OPTIMALITY OF STATE DEPENDENT SINGLE SERVER QUEUE WITH MMPP INPUT

R. SAKTHI¹, R. RAGHAVENDRAN, AND V. VIDIHYA

Abstract. In this paper, we study the optimal control of service rate in a single server state dependent queue with MMPP input. The objective of this work is find the optimal service rate that minimizes the service rate using matrix method. A performance difference formula is derived to quantify the difference of service rate based on varying arrival in the long run.

1. Introduction

Queuing systems is a mathematical phenomenon that has wide range of applications in computer systems and communication networks. It characterizes the system as the arrival, service mechanism, the number of servers and service discipline with finite or infinite service. The customer (packets) arrive to the system and wait in the buffer to be serviced, if the server fails to provide service immediately where there are many packets to be transferred that may suffer a long delay which lead to poor performance. This situation can be represented by performance of queuing system with waiting length dependent on increased arrival or poor services. If the waiting length exceeds a threshold value the queue may be overloaded so that in order to reduce the queue length, the service rate may be increased. In this paper we study a state dependent service queuing system in

¹corresponding author

2010 Mathematics Subject Classification. 60K25, 68M20, 90B22.

Key words and phrases. Markov Modulated Poisson process (MMPP), quasi-birth-death, performance analysis.
which the customer arrival follows Poisson process, and the system has only one server to provide the service to the arriving customer.

Queueing models with change of arrival rate are widely used in the telecommunications industries to study congestion problems related to internet and mobile networks. Among some of the methods, one method for capturing the variation behavior of the arrival process is via Markov-modulated Poisson processes (MMPPs). In the stochastic processes theory, the counting process has gained its importance in the field such as science, engineering etc. A number of counting processes have been introduced in order to capture the characteristics of the actual stochastic processes, some of the well-known counting processes are Poisson processes, compound Poisson processes, Markov Modulated Poisson processes (MMPP), renewal processes, and semi-Markov processes. Markovian arrival processes are considered as generalizations of Poisson processes, compound Poisson processes and Markov modulated Poisson processes. Recently, a significant amount of research effort is being put to the optimization of queueing systems, whereby considering the joint study of the performance analysis and the performance optimization [1], [4]. Service rate control is one of the dominant research topic in the performance optimization of queueing systems and it has attracted a lot of attention from researchers and practitioners [2], [3], [5], [6]. A system with service rate depends on workload have been studied in [9] and [10]. For results on computing the steady state distribution using matrix geometric solution of state dependent queue, one can refer to the paper [11] in which the various performance measures of $M/E_k/1$ has been investigated.

2. Model Description

In this paper, we study the model $MMPP/M/1$ queue with infinite capacity. A $MMPP$ is defined by two matrices $D_0$ and $D_1$ with order $m \times m$, where $m$ denotes the states of the $MMPP$. Let $D_0 + D_1$ represent the infinitesimal generator which is irreducible under the underlying Continuous Time Markov Chain (CTMC). The matrices $D_0$ and $D_1$ are given by

$$D_0 = \begin{bmatrix} -c_{11} & c_{12} \\ c_{21} & -c_{22} \end{bmatrix}, \quad D_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$c_{11} = \lambda_1 + c_{11}$, $c_{22} = \lambda_2 + c_{21}$. The steady state average arrival of $MMPP$ is
\[ \lambda = \pi D_1 e, \] where \( \pi \) is a row vector representing the steady state distribution of phases satisfying \( \pi D = 0 \) and \( \pi e = 1 \), and \( e \) is a column vector with all entries equal to 1. A customer that arrives to the system is in a random environmental state \( i = 1, 2, \ldots, m \). The service discipline is first come first served.

