Cyclic Change of Server's Performance: Impacts and Applications

Boray Huang, Jingui Xie, and Qi-Ming He

Abstract—**This paper studies a stochastic system where the performance of the server changes stochastically and cyclically. We first investigate the performance measures of the system, including the queue length and the overall cost. In particular, we derive an exact expression for the expected length of the renewal cycle, and present closed matrix forms for the mean and variance of the queue length. We then develop an explicit method to tackle a workload control problem, based on an M/G/1 queue approximation. Numerical examples are presented to illustrate the effectiveness of the method.**

Index Terms—**Cyclic change behaviour, M/G/1 queue, optimal workload control, system performance.**

NOMENCLATURE

λ	Arrival rate.
l	$\{0, 1, \ldots, l\}$ service stages.
S_i	Service time at stage i $(0 \leq i \leq l)$.
$B_i(\cdot)$	Distribution function of service time S_i.
μ_i	Service rate at stage i.
p_i	Server transition probability from stage i to $i+1$ or from stage $i = l$ to 0.
Y_n	Service time of the nth job.
Y_∞	Service time of an arbitrary job.
$X(t)$	Number of jobs in the system at time t.
N	$\lim_{t\to\infty} X(t)$.
X_n	Number of jobs in the system immediately after the nth departure.
X_a	Number of jobs in the system seen by an arbitrary arriving job.
X_∞	$\lim_{n\to\infty} X_n$.
$\widetilde{N}$	Queue length in the $\mathrm{M/G_{app}/1}$ system in steady state.
I_n	Stage of the server immediately after the nth departure.
I_∞	$\lim_{n\to\infty} I_n$.
L_C	Length of a renewal cycle.
α	$\left(\sum_{i=0}^{l}(p_i\mu_i)^{-1}\right)^{-1}$
β	$\left(\sum_i p_i^{-1}\right)^{-1}$.
$\overline{\mu}$	$\alpha\beta^{-1}$ average service rate.
ρ_i	λ/μ_i.
ρ	$\lambda/\overline{\mu}$ system traffic intensity.
P	Transition probability matrix of the embedded Markov chain at departure epochs.
A	$\sum_{i\geq 0} A_i$, an irreducible stochastic matrix.
G	G-matrix of the embedded Markov chain.
$\boldsymbol{w}$	$\beta(p_0^{-1}, p_1^{-1}, \cdots, p_l^{-1})$ invariant vector of A, that is, $\boldsymbol{w}A = \boldsymbol{w}$, $\boldsymbol{w}\boldsymbol{e} = 1$.
$\boldsymbol{\eta}$	$\sum_{i=1}^{\infty} iA_i\boldsymbol{e} = (\rho_0, \cdots, \rho_l)^T$, $\rho = \boldsymbol{w}\boldsymbol{\eta}$.
$\boldsymbol{g}$	the invariant probability vector of G, i.e., $\boldsymbol{g}G = \boldsymbol{g}$, $\boldsymbol{g}\boldsymbol{e} = 1$.
$\boldsymbol{v}_0$	$(1-\rho)\boldsymbol{g}$ boundary steady-state probability of embedded Markov chain, $\boldsymbol{v}_0\boldsymbol{e} = 1 - \rho$.
H	$(I - A + \boldsymbol{e}\boldsymbol{w})^{-1}$. (Note: $\boldsymbol{w}H = \boldsymbol{w}$).

I. INTRODUCTION

IN manufacturing and service systems, a machine (or server) often has a cyclic change in its performance. The cyclic behavior may or may not follow a typical pattern. In some cases, the change of the server's performance is monotonic. For instance, a machine may need a warm-up period after a startup to achieve its regular throughput rate. A server's performance may also decrease over time due to fatigue or wear. These phenomena are not uncommon in practice. (See [1] and [13] for examples in data-transmission systems and semiconductor fabrication processes.) In other cases, the cyclic pattern of the server's performance is not strictly monotonic. Reverse bathtub functions are often used to describe the effective production

Manuscript received June 06, 2012; revised September 30, 2012, May 30, 2013, September 30, 2013, and October 17, 2013; accepted October 17, 2013. Date of publication October 24, 2013; date of current version February 19, 2014. This work was supported in part by the Singapore Ministry of Education Academic Research Fund FRC-R-266-000-050-133 and in part by NSFC under Grant 71201154, NSFC major program under Grant 71090401/71090400, and CPSF under Grants 2012M521260 and 2013T60627. Recommended by Associate Editor Loo Hay Lee.

