Optimal Control of State-Dependent Service Rates in a MAP/M/1 Queue

Li Xia, *Senior Member, IEEE*, Qi-Ming He, and Attahiru Sule Alfa

***Abstract*—In this paper, we study the optimal control of service rates in a queueing system with a Markovian arrival process (MAP) and exponential service times. The service rate is allowed to be state dependent, i.e., we can adjust the service rate according to the queue length and the phase of the MAP. The cost function consists of holding cost and operating cost. The goal is to find the optimal service rates that minimize the long-run average total cost. To achieve that, we use the matrix-analytic methods (MAM) together with the sensitivity-based optimization (SBO) theory. A performance difference formula is derived, which can quantify the difference of the long-run average total cost under any two different settings of service rates. Based on the difference formula, we show that the long-run average total cost is monotone in the service rate and the optimal control is a bang–bang control. We also show that, under some mild conditions, the optimal control policy of service rates is of a quasi-threshold-type. By utilizing the MAM theory, we propose a recursive algorithm to compute the value function related quantities. An iterative algorithm to efficiently find the optimal policy, which is similar to policy iteration, is proposed based on the SBO theory. We further study some special cases of the problem, such as the optimality of the threshold-type policy for the M/M/1 queue. Finally, a number of numerical examples are presented to demonstrate the main results and explore the impact of the phase of the MAP on the optimization in the MAP/M/1 queue.**

***Index Terms*—Markov decision process (MDP), matrix-analytic methods (MAM), queueing system, sensitivity-based optimization (SBO), service rate control.**

Manuscript received July 1, 2015; revised February 10, 2016 and September 4, 2016; accepted February 9, 2017. Date of publication March 7, 2017; date of current version September 25, 2017. This work was supported in part by the National Key Research and Development Program of China (2016YFB0901902), in part by the National Natural Science Foundation of China under Grant 61573206, Grant 61203039, and Grant U1301254, and in part by the Suzhou-Tsinghua Innovation Leading Action Project. Recommended by Associate Editor C. M. Lagoa.

L. Xia is with the Center for Intelligent and Networked Systems, Department of Automation, TNList, Tsinghua University, Beijing 100084, China (e-mail: xial@tsinghua.edu.cn).

Q.-M. He is with the Department of Management Sciences, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: q7he@uwaterloo.ca).

A. S. Alfa is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada, and also with the Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria 0002, South Africa (e-mail: attahiru.alfa@umanitoba.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2017.2679139

I. Introduction

TRADITIONALLY, queueing theorists have focused on modeling queueing systems and analyzing performance measures. Recently, a considerable amount of research effort is being devoted to the optimization of queueing systems, whereby the study of the performance analysis and the performance optimization are being considered jointly. Service rate control is one of the most important research topics in the performance optimization of queueing systems and it has attracted a lot of attention from researchers and practitioners [8], [12], [22], [24], [25].

Different studies on the service rate control problem focus on different queueing models. For an M/M/1 queue, it is proved that the optimal service rate is nondecreasing with respect to (w.r.t.) the number of customers in the system when the holding cost is nondecreasing and the criterion is to minimize the long-run average cost [24]. A recent study shows that such a monotone structure holds for systems with state-dependent service rate and/or state-dependent arrival rate [3]. The effect of pricing on the optimization result is also considered in [3]. For a tandem queue with two servers, it is proved that the optimal state-dependent service rate has a structure of the bang–bang control determined by a switch function [21]. The monotone structure of optimal service rates for a cyclic queue, where the service rate is assumed to be independent of the state, is studied in [26]. It is proved in [15] and [31] that for a closed Jackson network, the optimal service rates are located on the boundary of their value domains (bang–bang control) when the cost function is linear in the service rate. Recently, the above property is further extended under a more general condition where the cost function is only required to be concave w.r.t. the service rate [30]. There are also some works studying the service rate control problem from a game theoretic perspective [14], [17], [28].

Most of the existing works assume that the arrival process is a stationary Poisson process. However, many practical service systems do not satisfy this assumption. It is very common to see that the customer arrival process is neither Poisson nor stationary, e.g., Internet traffic with bursty and self-similarity properties [23]. Therefore, it is useful to study the service rate control problem, where the arrival process is not Poisson. Recent work [13] studies the dynamic service rate control problem of a queue, where the arrival process is a Markov modulated Poisson process (MMPP) and the convexity and/or monotonicity assumptions on cost functions are assumed. In this paper, we study the service rate control problem in a MAP/M/1 queue, which is a very general scenario compared with the work in the literature. The assumptions on cost functions are also significantly weaker. By utilizing structural properties of queues, we develop an efficient algorithm for computing the optimal policy.

Markovian arrival process (MAP) is a very general stochastic model for customer arrivals in queueing systems [2]. It can model almost all arrival processes, such as Poisson process, MMPP, phase-type (PH) renewal process, etc. The coefficient of variation of a MAP can be any positive number. Thus, to a certain extent, a MAP can model the arrival process with bursty, self-similarity, or heavy-tail property. It is therefore interesting to study the service rate control of queueing systems with MAP arrival, which can be used to model many practical problems. However, MAP arrival increases the model complexity since we have to consider the MAP's phase status in the analysis. Matrix-analytic methods (MAM) are a set of well-known techniques that can handle such complexity. MAM uses a block matrix structure to organize the transition rate matrix of queueing models and proposes efficient computation algorithms to compute the steady-state performance measures [1], [9], [18], [19]. Utilizing the MAM theory can not only simplify the representation of such queueing models, but also provide efficient computation algorithms for performance analysis. On the other hand, the sensitivity-based optimization (SBO) theory is a new tool proposed for the performance optimization of Markov systems [6], [7]. The key result is the performance difference formula that can quantify the performance difference of Markov systems under any two policies or parameters. The difference formula gives a clear relation that explains how the system performance is varying according to different policies or parameters. The SBO theory is especially effective for the performance optimization of queueing systems because it can efficiently utilize the structure properties of queueing models. Recently, the SBO theory has been applied to the optimization of queueing systems successfully [28]–[30], and some interesting results that are difficult to obtain in classical queueing theory have been derived. But these works focus on the Jackson network and cannot handle the non-Poisson arrival process. Therefore, it is interesting to apply the SBO theory to study the service rate control in the MAP/M/1 queue.

In this paper, we combine MAM and SBO theories to study the service rate control in the MAP/M/1 queue. We use MAM techniques to organize the transition rate matrix in a quasi-birth-death (QBD) form. Based on the concise form of matrix organization, we apply the SBO theory and derive a difference formula under different service rates. With the difference formula, some interesting and insightful results are obtained, such as a necessary and sufficient condition of optimal service rates, a monotonicity property of the system performance w.r.t. the service rate, and an optimality of quasi-threshold-type policy. By utilizing the MAM theory, we propose an efficient computation algorithm to recursively compute the value function related quantities. We propose an efficient optimization algorithm, based on the SBO theory, to iteratively find the optimal policy. Some further optimality properties are also obtained in special cases. Finally, we conduct numerical experiments to verify the main results and explore the impact of the phase of the MAP on the control of the service rate.

The rest of the paper is organized as follows. In Section II, we give a mathematical formulation of the service rate control problem in the MAP/M/1 queue, where we utilize MAM techniques to represent the QBD structure of transition rate matrix. In Section III, we apply the SBO theory and derive the difference formula for the optimization problem. Some optimality properties are also derived in this section. We further characterize the optimal control policy and develop efficient algorithms for searching the optimal policy in Section IV. In Section V, we conduct numerical experiments to gain insights into the impact of the phase of the customer arrival process on the optimization result. Finally, we summarize the results obtained in this paper and provide some brief discussions about possible future work in Section VI.

II. Problem Formulation

We consider a MAP/M/1 queue. Customer arrivals obey a MAP that is a very general arrival process [18]. A MAP is defined by two m-by-m matrices $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$, where m is the number of phases (states) of the MAP. The MAP can also be viewed as a class of hidden Markov models, where the customer arrivals depend on the phase of an underlying continuous-time Markov chain (CTMC). Let $\boldsymbol{D} = \boldsymbol{D}_0 + \boldsymbol{D}_1$ be an irreducible infinitesimal generator of the underlying CTMC of the MAP. The phase transitions of the underlying CTMC are classified into two types, one is a hidden transition that does not lead to a customer arrival and the other is an observable transition that leads to a customer arrival. $D_0(i, j)$ is the rate of the hidden transition from phase i to phase j, whereas $D_1(i, j)$ is the rate of the observable transition from phase i to phase $j, i, j = 1, 2, \ldots, m$. Both types of transitions can change the current phase of the MAP. The hidden transition can change the rate of customer arrivals and it has physical meanings, such as different arrival modes or breakdowns of customer generating sources. The diagonal elements of $\boldsymbol{D}_0$ are negative. All the other elements of $\boldsymbol{D}_0$ and all the elements of $\boldsymbol{D}_1$ are nonnegative. Below, we give an example of $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ for a two-phase MAP:

$$\boldsymbol{D}_0 = \begin{bmatrix} -\sigma_1 & \gamma_{1,2} \\ \gamma_{2,1} & -\sigma_2 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} \\ \lambda_{2,1} & \lambda_{2,2} \end{bmatrix}$$

where $\gamma_{i,j} \geq 0$ and $\lambda_{i,j} \geq 0$ for all i, j, $\sigma_1 = \gamma_{1,2} + \lambda_{1,1} + \lambda_{1,2} > 0$, $\sigma_2 = \gamma_{2,1} + \lambda_{2,1} + \lambda_{2,2} > 0$. It is well known that the steady-state average arrival rate of a MAP is $\lambda = \boldsymbol{\omega}\boldsymbol{D}_1\boldsymbol{e}$, where $\boldsymbol{\omega}$ is a row vector representing the steady-state distribution of phases and satisfying $\boldsymbol{\omega}\boldsymbol{D} = \boldsymbol{0}$ and $\boldsymbol{\omega}\boldsymbol{e} = 1$, and $\boldsymbol{e}$ is a proper dimension column vector with all the elements as 1.

A customer that arrives to find the server busy will wait in the waiting room. We assume that the waiting room capacity is infinite. The service discipline is first-come-first-served. The service is provided by a single server. The service time follows an exponential distribution. The service rate is state dependent and it depends on the queue length at the server and the phase status of the customer arrival process. We denote the service rate as $\mu_{n,j}$, where n is the number of customers at the server and j

is the phase status of the MAP, $n = 0, 1, 2, \ldots, j = 1, 2, \ldots, m$. Obviously, we have $\mu_{n,j} = 0$ if $n = 0$. The service rate $\mu_{n,j}$ will change according to the value of n at the server and j of the arrival process. Suppose the server currently works at rate $\mu_{n,j}$. If n increases to $n+1$ (i.e. a new customer arrives) or the phase status changes to j', then the server will, respectively, change its rate to $\mu_{n+1,j}$ or $\mu_{n,j'}$ immediately. If n decreases to $n-1$ (i.e., a servicecompletion occurs), the server will change its rate to $\mu_{n-1,j}$. Such a scheme is reasonable since the exponential service time distribution has the memoryless property.[1]

The system state at time t is defined as $(N(t), J(t))$, where $N(t)$ is the number of customers at the server at time t and $J(t)$ is the phase of the MAP at time t, for $t \geq 0$. For a specific state, we denote it as (n, j), where $n = 0, 1, \ldots$ and $j = 1, 2, \ldots, m$. All the possible states compose the state space $\mathcal{X} := \{\text{all } (n, j) | n = 0, 1, \ldots \text{ and } j = 1, 2, \ldots, m\}$. Obviously, the state space $\mathcal{X}$ is infinite. All the service rates with the same queue length n consist an m-dimensional column vector as follows:

$$\boldsymbol{\mu}_n = (\mu_{n,1}, \mu_{n,2}, \ldots, \mu_{n,m})^T \tag{1}$$

where the superscript T indicates the transpose operation of matrix or vector. We further define an $m \times m$ diagonal matrix as

$$\text{diag}(\boldsymbol{\mu}_n) = \begin{pmatrix} \mu_{n,1} & 0 & 0 & \cdots & 0 \\ 0 & \mu_{n,2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \mu_{n,m} \end{pmatrix}. \tag{2}$$

It is easy to see that $\{(N(t), J(t)), t \geq 0\}$ is a QBD process. The transition rate matrix of this QBD process is given as (3) shown in the bottom of the page.

In the matrix (3), the system state (n, j) is sorted first according to the order of n, then according to the order of j. That is, the first row corresponds to the transition rate from the initial state $(0, 1)$ to the next state (n, j), the second one from $(0, 2)$ to (n, j), until the mth one from $(0, m)$ to (n, j), the $(m+1)$ st row from $(1, 1)$ to (n, j), etc., where $j = 1, 2, \ldots, m$ and $n = 0, 1, 2, \ldots$. From the QBD structure of (3), we know that if the diagonal matrix $\text{diag}(\boldsymbol{\mu}_n)$ is the same for all $n > n_0$ (where n_0 is a given constant), we can use the standard MAM to efficiently compute the steady-state distribution of the MAP/M/1 queue [1], [9], [18]. However, in this paper, we aim to study the performance optimization of the MAP/M/1 under different settings of service rates, i.e., the values of $\text{diag}(\boldsymbol{\mu}_n)$ are generally different. Therefore, the MAM technique cannot be used directly for this purpose. We combine the MAM theory with the SBO theory [6], [7] to study this optimization problem.