The service is provided by a single server with different service rates as its state dependent. The service rate depends on the switch queue length according to \( \lambda_1 \) and \( \lambda_2 \). The service rate is denoted as \( \mu_{n,j} \), where \( n \) is the number of customer and \( j \) is the state of MMPP queue, \( n = 0, 1, 2, \ldots, j = 1, 2, \ldots, m \) with \( \mu_{n,j} = 0 \) for \( n = 0 \). The service rate \( \mu_{n,j} \) will modulate according to the value of \( n \) and the arrival rate. As the exponential service rate satisfies memoryless property, it is admissible that if \( n \) increases to \( n+1 \) or \( j \) to \( j' \) then the server will change its rate \( \mu_{n+1,j} \) or \( \mu_{n,j'} \) immediately, otherwise if \( n \) decreases to \( n-1 \) the server will change its service rate \( \mu_{n-1,j} \). The state of the system at time \( t \) is described as bivariate process

\[ \{(N(t), J(t)) : t \geq 0 \}, \]

where \( N(t) \) the number of customers in the system, \( J(t) \) denotes the state of the MMPP at time \( t \). The bivariate process \( \{(N(t), J(t)) : t \geq 0 \} \) is a QBD process on the state phase \( S = Z_1 \times \{1, 2, \ldots, m\} \). The infinitesimal generator of the QBD is given by

\[
Q = \begin{pmatrix} B_{00} & F_0 \\ B_{10} & L_1 & F_1 \\ & B_1 & L_1 & F_1 \\ & & \ddots & \ddots & \ddots \\ & & & B_i & L_i & F_i \\ & & & & \ddots & \ddots & \ddots \\ & & & & & \ddots & \ddots & \ddots & \ddots \\ & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}
\]

where \( B_{00} = D_0, B_{10} = d(\mu_1), B_i = d(\mu_i), L_i = D_0 - d(\mu_i), F_i = D_1 \) for \( i = 1, 2, \ldots, m \). The transition rate matrix of this particular QBD process is given by

\[
Q = \begin{pmatrix} D_0 & D_1 & 0 & 0 & 0 & \ldots \\ d(\mu_1) & D_0 - d(\mu_1) & D_1 & 0 & 0 & \ldots \\ 0 & d(\mu_2) & D_0 - d(\mu_1) & D_1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}.
\]

For \( m = 2 \), all the service rate with equal queue length \( n \) consists an equal 2–dimensional column vector as follows \( \mu_n = (\mu_{n,1}, \mu_{n,2})^T \) and the \( 2 \times 2 \) diagonal
matrix is defined as \( \mathbf{d}(\mu_n) = \begin{pmatrix} \mu_{n,1} & 0 \\ 0 & \mu_{n,2} \end{pmatrix} \). For the equal values of the diagonal matrix \( \mathbf{d}(\mu_n) \) for some \( n > N_0 \) the matrix analytic method can be efficiently applied to compute the steady state distribution of \( \text{MMPP}/M/1 \) queue. As the service rate are different, the performance optimization of the \( \text{MMPP}/M/1 \) queue are obtained with the different approach of using matrix analytic method combined with sensitivity optimization approach. Let \( h_n,j \) be the holding cost per unit time in the state \( j \) with \( n \) customers. The cost function is given by

\[
c_{n,j} = h_{n,j} + \theta \mu_{n,j},
\]

where \( \theta > 0 \) is the load to balance the cost of the operating and holding. The service rate \( \mu_{n,j} \) is flexible in its domain \( M_{n,j} = [\mu_{n,j}^-, \mu_{n,j}^+] \), \( n = 1, 2, \ldots, j = 1, 2, \ldots, m \). The service rate which are different are listed in the matrix form as \( \mu = (\mu_1, \mu_2, \ldots) \) where \( \mu_n \) is an element in

\[
M_n = M_{n,1} \times M_{n,2} \times \ldots M_{n,m},
\]

where the constant values \( \mu_{n,j}^-; \mu_{n,j}^+ \) are the minimum and maximum values. The aim of this work is to compute the optimal service rate \( \mu^* \) that minimize the mean total cost i.e \( \mu^*_n = (\mu_{n,1}^+, \mu_{n,2}^+, \ldots, \mu_{n,m}^+) \), \( n \geq 1 \), \( \mu^* = (\mu_1^+, \mu_2^+, \ldots) \). In the long run the mean total cost of the system is defined as

\[
C = \lim_{T \to \infty} \frac{1}{T} E \left\{ \int_0^T c((N(t), J(t))) dt \right\} = \pi c.
\]

Then the potential function of the CTMC is defined as

\[
\mu^* = \arg\min_{\mu_{n,j}^- \leq \mu_{n,j} \leq \mu_{n,j}^+} \pi c
\]

\[
\phi(n, j) = \lim_{T \to \infty} \frac{1}{T} E \left\{ \int_0^T [c((N(t), J(t))) - \delta] dt / (N(0), J(0)) \right\}.
\]