B. Huang is with the Department of Industrial and Systems Engineering, National University of Singapore, Singapore 117576 (e-mail: isehb@nus.edu.sg).

J. Xie is with the School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China (e-mail: xiej@ustc.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).

Digital Object Identifier 10.1109/TAC.2013.2287111

rate over time in reliability literature. A deteriorated system may also receive *imperfect* maintenances and partially restore its capacity. But it would not go back to as-new status until it is replaced [21]. In each cycle of these systems, there would be some increase of performance over a generally deteriorating process. This paper is interested in the server's cyclic behavior and its impact on the system's overall outcomes, as well as the optimal control of system workload. The performance measures are derived by matrix-analytic methods (e.g., [17]). Based on an M/G/1 approximation, we also find an explicit method to tackle the optimization problem. Finally, both methodologies are illustrated in the analysis of two typical systems: a learning system and a deteriorating system.

A variety of changing patterns in the server's performance of a queueing system has been studied for a long time. For example, [33] studies an M/M/1 queue in which there are two exponentially distributed phases, and the arrival and service rates depend on the phases. (See [32] for an extension.) In [16], the model is generalized to an M/G/1 multistate system, which deviates from [33] by assuming the service time distribution depends only on the phase at the beginning of each service. In [9] and [10], a more general model is analyzed: an M/G/1 queue in a semi-Markovian environment. Reference [18] studies a queueing model with a server that changes its service rate according to a finite birth and death process. There are many other studies considering similar queues (e.g., [3], [6], [8], [11], [20], [23], [24], [26], [29]). However, very few studies focus on the *cyclic* change of the server's performance, or the optimal control of the system workload.

The rest of this paper is organized as follows. In Section II, we construct the queueing model with semi-Markov service times to describe the server's cyclic behavior and introduce the optimal workload control problem. In Section III, we apply matrix-analytic methods to analyze the queueing system. In Section IV, we discuss the optimization problem and provide an approximation method with explicit formulas. Examples are given in Section V. Conclusions are made in Section VI.

II. Queueing Model

Assume that jobs arrive according to a Poisson process with arrival rate λ. Then, the interarrival times are independent and exponentially distributed with a parameter λ. There is only one server in the system, and the service discipline is first come first served (FCFS). The service time depends on server status which changes according to a finite cyclic Markov chain.

The server status has, in total, $l+1$ stages $\{0, 1, \ldots, l\}$. The server only changes its status right after the completion of a service. If the server is at stage j $(0 \leq j < l)$ at the end of a service, its state changes to $j+1$ with probability p_j $(0 < p_j \leq 1)$, or remains j with probability $1-p_j$. If the server's current stage is l, the next stage is either l with probability $1-p_l$, or 0 with probability p_l. We call the transition of the server status from l to 0 a *server renewal*, which corresponds to a server replacement or repair in practice. The duration between two consecutive server renewals, that is, the period from stages 0 to 1, 1 to 2, ..., and l to 0, is called a *renewal cycle*, which is illustrated in Fig. 1. It is easy to see that transitions of the server status can be described by a Markov chain with $l+1$ states and transition probability matrix