The system cost consists of two parts, holding cost and operating cost. Let $\phi(n, j)$ be the holding cost per unit time, given the queue state n and the phase j. Also, let $\psi(\mu_{n,j})$ be the operating cost per unit time, which is defined as a function of the service rate, given the queue state n and the phase j. In this paper, we consider a specific form of the operating cost, $\psi(\mu_{n,j}) = b\mu_{n,j}$, so the cost function is defined as

$$f(n, j) = \phi(n, j) + b\mu_{n,j}, \quad n = 0, 1, \ldots, \; j = 1, 2, \ldots, m \tag{4}$$

where $b \; (> 0)$ is a weight to balance the prices of the operating cost and the holding cost. In the above cost function, $\phi(\cdot)$ has no relation to service rates. For the characterization of the optimal policy in this paper, in general, we do not have any requirement on $\phi(\cdot)$, such as the convexity (concavity) or nondecreasing (nonincreasing) conditions widely used in the literature [24]. However, we do impose mild conditions on $\phi(n, j)$ when n is big, which can be found later in this paper (see Theorem 3). Note that we assume a linear operating cost in this paper. Although this assumption is reasonable in the real world, there also exist many other cases where the operating cost is nonlinear in service rates, such as the exponential structure of power consumption w.r.t. transmission rates in wireless communication. It deserves further investigation for the case of nonlinear operating cost.

The service rate $\mu_{n,j}$ is adjustable and its value domain is denoted as $U_{n,j}$, $n = 1, 2, \ldots, j = 1, 2, \ldots, m$. All the service rates are organized in a matrix form as

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \ldots) \tag{5}$$

which has m rows and infinitely many columns. We shall call $\boldsymbol{\mu}$ a policy from the terminology of Markov decision process (MDP). The value domain of $\boldsymbol{\mu}_n$ is denoted as

$$\mathcal{U}_n = U_{n,1} \times U_{n,2} \times \cdots \times U_{n,m}, \tag{6}$$

where $\times$ is the Cartesian product and $n = 1, 2, \ldots$. The value domain of $\boldsymbol{\mu}$ is denoted as

$$\mathcal{U} = \mathcal{U}_1 \times \mathcal{U}_2 \times \cdots \tag{7}$$

where $\mathcal{U}$ is also called the policy space in MDP. We observe that $\boldsymbol{\mu}$ is a parametric policy with infinite dimension, so is the policy space $\mathcal{U}$. In this paper, we assume that the value domain $U_{n,j}$ is

[1]The phase status j may be unobservable in practice, which makes the phase-dependent service rate $\mu_{n,j}$ difficult to implement. We give more discussion in Examples 5 and 6 in Section V, where a sampling policy is proposed to approximate $\mu_{n,j}$.

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{D}_0 & \boldsymbol{D}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \text{diag}(\boldsymbol{\mu}_1) & \boldsymbol{D}_0 - \text{diag}(\boldsymbol{\mu}_1) & \boldsymbol{D}_1 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \text{diag}(\boldsymbol{\mu}_2) & \boldsymbol{D}_0 - \text{diag}(\boldsymbol{\mu}_2) & \boldsymbol{D}_1 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \text{diag}(\boldsymbol{\mu}_3) & \boldsymbol{D}_0 - \text{diag}(\boldsymbol{\mu}_3) & \boldsymbol{D}_1 & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{3}$$

a continuous interval, that is,

$$\mu_{n,j} \in U_{n,j} = [\mu_{n,j}^{\min},\ \mu_{n,j}^{\max}], \quad n = 1, 2, \ldots,\ j = 1, 2, \ldots, m \tag{8}$$

where $\mu_{n,j}^{\min}$ and $\mu_{n,j}^{\max}$ are constant. To guarantee the stability of the queue, we assume $\mu_{n,j}^{\max} > \lambda$. Note that we may further assume $\mu_{n,j}^{\min} = 0$ in some cases of this paper. Let

$$\begin{aligned} \boldsymbol{\mu}_n^{\max} &= (\mu_{n,1}^{\max}, \ldots, \mu_{n,m}^{\max})^T, \text{ for } n \geq 1; \\ \boldsymbol{\mu}^{\max} &= (\boldsymbol{\mu}_1^{\max}, \boldsymbol{\mu}_2^{\max}, \ldots). \end{aligned} \tag{9}$$

We assume that there exist policies for which the system is stable. The steady-state distribution is denoted as $\{\pi(n,j),\ n = 0, 1, \ldots$ and $j = 1, 2, \ldots, m\}$. We further define an m-dimensional row vector as

$$\boldsymbol{\pi}(n) = (\pi(n,1), \pi(n,2), \ldots, \pi(n,m)). \tag{10}$$

The infinite-dimensional row vector of the steady-state distribution is defined as

$$\boldsymbol{\pi} = (\boldsymbol{\pi}(0), \boldsymbol{\pi}(1), \boldsymbol{\pi}(2), \ldots). \tag{11}$$

It is well known that

$$\begin{aligned} \boldsymbol{\pi}\boldsymbol{B} &= \boldsymbol{0}, \\ \boldsymbol{\pi}\boldsymbol{e} &= 1. \end{aligned} \tag{12}$$

Similarly, we also define the following column vectors composed of element $f(n,j)$:

$$\begin{aligned} \boldsymbol{f}(n) &= (f(n,1), f(n,2), \ldots, f(n,m))^T, \text{ for } n = 0, 1, 2, \ldots; \\ \boldsymbol{f} &= (\boldsymbol{f}^T(0), \boldsymbol{f}^T(1), \boldsymbol{f}^T(2), \ldots)^T. \end{aligned} \tag{13}$$

Note that $\boldsymbol{f}$ is a column vector with infinitely many elements. The long-run average total cost of the system is defined as

$$\eta = \lim_{T\to\infty} \frac{1}{T} E\left\{\int_0^T f(N(t), J(t))dt\right\} = \boldsymbol{\pi}\boldsymbol{f} \tag{14}$$

where we assume that the Markov process $\{(N(t), J(t)), t > 0\}$ is a unichain, thus η is independent of the initial state $(N(0), J(0))$. The condition of $(N(t), J(t))$ being a unichain is complicated and one of the sufficient conditions is as follows: there exists a constant N such that $\mu_{n,j} > \lambda$, $\forall\ n > N$, where $j = 1, 2, \ldots, m$. We will further observe that the quasi-threshold optimal policy derived in Section IV satisfies such sufficient condition. The objective of this problem is to find the optimal service rates $\boldsymbol{\mu}^*$ from the value domain $\mathcal{U}$ that minimize the long-run average total cost, that is,

$$\boldsymbol{\mu}^* = \underset{\boldsymbol{\mu}\in\mathcal{U}}{\operatorname{argmin}}\{\eta\} = \underset{\substack{\mu_{n,j}^{\min} \leq \mu_{n,j} \leq \mu_{n,j}^{\max} \\ n=1,2,\ldots;\ j=1,\ldots,m}}{\operatorname{argmin}} \{\boldsymbol{\pi}\boldsymbol{f}\}. \tag{15}$$

Remark 1: The number of optimization variables in (15) is infinite. In the MDP terminology, the dimensions of the state space and the policy space of (15) are infinite. It is difficult to find the optimal policy of an MDP problem with infinite dimension.

III. Characterization of the Optimal Policy With the SBO Theory

In this section, we study the optimization problem (15) with the SBO theory [6], [7]. More specifically, we derive a difference formula that can quantify the performance difference of Markov systems under any two sets of service rates. The difference formula gives us advantage to study the relation between the performance and the parameters. Thus, some optimality properties can further be obtained.

First, we define the *performance potential* of the continuous time Markov process as

$$g(n,j) = \lim_{T\to\infty} E\left\{\int_{t=0}^T [f(N(t), J(t)) - \eta]dt \,\middle|\, (N(0), J(0)) = (n,j)\right\}. \tag{16}$$

From the above definition, we see that $g(n,j)$ quantifies the contribution of the initial state (n,j) to the long-run average performance of the Markov system. It is also called the *relative value function* or *bias* in the traditional MDP theory [20]. We further define a column vector $\boldsymbol{g}$ that is composed of elements $g(n,j),\ n = 0, 1, \ldots$ and $j = 1, 2, \ldots, m$. Specifically, we define

$$\begin{aligned} \boldsymbol{g}(n) &= (g(n,1), g(n,2), \ldots, g(n,m))^T, \text{ for } n = 0, 1, \ldots; \\ \boldsymbol{g} &= (\boldsymbol{g}^T(0), \boldsymbol{g}^T(1), \boldsymbol{g}^T(2), \ldots)^T. \end{aligned} \tag{17}$$

With (16), $g(n,j)$ can be decomposed as

$$\begin{aligned} & g(n,j) \\ &= E\{\tau\}[f(n,j) - \eta] + \lim_{T\to\infty} E\left\{\int_{t=\tau}^T [f(N(t), J(t)) - \eta]dt \,\middle|\, (N(0), J(0)) = (n,j)\right\} \\ &= E\{\tau\}[f(n,j) - \eta] + \sum_{\substack{(n',j')\in\mathcal{X}, \\ (n',j')\neq(n,j)}} Pr\Big\{(N(\tau), J(\tau)) = (n',j') \,\Big|\, (N(0), J(0)) = n,j)\Big\} \lim_{T\to\infty} E\left\{\int_{t=\tau}^T [f(N(t), J(t)) - \eta]dt \,\middle|\, (N(\tau), J(\tau)) = (n',j')\right\} \end{aligned} \tag{18}$$

where τ is the sojourn time of the Markov system staying at state (n,j). From the physical meaning of the transition rate of a Markov process, we know that

$$\begin{aligned} E\{\tau\} &= -\frac{1}{B((n,j),(n,j))}; \\ Pr\{(N(\tau), J(\tau)) &= (n',j') | (N(0), J(0)) = (n,j)\} \\ &= -\frac{B((n,j),(n',j'))}{B((n,j),(n,j))}. \end{aligned} \tag{19}$$

Substituting the above equation and (16) into (18), we have the following equation:

$$-B((n,j),(n,j))g(n,j) = f(n,j) - \eta + \sum_{\substack{(n',j')\in\mathcal{X},\\ (n',j')\neq(n,j)}} B((n,j),(n',j'))g(n',j') \quad (20)$$

Therefore, in a matrix form, we obtain the following equation:

$$\boldsymbol{B}\boldsymbol{g} = -\boldsymbol{f} + \eta\boldsymbol{e} \quad (21)$$

where the above equation has infinite dimension.

Suppose the policy of the Markov system is changed and the corresponding parameters under the new policy are denoted as $\boldsymbol{B}'$, $\boldsymbol{f}'$, $\boldsymbol{\pi}'$, and η', respectively. Below, we study the difference between η and η'. Premultiplying by $\boldsymbol{\pi}'$ on both sides of (21), we have

$$\boldsymbol{\pi}'\boldsymbol{B}\boldsymbol{g} = -\boldsymbol{\pi}'\boldsymbol{f}' + \boldsymbol{\pi}'(\boldsymbol{f}' - \boldsymbol{f}) + \eta\boldsymbol{\pi}'\boldsymbol{e}. \quad (22)$$

Since $\boldsymbol{\pi}'\boldsymbol{B}' = \boldsymbol{0}$, $\boldsymbol{\pi}'\boldsymbol{e} = 1$, and $\boldsymbol{\pi}'\boldsymbol{f}' = \eta'$, we can derive the *performance difference formula* as

$$\eta' - \eta = \boldsymbol{\pi}'[(\boldsymbol{B}' - \boldsymbol{B})\boldsymbol{g} + (\boldsymbol{f}' - \boldsymbol{f})]. \quad (23)$$

Remark 2: The right-hand side of the difference formula (23) is an inner product of two vectors with infinite dimension. That is, the SBO theory and the difference formula are also valid for the Markov system with an infinite state space.

The above difference formula is valid for the performance optimization of any Markov systems. Below, we apply it to study our specific problem (15).

Suppose the service rate is changed from $\boldsymbol{\mu}_n$ to $\boldsymbol{\mu}'_n$, $n = 1, 2, \ldots$. With the QBD structure of $\boldsymbol{B}$ in (3), we can write $\boldsymbol{B}' - \boldsymbol{B}$ as (24) that is shown in this page.

Substituting (24) and the cost function (4) into the difference formula (23), we have

$$\begin{aligned}\eta' - \eta &= \boldsymbol{\pi}'[(\boldsymbol{B}' - \boldsymbol{B})\boldsymbol{g} + (\boldsymbol{f}' - \boldsymbol{f})] \\ &= \sum_{n=1}^{\infty} \boldsymbol{\pi}'(n)\{\mathrm{diag}(\boldsymbol{\mu}'_n - \boldsymbol{\mu}_n)[\boldsymbol{g}(n-1) - \boldsymbol{g}(n)] + b(\boldsymbol{\mu}'_n - \boldsymbol{\mu}_n)\} \\ &= \sum_{n=1}^{\infty}\sum_{j=1}^{m} \pi'(n,j)(\mu'_{n,j} - \mu_{n,j})[g(n-1,j) - g(n,j) + b]. \quad (25)\end{aligned}$$

For simplicity, we further define a quantity $G(n,j)$ as

$$G(n,j) := g(n-1,j) - g(n,j) + b. \quad (26)$$

We can see that $G(n,j)$ quantifies the difference among two adjacent performance potentials $g(n,j)$ and $g(n-1,j)$. It measures the long-term effect on the average performance when the system state is changed from (n,j) to $(n-1,j)$, which indicates the occurrence of a service completion event. It is similar to a concept called *perturbation realization factor* in the theory of perturbation analysis [6]. We will further study how to efficiently compute this important quantity using MAM techniques in Section IV.