Also, we define a column vector \( \Phi = (n, j) \) that consist of an element of \( \phi(n, j) \), \( n = 0, 1, 2, \ldots \) and \( j = 1, 2, \ldots, m \) as

\[
\Phi(n) = (\phi(n, 1), \phi(n, 2), \ldots, \phi(n, m))^T, \text{for } n = 0, 1, 2, \ldots
\]

\[
\Phi = (\Phi^T(0), \Phi^T(1), \ldots).
\]

Assumption in this model is about performance optimization of state dependent queue i.e the performance measures are studied when service rate is changed from
increased to decreased state or decreased to increased state. To study this relative function in the Markov decision theory [7] is adapted and the difference in cost function for MAP arrival obtained by [8] is given by

$$C' - C = \pi'[Q' - Q] \cdot \phi + h' - h$$

where

$$Q' - Q = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ \end{pmatrix}$$

$$= \sum_{n=1}^{\infty} \pi'(n) [d(\mu'_n - \mu_n)[\phi(n-1) - \phi(n)] + \theta(\mu'_n - \mu_n)]$$

$$= \sum_{n=1}^{\infty} \sum_{j=1}^{m} \pi'(n, j)(\mu'_{n,j} - \mu_{n,j})[\phi(n-1, j) - \phi(n, j) + \theta]$$

$$(2.1)$$

$$= \sum_{n=1}^{\infty} \sum_{j=1}^{m} (\mu'_{n,j} - \mu_{n,j}) \pi'(n, j) \Phi(n, j)$$

where $\Phi(n, j) = \phi(n-1, j) - \phi(n, j) + \theta$. The necessary and sufficient condition of the optimal service rate $\mu^*_{n,j}$ stated in the following lemma that has been proved in [8].

Lemma 2.1. The service rate $\mu^*_{n,j}$ is the optimal service rate iff

$$\mu^*_{n,j} \Phi^*(n, j) \leq \mu'_{n,j} \Phi^*(n, j), \forall \mu'_{n,j} \in [\mu_{n,j}^-, \mu_{n,j}^+], n = 1, 2, ..., j = 1, 2, ..., m.$$ 

Theorem 2.1. In the long run, the total mean cost $C$ is monotone with respect to the service rate $\mu_{n,j}$, for all $n = 1, 2, ...$ and $j = 1, 2, ..., m$.

Proof. Without any loss of generality assume that the service rate $\mu_{r,s}$ at the state $(r, s)$ is not transient. Let the current service rate be $\mu$. Consider the change in service rates $\mu_{r,s}$ to $\mu'_{r,s}$ and keep the other service rate fixed as $\mu_{n,s}$. With this change the total mean cost in the long run is given by

$$C' - C = (\mu'_{r,s} - \mu_{r,s}) \pi'(r, s) \Phi(r, s).$$

(2.2)

Now the same situation considered above is treated as vice versa. Suppose the current service rate is $\mu'$ and change the service rate $\mu'_{r,s}$ to $\mu_{r,s}$ and maintain the other service rate fixed. So the total mean cost in the long run is given by

$$C - C' = (\mu_{r,s} - \mu'_{r,s}) \pi(r, s) \Phi'(r, s).$$

(2.3)
Comparing (2.2) and (2.3), we get
\[ \frac{\Phi'(r,s)}{\Phi(r,s)} = \frac{\pi'(r,s)}{\pi(r,s)} > 0. \]
Hence \( \Phi(r,s) \) is fixed when \( \mu_{rs} \) is changed. So the equation (2.1) can be written as
\[ \Delta h = \pi(r,s)\Phi(r,s). \]
As the sign of this equation is not changed, \( h \) is monotone with respect to \( \mu_{rs} \) and hence the theorem is proved. \hfill \Box