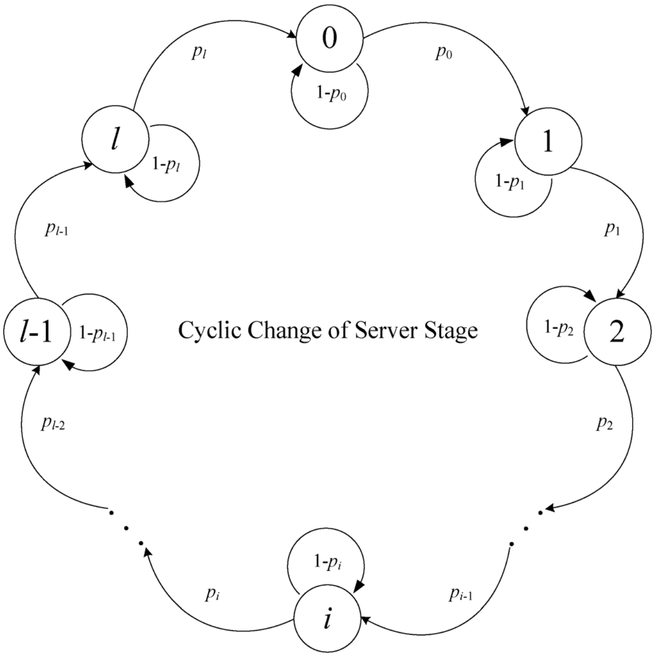


Fig. 1. Renewal cycle with multiple service stages.

$$A = \begin{pmatrix} 1-p_0 & p_0 & & & \\ & 1-p_1 & p_1 & & \\ & & \ddots & \ddots & \\ p_l & & & & 1-p_l \end{pmatrix}. \tag{1}$$

If the server is at stage j, the service time S_j is generally distributed with a distribution function $B_j(\cdot)$ whose Laplace–Stieltjes transform (LST) is $B_j^*(\cdot)$, for $j = 0, 1, \ldots, l$. Let $1/\mu_j$ be the mean of S_j. Then, μ_j is called the *service rate*. Note that the server status remains the same during a service, so the distribution function does not change during the service. We assume that all service times are independent random variables. The service times are also independent of the Poisson arrival process.

Let $X(t)$ be the number of jobs in the system at time t (≥ 0). Assume that a fixed cost c is incurred when the server is renewed/replaced, a holding cost h is charged per job in the system per-unit time, and a reward r is earned after each successful service completion. Denote $D(t)$ and $R(t)$ as the number of service completions and the number of server renewals in $[0, t]$, respectively. Given the system workload λ, the system long-run average profit can be written as

$$J(\lambda) = \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}\left[rD(T) - cR(T) - h\int_0^T X(t)dt \right]. \tag{2}$$

One goal of this paper is to find the optimal workload λ^* such that the long-run average profit is maximized.

III. Properties of the Queueing Model

To analyze the performance of the system, we consider an embedded Markov chain associated with the queue length observed by departing jobs. Denote X_n and I_n as the number of jobs in the system (i.e., the queue length) and the server stage immediately after the departure of the nth job, respectively. Denote ξ_n as the number of new arrivals during the service of the nth job. Then, we have

$$X_{n+1} = \begin{cases} X_n - 1 + \xi_{n+1}, & \text{if } X_n > 0, \\ \xi_{n+1}, & \text{if } X_n = 0. \end{cases} \tag{3}$$

It is easy to see that $\{I_n, n \geq 0\}$ is a Markov chain with a transition probability matrix A defined in (1).

It can be verified that the embedded Markov chain $\{(X_n, I_n), n \geq 0\}$ is irreducible. Denote its transition probability matrix and stationary distribution vector by P and $\boldsymbol{v} = (\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots)$, respectively, where $\boldsymbol{v}_i = (v_{i,0}, v_{i,1}, \cdots, v_{i,l})$, for $i \geq 0$. If the Markov chain is irreducible and positive recurrent, we must have $\boldsymbol{v}P = \boldsymbol{v}$ and $\boldsymbol{ve} = 1$, where $\boldsymbol{e}$ is the column vector with all elements equal to one (the column vector $\boldsymbol{e}$ is so defined throughout the paper with an appropriate dimension). Without loss of generality, the states $\{(i,j), i \geq 0, 0 \leq j \leq l\}$ are ordered lexicographically, where a level i consists of state $\{(i,j), 0 \leq j \leq l\}$. The embedded Markov chain is then an M/G/1-type Markov chain with transition probability matrix