With (26), we obtain the following performance difference formula when the service rate is changed from $\mu_{n,j}$ to $\mu'_{n,j}$, where $n = 1, 2, \ldots$, and $j = 1, 2, \ldots, m$.

$$\boxed{\eta' - \eta = \sum_{n=1}^{\infty}\sum_{j=1}^{m}(\mu'_{n,j} - \mu_{n,j})\pi'(n,j)G(n,j).} \quad (27)$$

Remark 3: In (27), $G(n,j)$ is a quantity related to the performance potential of the current system and it can be computed or estimated from sample paths. Although the value of $\pi'(n,j)$ is unknown, it is always nonnegative. Thus, (27) clearly describes how the average cost η will be changed when the service rates $\mu_{n,j}$'s are changed. Difference formula (27) is the key result in solving the service rate control problem of the MAP/M/1 queue and it is the basis of all the following analysis for the optimization problem.

Based on (27), we can obtain the *necessary and sufficient condition* of the optimal service rate $\mu^*_{n,j}$ with the following lemma.

Lemma 1: $\mu^*_{n,j}$ is the optimal service rate, if and only if it satisfies the following condition:

$$\begin{aligned}&\mu^*_{n,j}G^*(n,j) \le \mu'_{n,j}G^*(n,j) \quad \forall\, \mu'_{n,j} \in [\mu^{\min}_{n,j}, \mu^{\max}_{n,j}], \\ &\text{and } n = 1, 2, \ldots, j = 1, 2, \ldots, m.\end{aligned} \quad (28)$$

where $G^*(n,j)$ is the quantity defined in (26) with service rates $\mu^*_{n,j}$.

Proof: This lemma can be proved by contradiction using (27). First, we prove the necessary condition. Assume that there exists a situation such that (28) does not hold. For example, suppose that we can find a particular service rate $\mu'_{k,l}$ such that $\mu^*_{k,l}G^*(k,l) > \mu'_{k,l}G^*(k,l)$. Then, we can construct a new policy $\boldsymbol{\mu}'$ as follows: choose $\mu'_{n,j} = \mu^*_{n,j}$ for state $(n,j) \neq (k,l)$ and choose $\mu'_{k,l}$ for state $(n,j) = (k,l)$. With (27), we can see $\eta' - \eta^* = (\mu'_{k,l} - \mu^*_{k,l})\pi'(k,l)G^*(k,l) < 0$. That is, $\eta' < \eta^*$ and it contradicts the fact that μ^* is optimal. Therefore, the assumption that (28) does not hold must be wrong and the necessary condition is proved.

Second, we prove the sufficient condition. Since (28) holds for any $\boldsymbol{\mu}'$, with (27) we have $\eta' - \eta^* = \sum_{n=1}^{\infty}\sum_{j=1}^{m}(\mu'_{n,j} - \mu^*_{n,j})\pi'(n,j)G^*(n,j) \ge 0$. That is, $\eta^* \le \eta'$ for any $\boldsymbol{\mu}'$. We can see that $\boldsymbol{\mu}^*$ is truly the optimal service rate. The lemma is proved. ■

$$\boldsymbol{B}' - \boldsymbol{B} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \mathrm{diag}(\boldsymbol{\mu}'_1 - \boldsymbol{\mu}_1) & \mathrm{diag}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}'_1) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}'_2 - \boldsymbol{\mu}_2) & \mathrm{diag}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}'_2) & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}'_3 - \boldsymbol{\mu}_3) & \mathrm{diag}(\boldsymbol{\mu}_3 - \boldsymbol{\mu}'_3) & \boldsymbol{0} & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \quad (24)$$

Difference formula (27) provides us a tool to thoroughly study the structure of the optimal service rate of the problem. Based on (27), we can further derive a monotonicity property in the following theorem.

Theorem 1: The long-run average total cost η is monotone w.r.t. the service rate $\mu_{n,j}$, for all $n = 1, 2, \ldots$ and $j = 1, 2, \ldots, m$.

Proof: Without loss of generality, we study a particular service rate, say $\mu_{k,l}$ at state (k, l). We assume state (k, l) is not transient, otherwise it has no effects on the long-run average total cost. Suppose the current service rate is $\boldsymbol{\mu}$. We only change the service rate $\mu_{k,l}$ to $\mu'_{k,l}$ and maintain other service rates the same as $\mu_{n,j}$. That is, for the new service rate $\boldsymbol{\mu}'$, we have $\mu'_{n,j} = \mu_{n,j}$ for $(n, j) \neq (k, l)$. With (27), the change in the long-run average total cost is given as

$$\eta' - \eta = (\mu'_{k,l} - \mu_{k,l})\pi'(k,l)G(k,l). \tag{29}$$

On the other hand, we consider the above scenario reversely. Suppose the current service rate is $\boldsymbol{\mu}'$ that is defined as above. We change the service rate $\mu'_{k,l}$ to $\mu_{k,l}$ and maintain other service rates unvaried. That is, the service rate is changed from $\boldsymbol{\mu}'$ to $\boldsymbol{\mu}$. With (27), the change in the long-run average total cost can be written as

$$\eta - \eta' = (\mu_{k,l} - \mu'_{k,l})\pi(k,l)G'(k,l). \tag{30}$$

Comparing (29) and (30), we have

$$\frac{G'(k,l)}{G(k,l)} = \frac{\pi'(k,l)}{\pi(k,l)} > 0 \tag{31}$$

where $\pi'(k, l)$ and $\pi(k, l)$ are always positive since state (k, l) is not transient. Therefore, we can see that the sign of $G(k, l)$ remains unvaried when $\mu_{k,l}$ is changed. With (29), we can obtain the derivative formula as

$$\frac{d\eta}{d\mu_{k,l}} = \pi(k,l)G(k,l). \tag{32}$$

Since the sign of $G(k, l)$ remains unvaried when $\mu_{k,l}$ is changed, the sign of the above derivative is also unvaried. That is, η is monotone w.r.t. $\mu_{k,l}$ and the theorem is proved.

Remark 4: Although Theorem 1 claims the monotonicity of the long-run average total cost, we do not know if it is monotonically increasing or decreasing. Actually, η might be monotonically increasing in some $\mu_{n,j}$ and monotonically decreasing in other $\mu_{n,j}$, which is demonstrated in a numerical experiment later (see Example 1 in Section V).

With Lemma 1 and Theorem 1, we can directly obtain the following theorem.

Theorem 2: The optimal service rate $\mu^*_{n,j}$ can be either $\mu^{\min}_{n,j}$ or $\mu^{\max}_{n,j}$, for all $n = 1, 2, \ldots$ and $j = 1, 2, \ldots, m$. More specifically, we have

$$\begin{array}{l} \text{if } G^*(n,j) \geq 0, \text{ then } \mu^*_{n,j} = \mu^{\min}_{n,j} \\ \text{if } G^*(n,j) < 0, \text{ then } \mu^*_{n,j} = \mu^{\max}_{n,j}. \end{array} \tag{33}$$

That is, the bang–bang control is optimal for the service rate control of the MAP/M/1 queue.

The proof of the above theorem is straightforward and we omit it for simplicity. With Theorem 2, to find the optimal policy, the value domain of $\mu_{n,j}$ can be reduced from a continuous interval $[\mu^{\min}_{n,j}, \mu^{\max}_{n,j}]$ to a two-element set $\{\mu^{\min}_{n,j}, \mu^{\max}_{n,j}\}$, for all $n = 1, 2, \ldots$ and $j = 1, 2, \ldots, m$. This is a significant reduction of the optimization complexity for problem (15).

IV. Threshold-Type Policy and Optimization Algorithm

In this section, we combine the MAM theory and the SBO theory to study the computation and optimization related issues of the service rate control problem of the MAP/M/1 queue. Based on the QBD structure of $\boldsymbol{B}$, we utilize the MAM theory to numerically compute $G(n, j)$'s in a recursive manner. The optimality of threshold-type policy is further proved under some mild conditions. Finally, an iterative optimization algorithm is proposed to efficiently find the optimal policy. Note that in some places of this section, we assume that $\mu^{\max}_{n,j}$ and $\phi(n, j)$ are nondecreasing in n, which is quite justifiable in practice. We also assume $\mu^{\min}_{n,j} = 0$, for all n and j, in the remainder of this paper.

A. Optimality of Quasi-Threshold-Type Policy in MAP/M/1

From the analysis presented in Section III, we see that all the results and the algorithm depend on $G(n, j)$ defined in (26). One of its intuitive explanations is that $G(n, j)$ quantifies the long-term effect on the average performance when the (initial) system state is changed from (n, j) to $(n - 1, j)$, which indicates a service completion. This quantity is fundamental for our optimization approach. We denote its vector form as

$$\boldsymbol{G}(n) = (G(n,1), G(n,2), \ldots, G(n,m))^T, \quad n = 1, 2, \ldots. \tag{34}$$

By definition, we have $\boldsymbol{G}(n) = \boldsymbol{g}(n-1) - \boldsymbol{g}(n) + b\boldsymbol{e}$, for $n \geq 1$. Below, we discuss how to compute $\boldsymbol{G}(n)$.

By (20), we obtain

$$\begin{aligned} -D_0(i,i)g(0,i) &= f(0,i) - \eta \\ &+ \sum_{j=1:\, j\neq i}^{m} D_0(i,j)g(0,j) + \sum_{j=1}^{m} D_1(i,j)g(1,j) \end{aligned} \tag{35}$$

and, for $n \geq 1$

$$\begin{aligned} -(D_0(i,i) - \mu_{n,i})g(n,i) &= f(n,i) - \eta + \mu_{n,i}g(n-1,i) \\ + \sum_{j=1:\, j\neq i}^{m} D_0(i,j)g(n,j) &+ \sum_{j=1}^{m} D_1(i,j)g(n+1,j). \end{aligned} \tag{36}$$

In matrix form, the above equations become

$$\begin{aligned} &\eta\boldsymbol{e} = \boldsymbol{f}(0) + \boldsymbol{D}_0\boldsymbol{g}(0) + \boldsymbol{D}_1\boldsymbol{g}(1); \\ &\eta\boldsymbol{e} = \boldsymbol{f}(n) + \text{diag}(\boldsymbol{\mu}_n)\boldsymbol{g}(n-1) + (\boldsymbol{D}_0 - \text{diag}(\boldsymbol{\mu}_n))\boldsymbol{g}(n) \\ &\quad + \boldsymbol{D}_1\boldsymbol{g}(n+1), \qquad \text{for } n \geq 1. \end{aligned} \tag{37}$$

The above equations can be rewritten as

$$\begin{aligned} &\boldsymbol{0} = \eta\boldsymbol{e} - \boldsymbol{f}(0) - \boldsymbol{D}_0\boldsymbol{g}(0) - \boldsymbol{D}_1\boldsymbol{g}(1); \\ &\text{diag}(\boldsymbol{\mu}_n)(\boldsymbol{g}(n-1) - \boldsymbol{g}(n)) = \eta\boldsymbol{e} - \boldsymbol{f}(n) - \boldsymbol{D}_0\boldsymbol{g}(n) \\ &- \boldsymbol{D}_1\boldsymbol{g}(n+1), \qquad \text{for } n \geq 1. \end{aligned} \tag{38}$$

By adding $\mathrm{diag}(\boldsymbol{\mu}_n)b\boldsymbol{e} = b\boldsymbol{\mu}_n$ on both sides for $n \geq 1$, we obtain

$$\begin{aligned}&\mathrm{diag}(\boldsymbol{\mu}_n)(\boldsymbol{g}(n-1) - \boldsymbol{g}(n) + b\boldsymbol{e})\\ &= \eta\boldsymbol{e} - \boldsymbol{f}(n) - \boldsymbol{D}_0\boldsymbol{g}(n) - \boldsymbol{D}_1\boldsymbol{g}(n+1) + b\boldsymbol{\mu}_n \end{aligned} \tag{39}$$

or, equivalently

$$\mathrm{diag}(\boldsymbol{\mu}_n)\boldsymbol{G}(n) = \eta\boldsymbol{e} - \boldsymbol{\phi}(n) - \boldsymbol{D}_0\boldsymbol{g}(n) - \boldsymbol{D}_1\boldsymbol{g}(n+1) \tag{40}$$

where $\boldsymbol{\phi}(n) = (\phi(n,1), \phi(n,2), \ldots, \phi(n,m))^T = \boldsymbol{f}(n) - b\boldsymbol{\mu}_n$, which is indicated by (4). Using the above equations for n and $n+1$, we obtain, for $n \geq 1$

$$\begin{aligned}&\mathrm{diag}(\boldsymbol{\mu}_n)\boldsymbol{G}(n) - \mathrm{diag}(\boldsymbol{\mu}_{n+1})\boldsymbol{G}(n+1)\\ &= \boldsymbol{\phi}(n+1) - \boldsymbol{\phi}(n) - \boldsymbol{D}_0(\boldsymbol{g}(n) - \boldsymbol{g}(n+1))\\ &\quad - \boldsymbol{D}_1(\boldsymbol{g}(n+1) - \boldsymbol{g}(n+2))\\ &= d\boldsymbol{\phi}(n+1) - \boldsymbol{D}_0(\boldsymbol{g}(n) - \boldsymbol{g}(n+1) + b\boldsymbol{e})\\ &\quad - \boldsymbol{D}_1(\boldsymbol{g}(n+1) - \boldsymbol{g}(n+2) + b\boldsymbol{e})\\ &= d\boldsymbol{\phi}(n+1) - \boldsymbol{D}_0\boldsymbol{G}(n+1) - \boldsymbol{D}_1\boldsymbol{G}(n+2) \end{aligned} \tag{41}$$

where we define $d\boldsymbol{\phi}(n+1) := \boldsymbol{\phi}(n+1) - \boldsymbol{\phi}(n)$. Note that we used the fact $(\boldsymbol{D}_0 + \boldsymbol{D}_1)\boldsymbol{e} = \boldsymbol{0}$ in the second equality of the above equation. The above equation leads to the following useful equation for $\boldsymbol{G}(n+1)$, for $n \geq 0$

$$\begin{aligned}\boldsymbol{G}(n+1) = &\ (\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0)^{-1}\Big(- d\boldsymbol{\phi}(n+1)\\ &+ \mathrm{diag}(\boldsymbol{\mu}_n)\boldsymbol{G}(n) + \boldsymbol{D}_1\boldsymbol{G}(n+2)\Big). \end{aligned} \tag{42}$$

Note that $\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0$ is an M-matrix with dominating diagonal elements for all $\boldsymbol{\mu}_{n+1}(\geq 0)$ and, consequently, is invertible. It is also well known that all elements of the inverse matrix are positive [16].