**Lemma 2.2.** For a positive integer \( n \), if
\( \text{(i) } \mu_n \text{ is non decreasing in } n \)
\( \text{(ii) } \lambda_1 + \lambda_2 < \gamma \mu_n \text{ for } n \geq N \)
then \( \Phi(n) \) can be obtained as, for \( n \geq N \)
\[ \Phi(n + 1) = \beta(n + 1) + B_{n+1}d(\mu_n)\Phi(n), \]
where \( B_{n+1} \) is given by
\[ B_{n+1} = (d(\mu_{n+1}) - D_0 - D_1B_{n+2}d(\mu_{n+1}))^{-1}, \]
\[ \beta(n + 1) = B_{n+1}(-(h(n + 1) - h(n)) + D_1\beta(n + 2)). \]

**Proof.** Consider the decomposition of \( \Phi(n + 1) \) into two points \( \beta(n + 1) \) – the total expected cost during the transitions from level \( n + 1 - N \) to level \( n - N \) and \( B_{n+1} \) in the total expected cost during the transition from level \( n - N \) to the absorption state. So \( \Phi(n + 1) \) is given as
\[ \Phi(n + 1) = \beta(n + 1) + \frac{\Phi'}{\Phi(n)} \]
where the order of \( \frac{\Phi'}{\Phi} \) is \( m \times m \) with \( (i, j) \)th position values is the probability of \( Q \) reaches \( n - N \) for the first time at phase \( j \), given that the Markov Chain was in level \( n + 1 - N \) at phase \( j \). For the above assumption, we get
\[ \frac{\Phi'}{\Phi} = (d(\mu_{n+1}) - D_0 - D_1B_{n+2}d(\mu_{n+1})). \]
Let \( B_{n+1} = \frac{\Phi'}{\Phi} \), as \( \frac{\Phi'}{\Phi} \leq e \) for \( n \geq N \),
\[ 0 \leq \mu_{n+1} \leq (d(\mu_{n+1}) - D_0 - D_1B_{n+2}d(\mu_{n+1}))e \]
which concludes that inverse exists and matrix \( \Phi' \) can be computed recursively using (2.5). Hence \( \beta(n + 1) \) and \( \beta(n + 2) \) in (2.5) are obtained as
\[ \beta(n + 1) = (d(\mu_{n+1}) - D_0)^{-1}(-\Delta h + D_1(\beta(n + 2) + B_{n+2}\beta(n + 1))). \]
\hfill \Box
3. Computation and Optimality of MMPP/M/1

The optimality of the MMPP/M/1 is discussed using recursive procedure. For a given $\mu$, consider the following:

(i) $\mu_n = \mu^+ = \mu^+_{\infty}, n \geq N$
(ii) $\Delta h(n) = \Delta h_\infty, n \geq N$
(iii) for $1 \leq n \leq N, j = 1, 2, ..., m$

$$\mu_{n,j} = \begin{cases} 0, & \text{if } \Phi(n,j) \geq 0 \\ \mu^+_n, & \text{otherwise} \end{cases}$$

(iv) for $n \geq N, B_n$ is denoted as $B_\infty$.

**Recursive algorithm to compute the optimality of MMPP/M/1**

(i) Input the values of $m, D_0, D_1, \mu^+, \mu', C$ to get $N, \mu^+_{\infty}$ and $\Delta C_\infty$;
(ii) Compute $\overline{B}_\infty, B_\infty$ and $\beta_\infty$ stated in Lemma 2;
(iii) Repeat:
   - Let $\mu' \rightarrow \mu$;
   - Compute $B_n: n = N, N - 1, ..., 2, 1$ and $\beta(n), n = N, N - 1, ..., 2, 1$ using (2.5) and (2.6)
   - Assume $\Phi(1) = \beta(1)$ and compute $\Phi_n, n = 2, ..., N$ using (2.4);
(iv) Output the value $\mu$ as the optimum $\mu^*$.

4. Conclusion

In this work, we study the service rate control of MMPP/M/1 queue. The optimality of the service rate can be computed by recursive algorithm developed to compute the optimal policy. For future work, we may extend the ideas discussed so that it can be experimented to obtain the computational values.

**References**


Department of Humanities and Sciences
Rajalakshmi Engineering College
Chennai, India
E-mail address: rsakth@gmail.com

Department of Science and Humanities
Saveetha School of Engineering, SIMATS
Chennai, India
E-mail address: raghavendran@saveetha.com

School of Advanced sciences
Vellore Institute of Technology
Chennai, India
E-mail address: vidhya.v@vit.ac.in