$$P = \begin{pmatrix} A_0 & A_1 & A_2 & \cdots \\ A_0 & A_1 & A_2 & \cdots \\ & A_0 & A_1 & \cdots \\ & & \ddots & \ddots \end{pmatrix} \tag{4}$$

where

$$A_i = \begin{pmatrix} a_0(i)(1-p_0) & a_0(i)p_0 & & \\ & a_1(i)(1-p_1) & a_1(i)p_1 & \\ & & \ddots & \ddots \\ a_l(i)p_l & & & a_l(i)(1-p_l) \end{pmatrix} \tag{5}$$

and

$$a_j(i) = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^i}{i!} dB_j(t), i \geq 0, 0 \leq j \leq l. \tag{6}$$

It is easy to verify that $A = \sum_{i \geq 0} A_i$. It can be shown that $\boldsymbol{w} = \beta(p_0^{-1}, p_1^{-1}, \cdots, p_l^{-1})$ is the invariant probability vector of A, that is, $\boldsymbol{w}A = \boldsymbol{w}$ and $\boldsymbol{we} = 1$, where $\beta = (\sum_j p_j^{-1})^{-1}$. Define $\boldsymbol{\eta} = \sum_{i=1}^\infty iA_i\boldsymbol{e}$ and $\rho_j = \lambda/\mu_j, j = 0, 1, \cdots, l$. We can show that $\boldsymbol{\eta} = (\rho_0, \cdots, \rho_l)^T$, where the superscript "T" means a transpose of matrices or vectors. By [17, Theor. 3.2.1], the Markov chain is positive recurrent if and only if

$$\rho = \boldsymbol{w\eta} < 1. \tag{7}$$

A. Renewal Cycle

We first analyze the renewal cycles during the busy periods of the queueing system. Suppose there are plenty of jobs waiting for services all the time. The server is therefore always busy. Denote Y_n as the service time of the nth job, and I_n^Y as the server renewal stage at the beginning of the service of the nth job, respectively, where $n \geq 1$. We have $Y_n = S_{I_n^Y}$. If $I_n^Y = j$, the distribution function of the service time Y_n is $B_j(\cdot)$, that is, $B_j(x) = \mathbb{P}\{Y_n \leq x \mid I_n^Y = j\}, x > 0$. The random variables $\{Y_1, \cdots, Y_n\}$ are conditionally independent given the values $\{I_1^Y, \cdots, I_n^Y\}$, for all $n \geq 1$. It is easy to see that $\{I_n^Y, n \geq 1\}$, similar to $\{I_n, n \geq 0\}$, is a Markov chain with a transition probability matrix A given in (1).

When the server is at stage j $(0 \leq j \leq l)$, the number of services that are completed before the server changes its status is geometrically distributed with a parameter p_j. Therefore, once the server enters stage i, the average number of services completed at stage j is p_j^{-1}, and the average time spent at stage j is $p_j^{-1}\mathbb{E}[S_j] = 1/(p_j\mu_j)$. Denote L_B and N_B as the total length and the number of services completed in a renewal cycle, respectively, since the server is always busy. We obtain the average length of renewal cycle $\mathbb{E}[L_B] = \sum_{j=0}^l (p_j\mu_j)^{-1}$ and the average number of services completed in a renewal cycle $\mathbb{E}[N_B] = \sum_{j=0}^l (p_j)^{-1} = \beta^{-1}$. Let $\alpha = 1/\sum_{j=0}^l (p_j\mu_j)^{-1}$, so $\mathbb{E}[L_B] = \alpha^{-1}$.

Let Y_∞ be the service time of an arbitrary job. Then, with the renewal theory, the average service time per job is $\mathbb{E}[Y_\infty] = \mathbb{E}[L_B]/\mathbb{E}[N_B] = \alpha^{-1}\beta$. Denote $\overline{\mu}$ as the average service rate. Then, we have $\overline{\mu} = \alpha\beta^{-1}$. With the definition of ρ in (7), we have $\rho = \lambda/\overline{\mu}$, which can be interpreted as the system traffic intensity. We assume that $\rho < 1$ throughout this paper to ensure that the queueing system is stable.