Lemma 2: We assume that there exists a positive integer N such that 1) $\boldsymbol{\mu}_n$ is nondecreasing in n; and 2) $\lambda < \boldsymbol{\omega}\boldsymbol{\mu}_n$, for $n \geq N$. Then, vector $\boldsymbol{G}(n)$ can be obtained as, for $n \geq N$

$$\boldsymbol{G}(n+1) = \boldsymbol{\xi}(n+1) + \boldsymbol{A}_{n+1}\mathrm{diag}(\boldsymbol{\mu}_n)\boldsymbol{G}(n) \tag{43}$$

where $\boldsymbol{A}_{n+1}$ and $\boldsymbol{\xi}(n+1)$ can be computed recursively as

$$\begin{aligned}\boldsymbol{A}_{n+1} = &\ \left(\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+2}\mathrm{diag}(\boldsymbol{\mu}_{n+1})\right)^{-1};\\ \boldsymbol{\xi}(n+1) = &\ \boldsymbol{A}_{n+1}\left(-d\boldsymbol{\phi}(n+1) + \boldsymbol{D}_1\boldsymbol{\xi}(n+2)\right). \end{aligned} \tag{44}$$

Proof: We construct a level-dependent GI/M/1 type CTMC with transition rate matrix (45), where $d\boldsymbol{\mu}_{n+1} := \boldsymbol{\mu}_{n+1} - \boldsymbol{\mu}_n$.

It is readily seen that the first state of the Markov chain $\boldsymbol{Q}_N$ is an absorbing state. We claim that 1) the absorption probability from any state to that state is one, and 2) the expected absorption time is finite. To show the claims, we construct two more CTMCs $\tilde{\boldsymbol{Q}}_N$ and $\tilde{\tilde{\boldsymbol{Q}}}_N$ as (46) and (47) (45) shown at the bottom of the page. These two matrices have a similar structure as that of $\boldsymbol{Q}_N$ in (45).

If the two processes $\boldsymbol{Q}_N$ and $\tilde{\boldsymbol{Q}}_N$ both start in the same state, a sample path comparison shows that the process $\boldsymbol{Q}_N$ will be absorbed earlier than $\tilde{\boldsymbol{Q}}_N$. Thus, if the two claims hold for $\tilde{\boldsymbol{Q}}_N$, they hold for $\boldsymbol{Q}_N$. Since $\boldsymbol{\mu}_{n+1} \geq \boldsymbol{\mu}_{N+1}$, for $n \geq N$, it is easy to see that if the two claims hold for the Markov chain $\tilde{\tilde{\boldsymbol{Q}}}_N$, they hold for $\tilde{\boldsymbol{Q}}_N$. Thus, if the two claims hold for $\tilde{\tilde{\boldsymbol{Q}}}_N$, they hold for both $\tilde{\boldsymbol{Q}}_N$ and $\boldsymbol{Q}_N$. Since $\tilde{\tilde{\boldsymbol{Q}}}_N$ is a level-independent (absorption) QBD process, under the condition $\lambda < \boldsymbol{\omega}\boldsymbol{\mu}_{N+1}$, the two claims hold (see [19, ch. 1 and 3]). Consequently, the two claims hold for $\boldsymbol{Q}_N$, a fact that is used in the rest of the proof.

Now, we are back to $\boldsymbol{Q}_N$. In (42), if we define $-d\boldsymbol{\phi}(n)$ as some kind of cost, then $\boldsymbol{G}(n)$ can be interpreted as the expected total cost incurred during the first passage time (i.e., fundamental period) from level $n-N$ to the absorption state in the CTMC $\boldsymbol{Q}_N$. Since the expected absorption time is finite, the expected total cost is finite as well. Consequently, we have shown that $\boldsymbol{G}(n)$ exists for $n \geq N+1$.

$$\boldsymbol{Q}_N = \begin{pmatrix} 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{\mu}_{N+1} & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ d\boldsymbol{\mu}_{N+2} & \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+2}) & \boldsymbol{D}_1 & \boldsymbol{0} & \cdots \\ d\boldsymbol{\mu}_{N+3} & \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}_{N+2}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+3}) & \boldsymbol{D}_1 & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{45}$$

$$\tilde{\boldsymbol{Q}}_N = \begin{pmatrix} 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{\mu}_{N+1} & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}_{N+2}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+2}) & \boldsymbol{D}_1 & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}_{N+3}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+3}) & \boldsymbol{D}_1 & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{46}$$

$$\tilde{\tilde{\boldsymbol{Q}}}_N = \begin{pmatrix} 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{\mu}_{N+1} & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_1 & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_1 & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_0 - \mathrm{diag}(\boldsymbol{\mu}_{N+1}) & \boldsymbol{D}_1 & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{47}$$

To prove (43), we decompose $\boldsymbol{G}(n+1)$ into two parts: the expected total cost incurred during the transitions from level $n+1-N$ to level $n-N$, which is denoted by $\boldsymbol{\xi}(n+1)$, and the expected total cost incurred during the transitions from level $n-N$ to the absorption state, which is $\boldsymbol{G}(n)$. Then, we can obtain

$$\boldsymbol{G}(n+1) = \boldsymbol{\xi}(n+1) + \tilde{\boldsymbol{A}}_{n+1}\boldsymbol{G}(n) \tag{48}$$

where $\tilde{\boldsymbol{A}}_{n+1}$ is an $m \times m$ matrix whose (i,j)th element is the probability that the CTMC $\boldsymbol{Q}_N$ reaches level $n-N$ for the first time at phase j, given that the Markov chain was initially in level $n+1-N$ at phase i. By definition, it is easy to see $\tilde{\boldsymbol{A}}_n\boldsymbol{e} \le \boldsymbol{e}$.

Conditioning on the first transition, it can be shown that

$$\begin{aligned}\tilde{\boldsymbol{A}}_{n+1} = \ & (\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0)^{-1}\mathrm{diag}(\boldsymbol{\mu}_n) \\ & + (\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1\tilde{\boldsymbol{A}}_{n+2}\tilde{\boldsymbol{A}}_{n+1}. \end{aligned} \tag{49}$$

Let $\boldsymbol{A}_{n+1} = \tilde{\boldsymbol{A}}_{n+1}(\mathrm{diag}(\boldsymbol{\mu}_n))^{-1}$. By (49), it is easy to show that the equation for $\boldsymbol{A}_{n+1}$ and $\boldsymbol{A}_{n+2}$ in (44) holds, provided that the inverse on the right-hand side exists. To show that $\boldsymbol{A}_{n+1}$ exists and can be computed recursively using (44), we need to prove that matrix $\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+2}\mathrm{diag}(\boldsymbol{\mu}_{n+1})$ is invertible. Since $\tilde{\boldsymbol{A}}_{n+2} = \boldsymbol{A}_{n+2}\mathrm{diag}(\boldsymbol{\mu}_{n+1})$ and $\tilde{\boldsymbol{A}}_{n+2}\boldsymbol{e} \le \boldsymbol{e}$, for $n \ge N$, we have

$$\left(\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+2}\mathrm{diag}(\boldsymbol{\mu}_{n+1})\right)\boldsymbol{e} \ge \boldsymbol{\mu}_{n+1} \ge \boldsymbol{0}. \tag{50}$$

The matrix $\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+2}\mathrm{diag}(\boldsymbol{\mu}_{n+1})$ is an M-matrix with dominating diagonal elements. Thus, the matrix is invertible and, consequently, matrix $\boldsymbol{A}_n$ can be computed recursively using (44) for $n \ge N+1$.

Finally, the equation for $\boldsymbol{\xi}_{n+1}$ and $\boldsymbol{\xi}_{n+2}$ in (44) can be obtained by conditioning on the first transition of CTMC $\boldsymbol{Q}_N$:

$$\begin{aligned}\boldsymbol{\xi}(n+1) = \ & (\mathrm{diag}(\boldsymbol{\mu}_{n+1}) - \boldsymbol{D}_0)^{-1}(-d\boldsymbol{\phi}(n+1) \\ & + \boldsymbol{D}_1(\boldsymbol{\xi}(n+2) + \tilde{\boldsymbol{A}}_{n+2}\boldsymbol{\xi}(n+1)))\end{aligned} \tag{51}$$

which leads to the second equation in (44). This completes the proof of Lemma 2.

Based on Lemma 2, we have following observations on matrix $\boldsymbol{A}_n$ and vector $d\boldsymbol{\phi}(n)$

1) *Observation 1:* All elements of matrix $\boldsymbol{A}_n$ are positive.
2) *Observation 2:* If $d\boldsymbol{\phi}(n+k) \ge 0$ for all $k \ge 0$, then all elements of $\boldsymbol{\xi}(n)$ are nonpositive.

The above results lead to the following characterization of the optimal policy, which implies that the optimal policy $\boldsymbol{\mu}^*$ is of a *quasi-threshold-type* for individual phases.

Theorem 3: Assume the following:
1) $\boldsymbol{\mu}_{N+k}^{\max}$ is nondecreasing in k;
2) $\boldsymbol{\mu}_{N+k}^{\min} = \boldsymbol{0}$;
3) $\lambda < \boldsymbol{\omega}\boldsymbol{\mu}_{N+k}^{\max}$; and
4) $d\boldsymbol{\phi}(N+k) \ge 0$, for a fixed N and any $k \ge 0$.

If $\mu_{N,i}^* = \mu_{N,i}^{\max}$, for some i $(i = 1, 2, \ldots, m)$, then we have $\mu_{N+k,j}^* = \mu_{N+k,j}^{\max}$, for any $k > 0$ and $j = 1, 2, \ldots, m$. That is, the optimal service rate has a quasi-threshold-type.

Proof: By Lemma 2, $\boldsymbol{A}_{N+k}$ exists for $k \ge 0$. For $n \ge N$, we divide vector $\boldsymbol{G}(n)$ into $\boldsymbol{G}^-(n)$ and $\boldsymbol{G}^+(n)$ according to the sign of $G(n,j)$, where all elements of $\boldsymbol{G}^-(n)$ are negative and all elements of $\boldsymbol{G}^+(n)$ are nonnegative. We divide $\boldsymbol{\mu}_n^*$ into $\boldsymbol{\mu}_n^{*,-}$ and $\boldsymbol{\mu}_n^{*,+}$, and $\boldsymbol{\mu}_n^{\max}$ into $\boldsymbol{\mu}_n^{\max,-}$ and $\boldsymbol{\mu}_n^{\max,+}$, accordingly. By Theorem 2, we must have $\boldsymbol{\mu}_n^{*,-} = \boldsymbol{\mu}_n^{\max,-}$ and $\boldsymbol{\mu}_n^{*,+} = \boldsymbol{0}$. We divide $\boldsymbol{\xi}(n+1)$ into $\{\boldsymbol{\xi}^-(n+1), \boldsymbol{\xi}^+(n+1)\}$, and $\boldsymbol{A}_{n+1}$ into $\{\boldsymbol{A}_{n+1}^{-,-}, \boldsymbol{A}_{n+1}^{-,+}, \boldsymbol{A}_{n+1}^{+,-}, \boldsymbol{A}_{n+1}^{+,+}\}$, accordingly. By (43), we obtain

$$\begin{aligned}\begin{pmatrix}\boldsymbol{G}^-(n+1)\\ \boldsymbol{G}^+(n+1)\end{pmatrix} = \begin{pmatrix}\boldsymbol{\xi}^-(n+1)\\ \boldsymbol{\xi}^+(n+1)\end{pmatrix} + \begin{pmatrix}\boldsymbol{A}_{n+1}^{-,-} & \boldsymbol{A}_{n+1}^{-,+}\\ \boldsymbol{A}_{n+1}^{+,-} & \boldsymbol{A}_{n+1}^{+,+}\end{pmatrix} \\ \cdot \begin{pmatrix}\mathrm{diag}(\boldsymbol{\mu}_n^{\max,-}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{G}^-(n)\\ \boldsymbol{G}^+(n)\end{pmatrix}\end{aligned} \tag{52}$$

which leads to

$$\begin{aligned}\boldsymbol{G}^-(n+1) &= \boldsymbol{\xi}^-(n+1) + \boldsymbol{A}_{n+1}^{-,-}\mathrm{diag}(\boldsymbol{\mu}_n^{\max,-})\boldsymbol{G}^-(n) < 0,\\ \boldsymbol{G}^+(n+1) &= \boldsymbol{\xi}^+(n+1) + \boldsymbol{A}_{n+1}^{+,-}\mathrm{diag}(\boldsymbol{\mu}_n^{\max,-})\boldsymbol{G}^-(n) < 0\end{aligned} \tag{53}$$

where we utilize the observation that $\boldsymbol{A}_n$ is positive and $\boldsymbol{\xi}(n)$ is nonnegative. Therefore, by Theorem 2, we must have $\boldsymbol{\mu}_{n+1}^{*,-} = \boldsymbol{\mu}_{n+1}^{\max,-}$. The desired result follows.

Remark 5: With Theorem 3, we see that the optimal service rate of the MAP/M/1 queue has a quasi-threshold-type. That is, there exists a threshold N such that if $n > N$, $\mu_{n,j}^* = \mu_{n,j}^{\max}$, for $j = 1, 2, \ldots, m$.

