiley, New York, 1983.

[3] G. Choudhury and K.C. Madan, A batch arrival Bernoulli vacation queue with a random setup time under restricted admissibility policy, *International Journal of Operational Research* **2**:1 (2007) 81-97.

[4] E. Çinlar, *Introduction to Stochastic Processes*, Prentice Hall, New York, 1975.

[5] K.C. Madan and G. Choudhury, Steady state analysis of an $M^X/(G_1,G_2)/1$ queue with restricted admissibility and random setup time, *Information and Management Sciences* **17**:2 (2006) 33-56.

[6] K.C. Madan and G. Choudhury, An $M^X/G/1$ queue with a Bernoulli vacation schedule under restricted admissibility policy, *Sankhyā* **66**:1 (2004) 175-193.

[7] K.C. Madan and G. Choudhury, Steady State analysis of an $M^X/(G_1,G_2)/1$ queue with restricted admissibility of arriving batches and modified Bernoulli schedule server vacations, *Journal of Probability and Statistical Science* (JPSS) **2**:2 (2004) 167-186

[8] K.C. Madan and W. Abu-Dayyeh, Steady-state analysis of a single server bulk queue with general vacation times and restricted admissibility of arriving batches, *Revista Investigacion Operacional* **24**:2 (2003) 113-123.

[9] K.C. Madan and W. Abu-Dayyeh, Restricted admissibility of batches into an M/G/1 type bulk queue with modified Bernoulli schedule server vacations, *European Series in Applied and Industrial Mathematics: Probability and Statistics* (ESAIM: P&S), **6** (2002) 113-125.

[10] L. Tadj and G. Choudhury, Optimal design and control of queues, *TOP*, **13**:1 (2005) 359-414.