Now we consider the unconditional renewal cycles. Denote by L_C the length of an arbitrary renewal cycle. A simple relationship between the averages of L_C and L_B is established in the following theorem.

Theorem III.1: The average length of a (arbitrary) renewal cycle is given by

$$\mathbb{E}[L_C] = \mathbb{E}[L_B] + \frac{1-\rho}{\lambda\beta} = \mathbb{E}[L_B]\rho^{-1} = (\alpha\rho)^{-1}. \tag{8}$$

Proof: Define $(X_\infty, I_\infty) = \lim_{n\to\infty}(X_n, I_n)$, which exists under $\rho < 1$. It is easy to see that the distribution of (X_∞, I_∞) is $\boldsymbol{v}$ and the distribution of I_∞ is $\boldsymbol{w}$. Denote by τ_j the length of the interval between two consecutive departures, given that the server stage is j right after the first departure. Then, $\tau_j = I_{\{X_\infty > 0 | I_\infty = j\}} S_j + I_{\{X_\infty = 0 | I_\infty = j\}}(\Lambda + S_j) = S_j + I_{\{X_\infty = 0 | I_\infty = j\}}\Lambda$, where Λ has an exponential distribution with the parameter λ, and $I_{\{.\}}$ is the indicator function, and

$$\begin{aligned} \mathbb{E}[\tau_j] &= \mu_j^{-1} + \lambda^{-1} P\{X_\infty = 0 \mid I_\infty = j\} \\ &= \mu_j^{-1} + \lambda^{-1} \frac{P\{X_\infty = 0, I_\infty = j\}}{P\{I_\infty = j\}} \\ &= \mu_j^{-1} + \frac{v_{0,j}p_j}{\lambda\beta}. \end{aligned}$$

Therefore, the average length of a renewal cycle is given by

$$\begin{aligned} \mathbb{E}[L_C] &= \sum_{j=0}^l p_j^{-1} E(\tau_j) = \sum_{j=0}^l (p_j\mu_j)^{-1} + \frac{\boldsymbol{v}_0\boldsymbol{e}}{\lambda\beta} \\ &= \mathbb{E}[L_B] + \frac{1-\rho}{\lambda\beta} = \frac{1-\rho+\lambda\alpha^{-1}\beta}{\lambda\beta} \\ &= (\alpha\rho)^{-1}. \end{aligned}$$

■

Theorem III.1 shows that $\mathbb{E}[L_C] = \mathbb{E}[L_B]/\rho$, where ρ is the traffic intensity of the queueing system. It indicates that: When the traffic intensity is low (high), on average, the renewal cycle is longer (shorter). The reason is that when the traffic intensity is low, the probability that the server becomes idle is large. Since the server does not change its stage while idling, the more time it spends on the idle period, the longer the renewal cycle is.

B. Stationary Distribution

In the theory of Markov chains of M/G/1 type [17], the solution of the invariant probability vector $\boldsymbol{v}$ relies on a matrix G, which is the minimal non-negative solution to the matrix equation

$$G = \sum_{i=0}^{\infty} A_i G^i. \tag{9}$$

For a regular M/G/1-type Markov chain, [22] has proposed a stable recursive algorithm for the calculation of the vector $\boldsymbol{v}_i$. By applying the algorithm to our model, we have

$$\boldsymbol{v}_i = \left(\boldsymbol{v}_0 U_{i-1} + \sum_{j=1}^{i-1} \boldsymbol{v}_j U_{i-j} \right) (I - U_0)^{-1},\ \forall i \geq 1 \tag{10}$$

where I is the identity matrix, and

$$U_i = \sum_{k=i+1}^{\infty} A_k G^{k-i-1},\ \forall i \geq 0. \tag{11}$$

It is well known that $\boldsymbol{v}_0$ is the invariant vector of G for an M/G/1-type Markov chain where the first two rows in P are identical. With this and standard arguments for M/G/l-type Markov chains, the mean recurrent time to the boundary level (the level 0) is obtained by $(1-\rho)^{-1}$. We then have $\boldsymbol{v}_0 = (1-\rho)\boldsymbol{g}$, where $\boldsymbol{g}$ denotes the invariant probability vector of G, that is, $\boldsymbol{g}G = \boldsymbol{g}$ and $\boldsymbol{g}\boldsymbol{e} = 1$ (see [28]).