Note that in Theorem 3, we only require that $\boldsymbol{\mu}_n^{\max}$ and $\boldsymbol{\phi}(n)$ are nondecreasing in n when n is large enough. If $\boldsymbol{\mu}_n^{\max}$ and $\boldsymbol{\phi}(n)$ are always nondecreasing in n, then the optimal service rate of MAP/M/1 can have an even more special form called *strong quasi-threshold-type*. We can derive the following corollary directly based on Theorem 3.

Corollary 1: If $\boldsymbol{\mu}_n^{\max}$ and $\boldsymbol{\phi}(n)$ are always nondecreasing in n, $\boldsymbol{\mu}_n^{\min} = \boldsymbol{0}$, and $\lambda < \boldsymbol{\omega}\boldsymbol{\mu}_n^{\max}$ for all n, then the optimal service rate of the MAP/M/1 queue can be of the strong quasi-threshold-type. That is, there exists a threshold N such that if $n > N$, $\mu_{n,j}^* = \mu_{n,j}^{\max}$; if $n < N$, $\mu_{n,j}^* = 0$; if $n = N$, $\mu_{n,j}^*$ can be either 0 or $\mu_{n,j}^{\max}$ at different phase j.

B. Computation and Optimization Algorithm

The optimality of the quasi-threshold-type policy greatly reduces the optimization complexity of the problem. The form of the quasi-threshold-type policy is simple and it is easy to adopt in practice. Next, we develop an iterative algorithm to find the optimal policy. We further assume that for sufficiently large N, 1) $\boldsymbol{\mu}_n = \boldsymbol{\mu}_n^{\max} = \boldsymbol{\mu}_\infty^{\max}$ and 2) $d\boldsymbol{\phi}(n) = d\boldsymbol{\phi}_\infty$, for $n \ge N$. We use $\boldsymbol{A}_\infty$ to denote $\boldsymbol{A}_n$ for $n \ge N$. Then, we can rewrite (49) as follows:

$$\begin{aligned}\tilde{\boldsymbol{A}}_\infty = \ & (\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0)^{-1}\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) \\ & + (\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1(\tilde{\boldsymbol{A}}_\infty)^2\end{aligned} \tag{54}$$

where $\tilde{\boldsymbol{A}}_\infty = \boldsymbol{A}_\infty\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max})$. Using (54), the matrix $\tilde{\boldsymbol{A}}_\infty$ can be obtained numerically. The basic idea is described by Algorithm 1 as follows.

The computation procedure of Algorithm 1 is similar to that of the well-known matrix G for the QBD process and the convergence can be guaranteed (see [19, ch. 3]). In fact, $\tilde{\boldsymbol{A}}_\infty$ is the minimum nonnegative solution to (54).

Algorithm 1: A recursive numerical computation algorithm for $\tilde{\boldsymbol{A}}_\infty$ based on (54).

1) Initialize $\tilde{\boldsymbol{A}}_\infty$ arbitrarily, e.g., set $\tilde{\boldsymbol{A}}_\infty^{(0)} = \boldsymbol{0}$, $k = 0$, and $\epsilon > 0$.
2) Repeat:
   $\tilde{\boldsymbol{A}}_\infty^{(k+1)} = (\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0)^{-1}\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) + (\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1(\tilde{\boldsymbol{A}}_\infty^{(k)})^2$;
   $k \leftarrow k+1$;
   Until the stopping criterion $||\tilde{\boldsymbol{A}}_\infty^{(k)} - \tilde{\boldsymbol{A}}_\infty^{(k-1)}|| < \epsilon$ is satisfied.
3) Output $\tilde{\boldsymbol{A}}_\infty^{(k)}$ as the value of $\tilde{\boldsymbol{A}}_\infty$.

We use $\boldsymbol{\xi}_\infty$ to denote $\boldsymbol{\xi}(n)$ for $n \geq N$, and $d\boldsymbol{\phi}_\infty$ to denote $d\boldsymbol{\phi}(n)$ for $n \geq N$. If $\lambda < \boldsymbol{\omega}\boldsymbol{\mu}_\infty$, by (44), then we obtain

$$\begin{aligned}\boldsymbol{\xi}_\infty &= -(\boldsymbol{I} - \boldsymbol{A}_\infty\boldsymbol{D}_1)^{-1}\boldsymbol{A}_\infty d\boldsymbol{\phi}_\infty \\ &= -(\boldsymbol{A}_\infty^{-1} - \boldsymbol{D}_1)^{-1}d\boldsymbol{\phi}_\infty \\ &= -\left(\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0 - \boldsymbol{D}_1(\tilde{\boldsymbol{A}}_\infty + \boldsymbol{I})\right)^{-1} d\boldsymbol{\phi}_\infty.\end{aligned} \tag{55}$$

Note that it is well known from the QBD theory that $\tilde{\boldsymbol{A}}_\infty$ is finite and the matrix $\mathrm{diag}(\boldsymbol{\mu}_\infty^{\max}) - \boldsymbol{D}_0 - \boldsymbol{D}_1(\tilde{\boldsymbol{A}}_\infty + \boldsymbol{I})$ is invertible.

Once $\boldsymbol{A}_\infty$ and $\boldsymbol{\xi}_\infty$ are obtained, we extend (44) from $n \geq N$ to $n = N-1, \ldots, 1$ to compute $\{\boldsymbol{A}_n,\ n = N, N-1, \ldots, 1\}$ and $\{\boldsymbol{\xi}(n),\ n = N, N-1, \ldots, 1\}$, by assuming the existence of $\{\boldsymbol{A}_n,\ n = N, N-1, \ldots, 1\}$. By (40) and the first equality in (38), we obtain

$$\begin{aligned}\boldsymbol{G}(1) = &-(\mathrm{diag}(\boldsymbol{\mu}_1) - \boldsymbol{D}_0)^{-1}d\boldsymbol{\phi}(1) \\ &+ (\mathrm{diag}(\boldsymbol{\mu}_1) - \boldsymbol{D}_0)^{-1}\boldsymbol{D}_1\boldsymbol{G}(2).\end{aligned} \tag{56}$$

Replacing $\boldsymbol{G}(2)$ in the above equation with its expression in (43) yields

$$\boldsymbol{G}(1) = \boldsymbol{A}_1\left(-d\boldsymbol{\phi}(1) + \boldsymbol{D}_1\boldsymbol{\xi}(2)\right). \tag{57}$$

By comparing the above expression of $\boldsymbol{G}(1)$ and that of $\boldsymbol{\xi}(1)$ [see (44)], it is clear that $\boldsymbol{\xi}(1) = \boldsymbol{G}(1)$. Then, we extend (43) from large n (i.e., $n \geq N$) to $n = 2, 3, \ldots, N$, and compute $\{\boldsymbol{G}(n),\ n = 2, 3, \ldots, N\}$ recursively. The following lemma ensures the existence and finiteness of matrices $\{\boldsymbol{A}_n,\ n = N, N-1, \ldots, 1\}$, vectors $\{\boldsymbol{\xi}(n),\ n = N, N-1, \ldots, 1\}$, and vectors $\{\boldsymbol{G}(n),\ n = 1, 2, \ldots, N\}$.

Lemma 3: Assume that $\boldsymbol{D} = \boldsymbol{D}_0 + \boldsymbol{D}_1$ is irreducible. Then, $\{\boldsymbol{A}_n,\ n = 1, 2, \ldots\}$ exist and are finite. Given the existence of $\boldsymbol{A}_\infty$ and $\boldsymbol{\xi}_\infty$ (e.g., under the conditions 1) $\boldsymbol{\mu}_n = \boldsymbol{\mu}_n^{\max} = \boldsymbol{\mu}_\infty^{\max}$ and 2) $d\boldsymbol{\phi}(n) = d\boldsymbol{\phi}_\infty$, for $n \geq N$), then $\{\boldsymbol{\xi}(n),\ n = N, N-1, \ldots, 1\}$ and $\{\boldsymbol{G}(n),\ n = 1, 2, \ldots, N\}$ exist and are finite.

Proof: Since $\boldsymbol{D}$ is irreducible, every element of $\boldsymbol{\omega}$ (defined in Section II) is positive. Define $\hat{\boldsymbol{D}}_0 = \mathrm{diag}^{-1}(\boldsymbol{\omega})\boldsymbol{D}_0^T\mathrm{diag}(\boldsymbol{\omega})$, $\hat{\boldsymbol{D}}_1 = \mathrm{diag}^{-1}(\boldsymbol{\omega})\boldsymbol{D}_1^T\mathrm{diag}(\boldsymbol{\omega})$, and $\hat{\boldsymbol{A}}_n = \mathrm{diag}^{-1}(\boldsymbol{\omega})\boldsymbol{A}_n^T\boldsymbol{D}_1^T\mathrm{diag}(\boldsymbol{\omega})$. It is easy to verify that $\hat{\boldsymbol{D}}_0 + \hat{\boldsymbol{D}}_1$ is an irreducible infinitesimal generator. Premultiplying by $\boldsymbol{D}_1$ on both sides of the first equation in (44), we obtain

$$\boldsymbol{D}_1\boldsymbol{A}_n = \boldsymbol{D}_1\left(\mathrm{diag}(\boldsymbol{\mu}_n) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+1}\mathrm{diag}(\boldsymbol{\mu}_n)\right)^{-1} \tag{58}$$

which leads to

$$\begin{aligned}\boldsymbol{D}_1\boldsymbol{A}_n = &\boldsymbol{D}_1\left(\mathrm{diag}(\boldsymbol{\mu}_n) - \boldsymbol{D}_0\right)^{-1} \\ &+ \boldsymbol{D}_1\boldsymbol{A}_n\boldsymbol{D}_1\boldsymbol{A}_{n+1}\mathrm{diag}(\boldsymbol{\mu}_n)\left(\mathrm{diag}(\boldsymbol{\mu}_n) - \boldsymbol{D}_0\right)^{-1}.\end{aligned} \tag{59}$$

By routine calculations, we obtain

$$\hat{\boldsymbol{A}}_n = (\mathrm{diag}(\boldsymbol{\mu}_n) - \hat{\boldsymbol{D}}_0)^{-1}\left(\hat{\boldsymbol{D}}_1 + \mathrm{diag}(\boldsymbol{\mu}_n)\hat{\boldsymbol{A}}_{n+1}\hat{\boldsymbol{A}}_n\right). \tag{60}$$

Similar to the proof of Lemma 2, we construct the following CTMC in (61) shown in the bottom this page. Equation (60) implies that elements of $\hat{\boldsymbol{A}}_n$ are the transition probabilities of the phase of $\hat{\boldsymbol{Q}}$ during the fundamental period from level n to $n-1$. Consequently, $\hat{\boldsymbol{A}}_n\boldsymbol{e} \leq \boldsymbol{e}$, for $n \geq 1$.

Rewrite the matrix on the right-hand side of the first equation in (44) as

$$\begin{aligned}&\mathrm{diag}(\boldsymbol{\mu}_n) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+1}\mathrm{diag}(\boldsymbol{\mu}_n) \\ &= \mathrm{diag}^{-1}(\boldsymbol{\omega})\left(\mathrm{diag}(\boldsymbol{\mu}_n) - \hat{\boldsymbol{D}}_0^T - \hat{\boldsymbol{A}}_{n+1}^T\mathrm{diag}(\boldsymbol{\mu}_n)\right)\mathrm{diag}(\boldsymbol{\omega}) \\ &= \mathrm{diag}^{-1}(\boldsymbol{\omega})\left(\mathrm{diag}(\boldsymbol{\mu}_n) - \hat{\boldsymbol{D}}_0 - \mathrm{diag}(\boldsymbol{\mu}_n)\hat{\boldsymbol{A}}_{n+1}\right)^T\mathrm{diag}(\boldsymbol{\omega}).\end{aligned} \tag{62}$$

Similar to the proof of Lemma 2, it can be shown that matrix $\mathrm{diag}(\boldsymbol{\mu}_n) - \hat{\boldsymbol{D}}_0 - \mathrm{diag}(\boldsymbol{\mu}_n)\hat{\boldsymbol{A}}_{n+1}$ is invertible, which implies that matrix $\mathrm{diag}(\boldsymbol{\mu}_n) - \boldsymbol{D}_0 - \boldsymbol{D}_1\boldsymbol{A}_{n+1}\mathrm{diag}(\boldsymbol{\mu}_n)$ is invertible. Therefore, $\boldsymbol{A}_n$ exists for all $n \geq 1$. If $\boldsymbol{A}_\infty$ and $\boldsymbol{\xi}_\infty$ exist and are finite, it is easy to verify that $\{\boldsymbol{\xi}(n),\ n = N, N-1, \ldots, 1\}$ and $\{\boldsymbol{G}(n),\ n = 1, 2, \ldots, N\}$ exist and are finite.

Given a policy $\boldsymbol{\mu}$, suppose that the values of $\{\boldsymbol{G}(n),\ n = 1, 2, \ldots\}$ are computed using the above recursive procedure. Based on the difference formula (27) and Theorem 2, we can obtain an improved policy $\boldsymbol{\mu}'$ as follows: for $1 \leq n \leq N$ and $j = 1, 2, \ldots, m$

$$\mu'_{n,j} = \begin{cases} 0, & \text{if } G(n,j) \geq 0; \\ \mu_{n,j}^{\max}, & \text{otherwise.} \end{cases} \tag{63}$$

$$\hat{\boldsymbol{Q}} = \begin{pmatrix} 0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \hat{\boldsymbol{D}}_1\boldsymbol{e} & \hat{\boldsymbol{D}}_0 - \mathrm{diag}(\boldsymbol{\mu}_1) & \mathrm{diag}(\boldsymbol{\mu}_1) & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \hat{\boldsymbol{D}}_1 & \hat{\boldsymbol{D}}_0 - \mathrm{diag}(\boldsymbol{\mu}_2) & \mathrm{diag}(\boldsymbol{\mu}_2) & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{D}}_1 & \hat{\boldsymbol{D}}_0 - \mathrm{diag}(\boldsymbol{\mu}_3) & \mathrm{diag}(\boldsymbol{\mu}_3) & \cdots \\ \vdots & \vdots & \cdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{61}$$

Algorithm 2: An iterative algorithm to find the optimal service rate in MAP/M/1.

1) Initialize system parameters $(m, \boldsymbol{D}_0, \boldsymbol{D}_1)$, $\boldsymbol{\mu}^{\max}$, $\boldsymbol{\phi}$, b, and the initial policy $\boldsymbol{\mu}'$. Determine N, $\boldsymbol{\mu}_\infty^{\max}$, and $d\boldsymbol{\phi}_\infty$.
2) Compute $\tilde{\boldsymbol{A}}_\infty$ and $\boldsymbol{A}_\infty$ by using Algorithm 1. Compute $\boldsymbol{\xi}_\infty$ by using (55).
3) Repeat:
   Let $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}'$;
   Compute $\{\boldsymbol{A}_n,\ n = N, N-1, \cdots, 1\}$ and $\{\boldsymbol{\xi}(n),\ n = N, N-1, \cdots, 1\}$ with (44);
   Set $\boldsymbol{G}(1) = \boldsymbol{\xi}(1)$, compute $\{\boldsymbol{G}_n,\ n = 2, 3, \cdots, N\}$ by using (43);
   Use (63) to generate a new policy $\boldsymbol{\mu}'$;
   Until the stopping criterion $\boldsymbol{\mu} = \boldsymbol{\mu}'$ is satisfied.
4) Output $\boldsymbol{\mu}$ as the optimal policy $\boldsymbol{\mu}^*$.

Therefore, by combining the above computation and optimization procedures, we propose the following Algorithm 2 to find the optimal policy $\boldsymbol{\mu}^*$ of the MAP/M/1 queue.

Algorithm 2 is an iterative algorithm that is similar to the policy iteration in the MDP theory. The convergence of the algorithm can be simply interpreted as follows. After the values of $G(n, j)$ are computed, we use (63) to generate a new policy $\boldsymbol{\mu}'$. With the difference formula (27), we have

$$\eta' - \eta = \sum_{n=1}^{N}\sum_{j=1}^{m}(\mu'_{n,j} - \mu_{n,j})\pi'(n, j)G(n, j) < 0. \quad (64)$$

Note that by comparing the above equation and (27), we find that the summation of n is truncated from infinity to N, which is guaranteed by the quasi-threshold property in Theorem 3. The difficulty in computation due to the infinite dimension of policy space is then avoided. Therefore, the long-run average total cost is reduced after each iteration of Algorithm 2. Since the size of the policy space is simplified to $2^{N\times m}$, which is finite, Algorithm 2 will stop after a finite number of iterations. Theoretically, the worst case is that Algorithm 2 needs $2^{N\times m}$ iterations. In most cases, Algorithm 2 converges to the optimal policy within a very few iterations, which is also demonstrated by the numerical experiments in Section V.

Remark 6: The computation of value functions is the most important and time-consuming step of the policy iteration in MDP. How to efficiently compute value functions is a fundamental problem in the MDP theory and it attracts a lot of research effort, such as the function approximation and approximate dynamic programming [4]. Algorithm 2 utilizes the QBD structure and the MAM techniques to recursively compute the value of $G(n, j)$, which is an analog generated from value functions. Therefore, our approach extends the powerful computation capability of MAM to computing the value function in MDP. A similar idea for computing the value function is utilized in [10].

Remark 7: 1) Although we assume that $d\boldsymbol{\phi}(n) = d\boldsymbol{\phi}_\infty$, for $n \geq N$ in the algorithm, the optimal policy can be found if $d\boldsymbol{\phi}(n) \geq 0$ for $n \geq N$. Intuitively, under the condition $d\boldsymbol{\phi}(n) \geq 0$ for sufficiently large n, the optimal policy $\boldsymbol{\mu}_n^* = \boldsymbol{\mu}_n^{\max}$ for sufficiently large n. Thus, we can choose N sufficiently large in the algorithm to ensure that the resulting policy is optimal. 2) While the requirement on $d\boldsymbol{\phi}(n)$ for large n seems restrictive, there is no condition on $\boldsymbol{\phi}(n)$ and $\boldsymbol{\mu}_n^{\max}$ for small and moderate n. Thus, the application of the algorithm is not too restrictive.

Remark 8: Algorithms 1 and 2 are developed under the assumption that $d\boldsymbol{\phi}(n)$ is constant for $n \geq N$, for some fixed N. The algorithms can be used effectively under much weaker conditions, as long as $d\boldsymbol{\phi}(n)$ does not increase exponentially. The reason is that the tail of the stationary distribution $\boldsymbol{\pi}$ of the QBD process decays geometrically, if $\boldsymbol{\mu}_n = \boldsymbol{\mu}_\infty^{\max}$ for $n \geq N$. For such a case, we choose $\hat{\boldsymbol{\phi}}(n) = \boldsymbol{\phi}(\min\{n, N\})$, for $n \geq 1$. If N is sufficiently large, the difference between the long-run average total costs corresponding to $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ is negligible. Then, we can use $\hat{\boldsymbol{\phi}}$ instead of $\boldsymbol{\phi}$ to conduct Algorithm 1. The policy found by Algorithms 1 and 2 is optimal for both $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}$.

Until now, we have conducted a comprehensive investigation of the service rate control problem in the MAP/M/1 queue. In the rest of this section, we study some special situations where the optimization problem can further be simplified.

C. Optimality of Threshold-Type Policy in M/M/1

We consider a special MAP/M/1 queue in which the MAP has only one phase, i.e., $m = 1$. Thus, it becomes an M/M/1 queue. The arrival process is a Poisson process with rate λ. The system state can be simplified from (n, j) to n. All the scripts with phase status j can be omitted. The service rate is simplified as μ_n, which is also called *load-dependent* service rate. Recall that $\mu_n^{\min} = 0$. The service rate control problem (15) can be rewritten as

$$\boldsymbol{\mu}^* = \underset{\substack{0 \leq \mu_n \leq \mu_n^{\max} \\ n=1,2,\ldots}}{\operatorname{argmin}} \{\eta\}. \quad (65)$$

The performance difference formula (27) is also rewritten as

$$\eta' - \eta = \sum_{n=1}^{\infty}(\mu'_n - \mu_n)\pi'(n)G(n) \quad (66)$$

where $G(n) = g(n-1) - g(n) + b$.

Corollary 2: Since M/M/1 is a special case of MAP/M/1, all the previous results still hold for the service rate control of M/M/1. That is, Lemma 1, the monotonicity in Theorem 1, and the optimality of the bang–bang control in Theorem 2 are also correct for M/M/1.

Furthermore, we can prove the following theorem about the optimality of the threshold-type policy for the service rate control problem (65) in M/M/1.

Theorem 4: For the service rate control of M/M/1, the optimal service rate μ_n^* can have a threshold form. That is, there exists a threshold N such that if $n < N$, $\mu_n^* = 0$; otherwise, $\mu_n^* = \mu_n^{\max}$.

Proof: Since Theorem 2 still holds for M/M/1, we see that the optimal service rate μ_n^* is either 0 or $\mu_n^{\max}$, for all $n = 1, 2, \ldots$. Suppose for a specific n, say k, the optimal service rate is $\mu_k^* = 0$. When the number of customers at the server reaches k, the server will stop and n will increase definitely. That is, the system state n will never drop below k. In other words, all the system states $n < k$ are transient. Since the transient state has no contribution to the long-run average total cost, we can set $\mu_n^* = 0$ for all $n < k$. In summary, for any k with $\mu_k^* = 0$, we have proved that $\mu_n^* = 0$ for all $n < k$. Repeating this process, it is straightforward to show that there exists an N such that if $n < N$, $\mu_n^* = 0$; otherwise, $\mu_n^* = \mu_n^{\max}$. The optimality of the threshold-type policy of the optimal service rate is proved.

Theorem 4 guarantees the optimality of the threshold-type policy. Such a threshold-type policy is also called "N-policy" in the literature of queueing theory. The threshold-type policy is widely used since it is very simple and easy to implement in practice. The optimality of such type policy is of significance for both theoretical study and practical application. With Theorem 4, the optimization complexity of the service rate control problem (65) is reduced from an infinite-dimensional continuous search space $[0, \mu_n^{\max}]^\infty$ to a one-parameter discrete search space $N \in \{1, 2, \ldots\}$. Note that in this section, we do not require that the holding cost should be nondecreasing, which is usually required in the literature [24].

D. Analysis of Threshold-Type Policy in MAP/M/1

Although we have proved the optimality of quasi-threshold-type policy under mild conditions in Theorem 3, the threshold-type policy is not optimal for MAP/M/1 in general. However, since the threshold-type policy is very simple and widely used in practice, we limit our study to the threshold-type policy as a special case in this section. That is, for the service rate control of MAP/M/1, we adjust the service rate $\mu_{n,j}$ only according to the queue length n, which is also called the *load-dependent* service rate μ_n, $n = 1, 2, \ldots$. We have $\mu_{n,j} = \mu_n$, $\mu_{n,j}^{\max} = \mu_n^{\max}$, and $\mu_{n,j}^{\min} = 0$, for all $j = 1, 2, \ldots, m$. On the other hand, since the phase status j of the MAP is internal information, it is usually unobservable. Therefore, it might be infeasible to adjust the service rate considering the internal phase status j. The load-dependent service rate μ_n is more feasible than $\mu_{n,j}$ in the real world.

With the above assumption, the original service rate control problem (15) can be simplified as follows:

$$N^* = \underset{N=1,2,\cdots}{\operatorname{argmin}}\{\eta\}, \quad \text{where } \mu_{n,j} = 0, \text{ if } n < N;$$
$$\mu_{n,j} = \mu_n^{\max}, \text{ if } n \geq N. \quad (67)$$

For the above optimization problem, we can still apply the SBO theory and obtain the difference formula under different threshold-type policies. Suppose the threshold is changed from N to N'. Without loss of generality, we assume $N' > N$. According to the difference formula (27), we have

$$\begin{aligned}\eta' - \eta &= \sum_{n=N}^{N'-1} \sum_{j=1}^{m} (0 - \mu_{n,j}^{\max}) \pi'(n,j) G(n,j) \\ &= -\sum_{n=N}^{N'-1} \mu_n^{\max} \sum_{j=1}^{m} \pi'(n,j) G(n,j). \end{aligned} \quad (68)$$

For the system with threshold N', we see that the system states (n, j) with $n < N' - 1$ are all transient. Therefore, $\pi'(n, j)$ in (68) equals 0 when $n < N' - 1$. We obtain the following simplified version of the performance difference formula when the threshold is changed from N to N', $N' > N$

$$\eta' - \eta = -\mu_{N'-1}^{\max} \sum_{j=1}^{m} \pi'(N'-1, j) G(N'-1, j). \quad (69)$$

Although we derive the above difference formula for the optimization problem (67), we cannot develop a policy iteration algorithm similar to Algorithm 2 to solve it. This is because the load-dependent service rate in (67) makes the action selection correlated at different states, which violates the requirement of a standard MDP model [27]. Fortunately, we can still derive some optimality properties for this optimization problem. When the cost function and the system parameters are specified, we have the following theorem after further analysis.

Theorem 5: If $\mu_{n,j}^{\max}$ are equal for all $n = 1, 2, \ldots,$ $j = 1, 2, \ldots, m$, and the holding cost $\phi(n, j)$ is nondecreasing in n, then the optimal threshold is $N^* = 1$.

Proof: First, we consider a scenario where the threshold is $N = k$ and $k > 1$. We can see that all the states (n, j) with $n < k - 1$ are transient and the server always has more than $k - 1$ customers after the system reaches steady state. The steady-state distribution is denoted as $\pi(n, j)$ and we have $\pi(n, j) = 0$ for any $n < k - 1$. Second, we consider another scenario where the threshold is $N = 1$. The steady-state distribution is denoted as $\pi'(n, j)$. Since all of $\mu_{n,j}^{\max}$ are equal, the queueing system in the first scenario is equivalent to the second scenario, except the number of customers always has a gap $k - 1$. It looks like these two scenarios are the same except that there are $k - 1$ customers always waiting at the server forever in the first scenario. That is, we have $\pi(n + k - 1, j) = \pi'(n, j)$, for all $n = 0, 1, 2, \ldots$ and $j = 1, 2, \cdots, m$. Since the holding cost $\phi(n, j)$ is nondecreasing in n, we have $\phi(n + k - 1, j) \geq \phi(n, j)$. Therefore, we have $f(n + k - 1, j) \geq f(n, j)$ for all n and j. For the long-run average total cost, we have

$$\begin{aligned}\eta &= \sum_{n=0}^{\infty} \sum_{j=1}^{m} \pi(n,j) f(n,j) = \sum_{n=k-1}^{\infty} \sum_{j=1}^{m} \pi(n,j) f(n,j) \\ &= \sum_{n=0}^{\infty} \sum_{j=1}^{m} \pi(n+k-1,j) f(n+k-1,j) \\ &\geq \sum_{n=0}^{\infty} \sum_{j=1}^{m} \pi'(n,j) f(n,j) = \eta', \end{aligned} \quad (70)$$

where η and η' are the long-run average total cost in the first scenario and the second scenario, respectively. Therefore, we have $N^* = 1$ and the theorem is proved. ■

Remark 9: In the literature, it is often required that the holding cost should be nondecreasing. In Section III and Theorem 4, we can let the holding cost be any function in general. In Theorem 3, we assume the holding cost is nondecreasing after a sufficiently large N and we prove the optimality of quasi-threshold-type policy. In Theorem 5, we assume the holding cost is always nondecreasing and we prove the optimal threshold is $N^* = 1$. The conditions in these three scenarios are more

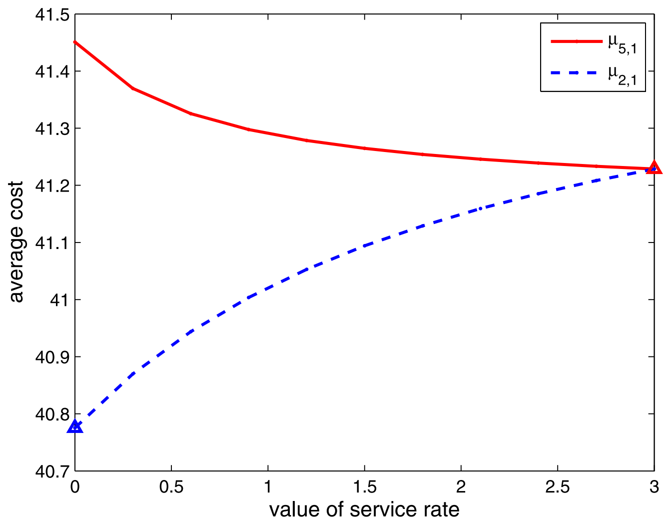


Fig. 1. Curve of the long-run average total cost w.r.t. the service rates.