Re-partitioning the state space by the server's stages, we have $\boldsymbol{v} = (\overline{\boldsymbol{v}}_0, \overline{\boldsymbol{v}}_1, \cdots, \overline{\boldsymbol{v}}_l)$, where $\overline{\boldsymbol{v}}_j = (v_{0,j}, v_{1,j}, \cdots)$, for $0 \leq j \leq l$. The corresponding transition probability matrix can be written as

$$\overline{P} = \begin{pmatrix} (1-p_0)P_0 & p_0 P_0 & & \\ & (1-p_1)P_1 & p_1 P_1 & \\ & & \ddots & \ddots \\ p_l P_l & & & (1-p_l)P_l \end{pmatrix} \tag{12}$$

where, for $0 \leq j \leq l$

$$P_j = \begin{pmatrix} a_j(0) & a_j(1) & a_j(2) & \cdots \\ a_j(0) & a_j(1) & a_j(2) & \cdots \\ & a_j(0) & a_j(1) & \cdots \\ & & \ddots & \ddots \end{pmatrix}. \tag{13}$$

Lemma III.1: For $0 \leq j \leq l$, we have $\overline{\boldsymbol{v}}_j \boldsymbol{e} = p_j^{-1}\beta$, where $\beta = (\sum_{k=0}^{l} p_k^{-1})^{-1}$.

Proof: $\overline{\boldsymbol{v}}_j \boldsymbol{e} = \sum_i v_{i,j} = \sum_i P\{X_\infty = i, I_\infty = j\} = P\{I_\infty = j\} = p_j^{-1}\beta$. This completes the proof. ■

Define X_a and I_a as the queue length (i.e., the number of jobs in the system) and the server renewal stage seen by an arbitrary arriving job, respectively. Denote by N the queue length at an arbitrary time, that is, $N = \lim_{t\to\infty} X(t)$, and L as the server renewal stage at an arbitrary time. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \cdots)$ be the joint stationary distribution vector of the queue length and the server renewal stage, where $\boldsymbol{\pi}_n = (\pi_{n,0}, \pi_{n,1}, \cdots, \pi_{n,l}), \forall n \geq 0$, and $\pi_{n,j} = P\{N = n, L = j\}, n \geq 0, 0 \leq j \leq l$. We have the following lemma showing that the queue length seen by arrivals, departures, or at arbitrary time, has the same distribution.

Lemma III.2: $\boldsymbol{\pi}_i \boldsymbol{e} = \boldsymbol{v}_i \boldsymbol{e}$, for $i \geq 0$. Especially, we have $\pi_{0,j} = v_{0,j}$, for $0 \leq j \leq l$.

Proof: With the Poisson arrivals see time averages (PASTA) property of Poisson arrivals [31], X_a and N have the same probability distribution, that is, $X_a \sim N$. Since the events $\{X_a = i\}$ and $\{X_\infty = i\}$ occur pairwisely, we have $P\{X_a = i\} = P\{X_\infty = i\}$. Hence, N and X_∞ have the same distribution, that is, $N \sim X_\infty$, which is known as the level crossing property in queues [7]. As a result, $P\{N = i\} = \boldsymbol{\pi}_i \boldsymbol{e} = P\{X_\infty = i\} = \boldsymbol{v}_i \boldsymbol{e}, \forall i \geq 0$. Similarly, the events $\{X_\infty = 0, I_\infty = j\}$ seen by a departing job will be soon observed by a new arriving job $\{X_a = 0, I_a = j\}$ since both events occur pairwisely. With the PASTA property, we have $P\{X_\infty = 0, I_\infty = j\} = v_{0,j} = P\{N = 0, L = j\} = \pi_{0,j}$, for $0 \leq j \leq l$. ■