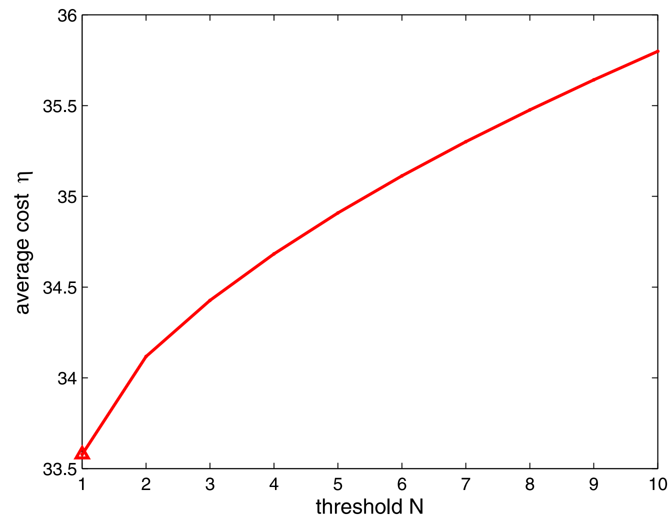


Fig. 2. Curve of the long-run average total cost w.r.t. the threshold when $\phi(n,j)$ is nondecreasing.

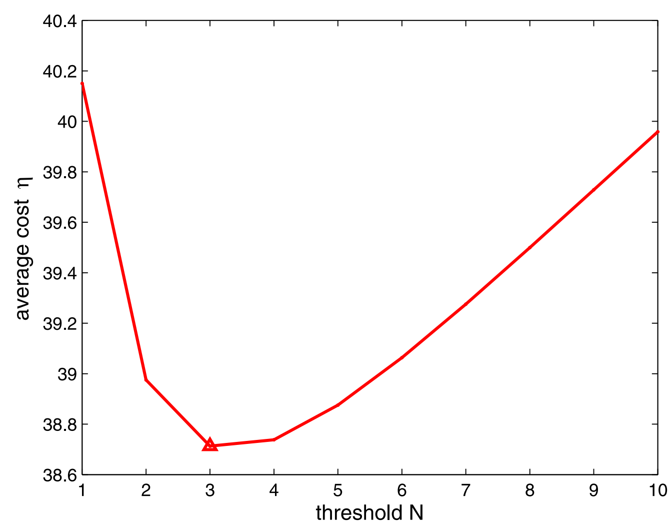


Fig. 3. Curve of the long-run average total cost w.r.t. the threshold when $\phi(n,j)$ is not always nondecreasing.

and more strict, thus the associated optimality properties also become more and more special.

V. Numerical Examples

In this section, we conduct some numerical experiments to gain insights into the queueing model of interest, especially the optimal policy.

Example 1: First, we verify whether the monotonicity property in Theorem 1 is correct in a numerical example. We consider a MAP/M/1 queue with $b = 10$

$$\begin{aligned} &m = 2,\ \boldsymbol{D}_0 = \begin{pmatrix} -2 & 0 \\ 1 & -3 \end{pmatrix},\ \boldsymbol{D}_1 = \begin{pmatrix} 1\,1 \\ 2\,0 \end{pmatrix}; \\ &\boldsymbol{\mu}^{\max} = \begin{pmatrix} 3\,3\,3\,3\,3\,3\,3\cdots \\ 3\,3\,3\,3\,3\,3\,3\cdots \end{pmatrix}; \quad \boldsymbol{\mu}^{\min} = \boldsymbol{0}; \\ &\phi(n,j) = \tfrac{10}{n+1} + 3\sqrt{n} + 10j, \ \ \text{for } n \geq 0,\ j = 1,2. \end{aligned} \tag{71}$$

By routine calculations, we obtain $\boldsymbol{\omega} = (0.75, 0.25)$. Thus, $\lambda = \boldsymbol{\omega}\boldsymbol{D}_1\boldsymbol{e} = 2$.

Without loss of generality, we arbitrarily choose the initial service rates as their maximal values, i.e., set $\mu_{n,j} = 3$ for all $n = 1, 2, \ldots$ and $j = 1, 2$. We arbitrarily choose certain service rates, say $\mu_{2,1}$ and $\mu_{5,1}$, and change their values in the domain $[0, 3]$. After numerical computation, we obtain Fig. 1 that illustrates the curves of η w.r.t. $\mu_{2,1}$ and $\mu_{5,1}$, respectively. The red solid line represents the curve of η w.r.t. $\mu_{5,1}$ and it shows that η is monotonically decreasing w.r.t. $\mu_{5,1}$. The blue dashed line represents the curve of η w.r.t. $\mu_{2,1}$ and it shows that η is monotonically increasing w.r.t. $\mu_{2,1}$. The triangles in Fig. 1 indicate the optimal values of service rates in the current scenario. Therefore, it is demonstrated that the average cost η is monotone w.r.t. the service rate $\mu_{n,j}$ (Theorem 1) and the optimal service rates are either 0 or $\mu^{\max}_{n,j}$ (Theorem 2).

Example 2 (Example 1 continued): We use the same parameter settings as those of Example 1, except we set $\phi(n,j) = \sqrt{n} + 10j$. Therefore, the cost function is always nondecreasing. To verify Theorem 5, we enumerate every threshold policy from $N = 1$ to $N = 10$ and obtain the system average costs under different thresholds. From Fig. 2, we can see that the optimal threshold is $N^* = 1$ and Theorem 5 is verified in this example.

Example 3 (Example 2 continued): As we discussed, Theorem 5 requires that the holding cost $\phi(n,j)$ should be always nondecreasing in n. When this condition is not satisfied, the optimal threshold may not be $N^* = 1$. Below, we use a numerical experiment to demonstrate this statement. We choose the holding cost as $\phi(n,j) = \frac{10}{n+1} + 2\sqrt{n} + 10j$. Other parameters are the same as those of Example 1. Obviously, when n increases from 0 to ∞, $\phi(n,j)$ will first decrease and then increase. We enumerate the threshold from $N = 1$ to $N = 10$ and obtain the long-run average total costs under the different thresholds. From Fig. 3, we can see that the optimal threshold is $N^* = 3$, which is different from the result of Theorem 5.

In Examples 2 and 3, we assume a threshold-type policy. Actually, the optimality of the threshold-type policy cannot be guaranteed for a general MAP/M/1 queue, whereas the optimality of the quasi-threshold-type policy is proved under mild conditions in Theorem 3. In Examples 4–6, we conduct experiments to verify Theorem 3 and Algorithms 1 and 2 for finding the optimal policy.

Example 4: We consider a MAP/M/1 with $b = 10$

$$m=2,\ \boldsymbol{D}_0 = \begin{pmatrix} -0.1 & 0 \\ 0.2 & -3 \end{pmatrix},\ \boldsymbol{D}_1 = \begin{pmatrix} 0.09 & 0.01 \\ 0 & 2.8 \end{pmatrix};$$
$$\boldsymbol{\mu}^{\max} = \begin{pmatrix} 3\ 5\ 4\ 4\ 2\ 3\ 3\ 4\ 2\ 2\ 2 \cdots \\ 8\ 3\ 2\ 5\ 2\ 2\ 5\ 5\ 7\ 7\ 7 \cdots \end{pmatrix};$$
$$\boldsymbol{\mu}^{\min} = \boldsymbol{0};\ \phi(n,j) = \frac{15}{n+1} + 2\sqrt{n} + 30j,\ \text{for } n \geq 0,\ j = 1,2. \tag{72}$$

By routine calculations, we obtain $\boldsymbol{\omega} = (0.9524, 0.0476)$ and $\lambda = 0.2286$. We can derive $d\boldsymbol{\phi}(n) = O(1/\sqrt{n})$ as $n \to \infty$. We also have $\boldsymbol{\mu}_\infty^{\max} = (2,7)^T$, and $d\boldsymbol{\phi}_\infty = (0,0)^T$. In computation, we choose $\hat{\boldsymbol{\phi}}(n) = \boldsymbol{\phi}(\min\{n, N\})$, for $n \geq 1$ and some sufficiently large N, say $N = 35$. As stated by Remark 8, if N is sufficiently large, the difference between cost functions corresponding to $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ is negligible. We can use $\hat{\boldsymbol{\phi}}$ instead of $\boldsymbol{\phi}$ to conduct Algorithm 1, where $d\hat{\boldsymbol{\phi}}_\infty = (0,0)^T$. By using Algorithm 2, we obtain the optimal policy as

$$\boldsymbol{\mu}^* = \begin{pmatrix} 0\ 0\ 0\ 0\ 0\ 3\ 3\ 4\ 2\ 2\ 2 \cdots \\ 0\ 0\ 0\ 0\ 2\ 2\ 5\ 5\ 7\ 7\ 7 \cdots \end{pmatrix}. \tag{73}$$

It is clear that the optimal policy $\boldsymbol{\mu}^*$ is of a quasi-threshold-type. More specifically, $\boldsymbol{\mu}^*$ is of a strong quasi-threshold-type, although it does not satisfy the condition in Corollary 1. If the queue length is 5, whether or not the maximum service rate should be used depends on the phase of the arrival process. We use the MAM to compute the average total cost per unit time. It is clear that the Markov process $\{(N(t), J(t)), t > 0\}$ is a level-independent QBD process. The stationary distribution $\boldsymbol{\pi}$ has a matrix-geometric solution (see [19]). For this case, the long-run average total cost is $\eta^* = 40.6958$.

To gain insights into the queueing model and the optimal policy, we consider the following variants. The main issue under investigation is the impact of the phase on queueing analysis and design. The motivation is that the phase of the MAP is fictitious and unobservable in queueing operation. Thus, applicable policy usually does not depend on the arrival phase. We next introduce some service policies that do not depend on the phase of the arrival process.

1) For the MAP/M/1 queue with policy $\boldsymbol{\mu}$, we construct an M/M/1 model with a Poisson arrival process with arrival rate λ and service rates (policy) $\boldsymbol{\mu}^{M/M/1}$ defined as $\mu_n^{M/M/1} = \min\{(\boldsymbol{\omega\mu})_n, \mu_{n,j}^{\max}, j = 1, \cdots, m\}$, for $n \geq 1$. The M/M/1 model has the same average arrival rate as that of the MAP/M/1 queue. By comparing the MAP/M/1 with policy $\boldsymbol{\mu}$ and the M/M/1 with policy $\boldsymbol{\mu}^{M/M/1}$, we explore the difference between simple and complicated queueing models. Note that the M/M/1 queue with policy $\boldsymbol{\mu}^{M/M/1}$ is a special case of the MAP/M/1 queue with policy $\boldsymbol{\mu}$.
2) For the MAP/M/1 queue, we apply policy $\boldsymbol{\mu}^{M/M/1}$. For policy $\boldsymbol{\mu}^{M/M/1}$, the service rate depends only on the queue length and is independent of the arrival phase. Note that the MAP/M/1 queue with policy $\boldsymbol{\mu}^{M/M/1}$ is a special case of the MAP/M/1 queue with policy $\boldsymbol{\mu}$.
3) For the MAP/M/1 queue with policy $\boldsymbol{\mu}$, we construct a new service policy $\boldsymbol{\mu}^{\boldsymbol{\omega}}$ as follows. Given the queue length n, the service rate is determined by the distribution $\boldsymbol{\omega}$ and $\boldsymbol{\mu}^{\max}$. That is, given $N(t) = n$, with probability $\omega(i)$, the service type, denoted by $I(t)$, is i (i.e., $I(t) = i$, $i = 1, 2, \ldots, m$) and the service rate is $\min\{\mu_{n,I(t)}, \mu_{n,J(t)}^{\max}\}$. Again, we explore the impact of the use of phase in the design of queueing models. For this model, the system can be represented by $\{(N(t), J(t), I(t)), t > 0\}$, where $I(t)$ is the type of service (which can be viewed as a sample of phase $J(t)$ obeying distribution $\boldsymbol{\omega}$). This process is still a QBD process and can be analyzed using MAM.