C. Mean and Variance of Queue Length

Define matrix $C_k = diag\{a_0(k), \cdots, a_l(k)\}, \forall k \geq 0$. We have $A_k = C_k A, \forall k \geq 0$. Define generating functions $V^*(z) = \sum_{i=0}^{\infty} \boldsymbol{v}_i z^i$ and $A^*(z) = \sum_{k=0}^{\infty} A_k z^k$. First, we have $A^*(z) = \sum_{k=0}^{\infty} A_k z^k = \sum_{k=0}^{\infty} C_k z^k A = P^*(z)A$, where $P^*(z) = diag\{B_0^*(\lambda(1-z)), \cdots, B_l^*(\lambda(1-z))\}$ is a diagonal matrix.

To further study the queue length N at an arbitrary time, we again use the embedded Markov chain at departures (i.e., (X_∞, I_∞)), whose steady-state equations are

$$\boldsymbol{v}_i = \boldsymbol{v}_0 A_i + \sum_{k=1}^{i+1} \boldsymbol{v}_k A_{i-k+1},\ \forall i \geq 0. \tag{14}$$

Equation (14) leads to the standard equation for the generating functions (see [17]), for $0 \leq z \leq 1$

$$V^*(z)\left(zI - A^*(z)\right) = \boldsymbol{v}_0 (z-1) A^*(z) \tag{15}$$

which can be simplified to

$$V^*(z)\left(zI - P^*(z)A\right) = (1-\rho)(z-1)\boldsymbol{g}P^*(z)A.$$

Differentiating both sides of (15) n times with respect to z and setting $z = 1^-$, we obtain

$$n\boldsymbol{v}_0 A^{*(n-1)}(1^-) = V^{*(n)}(1^-)(I-A) + nV^{*(n-1)}(1^-) - \sum_{i=1}^{n} \binom{n}{i} V^{*(n-i)}(1^-) A^{*(i)}(1^-). \tag{16}$$

Adding $V^{*(n)}(1^-)\boldsymbol{e}\boldsymbol{w}$ to both sides leads to

$$\begin{aligned} &V^{*(n)}(1^-) \\ &= L_n \boldsymbol{w} + n\boldsymbol{v}_0 A^{*(n-1)}(1^-) H \\ &+ \sum_{i=1}^{n} \binom{n}{i} V^{*(n-i)}(1^-) \times A^{*(i)}(1^-) H - nV^{*(n-1)}(1^-) H \end{aligned} \tag{17}$$

where $L_n = V^{*(n)}(1^-)\boldsymbol{e}$ and $H = (I - A + \boldsymbol{ew})^{-1}$. The invertibility of $I - A + \boldsymbol{ew}$ can be shown routinely. Postmultiplying both sides of (17) by $\boldsymbol{\eta}$, we have

$$V^{*(n)}(1^-)\boldsymbol{\eta} = \rho L_n + \theta_n \tag{18}$$

where

$$\theta_n = \left[n\boldsymbol{v}_0 A^{*(n-1)}(1^-) + \sum_{i=1}^{n} \binom{n}{i} V^{*(n-i)}(1^-)A^{*(i)}(1^-) \right. \\ \left. - nV^{*(n-1)}(1^-) \right] H\boldsymbol{\eta}.$$

To determine L_n, we write (16) for $n+1$ and postmultiply both sides by $\boldsymbol{e}$. This yields

$$(n+1)L_n = (n+1)\boldsymbol{v}_0\boldsymbol{\eta}_n + \sum_{i=1}^{n+1} \binom{n+1}{i} V^{*(n+1-i)}(1^-)\boldsymbol{\eta}_i \tag{19}$$

where $\boldsymbol{\eta}_i = A^{*(i)}(1^-)\boldsymbol{e}, \forall i \geq 1$ (note that $\boldsymbol{\eta}_1 = \boldsymbol{\eta}$). Together with (18) and (19), finally, we obtain for $n \geq 1$

$$L_n = \frac{(n+1)(\theta_n + \boldsymbol{v}_0\boldsymbol{\eta}_n) + \sum_{i=2}^{n+1} \binom{n+1}{i} V^{*(n+1-i)}(1^-)\boldsymbol{\eta}_i}{(n+1)(1-\rho)}. \tag{20}$$