TABLE I
LONG-RUN AVERAGE TOTAL COSTS FOR EXAMPLE 5

Model (policy)	$\boldsymbol{\mu}$	$\boldsymbol{\mu}^{M/M/1}$ (M/M/1)	$\boldsymbol{\mu}^{M/M/1}$ (MAP/M/1)	$\boldsymbol{\mu}^{\boldsymbol{\omega}}$
η	40.6958	40.7004	40.7373	40.7348

Example 5 (Example 4 continued): For Example 4, we compute the long-run average total costs for the original system and the three cases defined above, which are shown in Table I.

Table I demonstrates that the phase sensitive policies performs better than others. On the other hand, it also shows that the difference of the performance of these policies can be insignificant.

TABLE II
LONG-RUN AVERAGE TOTAL COSTS FOR EXAMPLE 6

Model (policy)	$\boldsymbol{\mu}$	$\boldsymbol{\mu}^{M/M/1}$ (M/M/1)	$\boldsymbol{\mu}^{M/M/1}$ (MAP/M/1)	$\boldsymbol{\mu}^{\boldsymbol{\omega}}$
η	35.2925	38.6545	40.8641	35.6888

Example 6: We consider a MAP/M/1 with $b = 5$, $m = 3$

$$\boldsymbol{D}_0 = \begin{pmatrix} -5 & 1 & 1 \\ 0 & -3 & 0.5 \\ 0.1 & 0 & -1 \end{pmatrix},\ \boldsymbol{D}_1 = \begin{pmatrix} 1 & 0 & 2 \\ 1.5 & 0 & 1 \\ 0.1 & 0.1 & 0.7 \end{pmatrix};$$
$$\mu^{\max} = \begin{pmatrix} 3 & 5 & 4 & 4 & 2 & 3 & 3 & 4 & 2 & 2 & 2 & \cdots \\ 8 & 3 & 2 & 5 & 2 & 2 & 5 & 5 & 7 & 7 & 7 & \cdots \\ 8 & 1 & 5 & 7 & 2 & 3 & 4 & 4 & 4 & 4 & 4 & \cdots \end{pmatrix};$$
$$\boldsymbol{\mu}^{\min} = \boldsymbol{0};\ \phi(n,j) = \frac{55}{n+1} + 2n^2,\ \text{for } n \geq 0,\ j = 1,2,3. \tag{74}$$

By routine calculations, we obtain $\omega = (0.0633,\ 0.0506,\ 0.8861)$ and $\lambda = 1.1139$. We also have $\boldsymbol{\mu}_\infty^{\max} = (2,7,4)^T$, and $d\boldsymbol{\phi}_\infty = (\infty, \infty, \infty)^T$. To handle the difficulty caused by ∞ in the computation of Algorithm 1, similar to Example 4, we choose $\hat{\boldsymbol{\phi}}(n) = \boldsymbol{\phi}(\min\{n, N\})$, for $n \geq 1$, and some sufficiently large N, say $N = 35$. For the cost function $\boldsymbol{\phi}$, we have $d\boldsymbol{\phi}(n) = O(n)$ if $n \to \infty$. As stated by Remark 8, if N is sufficiently large, the difference between long-run average total costs corresponding to $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}$ is negligible. We can use $\hat{\boldsymbol{\phi}}$ instead of $\boldsymbol{\phi}$ to conduct Algorithm 1, where $d\hat{\boldsymbol{\phi}}_\infty = (0,0,0)^T$. By using

Algorithm 2, we obtain the optimal policy as

$$\boldsymbol{\mu}^* = \begin{pmatrix} 0 & 5 & 4 & 4 & 2 & 3 & 3 & 4 & 2 & 2 & 2 & \cdots \\ 0 & 3 & 2 & 5 & 2 & 2 & 5 & 5 & 7 & 7 & 7 & \cdots \\ 0 & 0 & 5 & 7 & 2 & 3 & 4 & 4 & 4 & 4 & 4 & \cdots \end{pmatrix}. \tag{75}$$

The average total costs for the four models are given in Table II.

Example 6 demonstrates, again, that the optimal phase-dependent policy performs better than nonphase dependent policies. For this example, the nonphase policy $\boldsymbol{\mu}^{M/M/1}$ performs significantly poorer than the optimal policy, when it is applied to the MAP/M/1 queue. On the other hand, the sampling policy $\boldsymbol{\mu}^{\boldsymbol{\omega}}$ performs close to the optimal policy. Example 6 demonstrates that the benefit of considering the phase of the underlying process, even if it is fictitious, can be significant in queueing optimization. To handle the unobservability of phase status, we can also resort to other theories, such as the partial observable MDP [4], [11], [20] or the event-based optimization [5], [6], where we make decision based on partial observable states or observable events, instead of complete internal states. The optimization of such problem is much more complicated and it deserves further investigation.

VI. Conclusion

In this paper, we study the service rate control problem of the MAP/M/1 queue. Matrix-analytic methods are used to introduce a CTMC of queueing systems and recursively compute the fundamental quantity $G(n, j)$. The SBO method is used to analyze the optimality properties of the problem and develop an iterative algorithm to find the optimal policy. We obtain a performance difference formula that clearly describes how the long-run average total cost will change according to different service rates. Based on the difference formula, some insightful properties are derived, such as the monotonicity property and the optimality of the bang–bang control. We further prove the optimality of quasi-threshold-type policy under some mild conditions. A recursive computation algorithm and an iterative optimization algorithm are developed to find the optimal policy. The computation algorithm utilizes the powerful techniques of MAM and it solves the fundamental problem of efficiently calculating value functions in MDP. The optimization algorithm utilizes the optimality of quasi-threshold-type policy and it solves the difficulty of the infinite dimensionality of optimization variables in MAP/M/1. For some other special situations, we further study the optimality of threshold-type policy in M/M/1, which can be viewed as a special MAP/M/1 with only one phase. The property of threshold-type policy is also studied for MAP/M/1. Finally, we use some numerical experiments to demonstrate the main results of this paper. The impact of the phase of the customer arrival process on the optimal policy is also studied through numerical comparison.

This paper shows the potentials of combining MAM and SBO theories to study the performance optimization of complicated queueing models. For future work, we may further consider the server assignment problem, where we aim to determine the optimal number of working servers at every system state in a MAP/M/c queue. This problem has similarity to the problem studied in this paper. Another possible future topic is to study the service rate control of MAP/M/1/K with finite buffer and consider customer loss cost in the optimization problem. Some other topics, such as the performance optimization of queueing systems involving MAP arrival, PH service, etc., can further be studied with similar ideas. We hope our paper can shed light on those promising research topics.

References

[1] A. S. Alfa, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System*. New York, NY, USA: Springer, 2010.

[2] J. Artalejo, A. Gomez-Corral, and Q. M. He, "Markovian arrivals in stochastic modelling: A survey and some new results," *Statist. Oper. Res. Trans.*, vol. 34, no. 2, pp. 101–144, 2010.

[3] B. Ata and S. Shneorson, "Dynamic control of an M/M/1 service system with adjustable arrival and service rates," *Manage. Sci.*, vol. 52, pp. 1778–1791, 2006.

[4] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Vol. II (Approximate Dynamic Programming)*, 4th ed. MA, USA: Athena Sci., 2012.

[5] X. R. Cao, "Basic ideas for event-based optimization of Markov systems," *Discrete Event Dyn. Syst. Theory Appl.*, vol. 15, pp. 169–197, 2005.

[6] X. R. Cao, *Stochastic Learning and Optimization—A Sensitivity-Based Approach*, New York, NY, USA: Springer, 2007.

[7] X. R. Cao and H. F. Chen, "Perturbation realization, potentials, and sensitivity analysis of Markov processes," *IEEE Trans. Autom. Control*, vol. 42, pp. 1382–1393, Oct. 1997.

[8] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems–Queueing Theory in Action*. Cambridge, U.K: Cambridge Univ. Press, 2013.

[9] Q. M. He, *Fundamentals of Matrix-Analytic Methods*. New York, NY, USA: Springer, 2014.

[10] Q. M. He, E. M. Jewkes, and J. Buzacott, "An efficient algorithm for computing the optimal replenishment policy for an inventory-production system," *Advances in Matrix-Analytic Methods for Stochastic Models*. New Jersey, USA: Notable Publications, Inc., 1998, pp. 381–402.

[11] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artif. Intell.*, vol. 101, pp. 99–134, 1998.

[12] L. Kleinrock, *Queueing Systems, Vol. I: Theory*. Hoboken, NJ, USA: Wiley, 1975.

[13] R. Kumar, M. E. Lewis, and H. Topaloglu, "Dynamic service rate control for a single server queue with Markov modulated arrivals," *Naval Research Logistics*, vol. 60, pp. 661–677, 2013.

[14] A. Lazar, "Optimal flow control of a class of queueing networks in equilibrium," *IEEE Trans. Autom. Control*, vol. 28, no. 11, pp. 1001–1007, Nov. 1983.

[15] D. J. Ma and X. R. Cao, "A direct approach to decentralized control of service rates in a closed Jackson network," *IEEE Trans. Autom. Control*, vol. 39, no. 7, pp. 1460–1463, Jul. 1994.

[16] H. Minc, *Non-negative Matrix*. New York, NY, USA: Wiley, 1988.

[17] P. Naor, "The regulation of queue size by levying tolls," *Econometrica*, vol. 37, pp. 15–24, 1969.

[18] M. F. Neuts, "A versatile Markovian point process," *J. Appl. Probab.*, vol. 16, no. 4, pp. 764–779, 1979.

[19] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, MD, USA: Johns Hopkins Univ. Press, 1981.

[20] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 1994.

[21] Z. Rosberg, P. Varaiya, and J. Walrand, "Optimal control of service in tandem queues," *IEEE Trans. Autom. Control*, vol. 27, pp. 600–610, 1982.

[22] L. I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*. New York, NY, USA: Wiley, 1998.

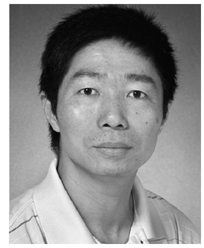
[23] C. Stathis and B. S. Maglaris, "Modeling the self-similar behavior of network traffic," *Comput. Netw.*, vol. 34, no. 1, pp. 37–47, 2000.

[24] S. J. Stidham and R. R. Weber, "Monotonic and insensitive optimal policies for control of queues with undiscounted costs," *Oper. Res.*, vol. 37, no. 4, pp. 611–625, 1989.

[25] S. J. Stidham and R. R. Weber, "A survey of Markov decision models for control of networks of queues," *Queueing Syst.*, vol. 13, pp. 291–314, 1993.
[26] R. R. Weber and S. J. Stidham, "Optimal control of service rates in networks of queues," *Adv. Appl. Probab.*, vol. 19, pp. 202–218, 1987.
[27] L. Xia and Q. S. Jia, "Parameterized Markov decision process and its application to service rate control," *Automatica*, vol. 54, pp. 29–35, 2015.
[28] L. Xia, "Service rate control of closed Jackson networks from game theoretic perspective," *Eur. J. Oper. Res.*, vol. 237, no. 2, pp. 546–554, 2014.
[29] L. Xia, "Event-based optimization of admission control in open queueing networks," *Discrete Event Dyn. Syst., Theory Appl.*, vol. 24, no. 2, pp. 133–151, 2014.
[30] L. Xia and B. Shihada, "Max-Min optimality of service rate control in closed queueing networks," *IEEE Trans. Autom. Control*, vol. 58, pp. 1051–1056, Apr. 2013.
[31] D. D. Yao and Z. Schechner, "Decentralized control of service rates in a closed Jackson network," *IEEE Trans. Autom. Control*, vol. 34, no. 2, pp. 236–240, Feb. 1989.

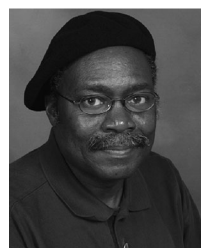
Li Xia (SM'16) received the Bachelor's and Ph.D. degrees in control theory from Tsinghua University, Beijing, China, in 2002 and 2007, respectively.

After graduation, he was with the IBM Research China as a research staff member (2007–2009) and with the King Abdullah University of Science and Technology, Saudi Arabia as a Postdoctoral Research Fellow (2009–2011). Then, he returned to Tsinghua University as a faculty member in 2011. He was a Visiting Scholar in Stanford University, the Hong Kong University of Science and Technology, etc. He is currently an Associate Professor with the Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University. He serves/served as an Associate Editor and program committee member of a number of international journals and conferences. His research interests include the methodology research in stochastic optimization, queueing theory, Markov decision processes, reinforcement learning, and the application research in building energy, energy Internet, production systems, Internet of things, etc.



Qi-Ming He received the Ph.D. degree in Operations Research and Control Theory from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China, in 1989, and the Ph.D. degree in Management Science from the Department of Management Science, University of Waterloo, Waterloo, ON, Canada, in 1996.

He is currently a Professor in the Department of Management Sciences, University of Waterloo. His main research interest include the algorithmic methods in applied probability, queueing theory, inventory control, and production management. In investigating various stochastic models, his favorite methods are matrix analytic methods. He is working on queueing systems with multiple types of customers, inventory systems with multiple types of demands, and representations of phase-type distributions and their applications.



Attahiru Sule Alfa received his Ph.D. degree in transportation & traffic engineering from the University of New South Wales, Sydney, Australia in 1980.

He is currently a Professor of telecommunication systems in the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB, Canada, and also a SARChI Chair Professor in the Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa. He has authored two books, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System* (Springer, 2010) and *Applied Discrete-Time Queue* (Springer, 2015) His research covers, but not limited to, the following areas: queueing theory and applications, optimization, performance analysis and resource allocation in telecommunication systems, modeling of communication networks, analysis of cognitive radio networks, modeling and analysis of wireless sensor networks, developing efficient decoding algorithms for LDPC codes, channel modeling, traffic estimation for the Internet, and cross-layer analysis. He also works in the application of queueing theory to other areas such as transportation systems, manufacturing systems, and healthcare systems.