Define $G_N(z) = \sum_{i \geq 0} z^i P\{N = i\}$ as the generating function of the queue length at an arbitrary time. By Lemma III.2, we know that the stationary queue length seen at departures and the one at an arbitrary time have the same distribution. Hence, we have $G_N^{(n)}(z) = V^{*(n)}(z)\boldsymbol{e}$. The expectation and variance of the stationary queue length are then found in the following theorem.

Theorem III.2: The expectation and variance of the stationary queue length are given by

$$\mathbb{E}[N] = \frac{(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2}{1-\rho} + \frac{\boldsymbol{w}D_2\boldsymbol{e}}{2(1-\rho)}, \tag{21}$$

$$Var(N) = \frac{\sum_{i=1}^{2}(\theta_i + \boldsymbol{v}_0 D_i\boldsymbol{e}) + V^{*(1)}(1^-)D_2\boldsymbol{e}}{1-\rho} \\ + \sum_{i=2}^{3} \frac{\boldsymbol{w}D_i\boldsymbol{e}}{i(1-\rho)} - (\mathbb{E}[N])^2. \tag{22}$$

where $D_i = \lambda^i diag\{\mathbb{E}[S_0^i], \cdots, \mathbb{E}[S_l^i]\}$, $i \geq 0$ and θ_1, and θ_2, $V^{*(1)}(1^-)$ are given by (24)–(26) in the following proof.

Proof: From (20), we obtain

$$\mathbb{E}[N] = G_N^{(1)}(1^-) = L_1 = \frac{2(\theta_1 + \boldsymbol{v}_0\boldsymbol{\eta}) + V^*(1^-)\boldsymbol{\eta}_2}{2(1-\rho)}. \tag{23}$$

Since $A^{*(n)}(z) = P^{*(n)}(z)A$, we have $A^{*(n)}(1^-) = D_nA$, $n \geq 0$. Since $V^*(1^-) = \boldsymbol{w}$ and $\boldsymbol{w}H = \boldsymbol{w}$, we have

$$\begin{aligned} \theta_1 &= \left[\boldsymbol{v}_0 A^*(1^-) + V^*(1^-)A^{*(1)}(1^-) - V^*(1^-)\right] H\boldsymbol{\eta} \\ &= (\boldsymbol{v}_0 A + \boldsymbol{w}D_1A - \boldsymbol{w})\, H\boldsymbol{\eta} \\ &= (\boldsymbol{v}_0 A + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho. \end{aligned} \tag{24}$$

Therefore, we have

$$\begin{aligned} \mathbb{E}[N] &= \frac{\theta_1 + \boldsymbol{v}_0\boldsymbol{\eta}}{1-\rho} + \frac{\boldsymbol{w}D_2A\boldsymbol{e}}{2(1-\rho)} \\ &= \frac{\theta_1 + (\boldsymbol{v}_0 - \boldsymbol{v}_0A + \boldsymbol{v}_0\boldsymbol{ew})H\boldsymbol{\eta}}{1-\rho} + \frac{\boldsymbol{w}D_2\boldsymbol{e}}{2(1-\rho)} \\ &= \frac{(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2}{1-\rho} + \frac{\boldsymbol{w}D_2\boldsymbol{e}}{2(1-\rho)}. \end{aligned}$$

From (17), we have

$$\begin{aligned} &V^{*(1)}(1^-) \\ &\quad = L_1\boldsymbol{w} + \left[\boldsymbol{v}_0A^*(1^-) + V^*(1^-)A^{*(1)}(1^-) - V^*(1^-)\right] H \\ &\quad = L_1\boldsymbol{w} + [\boldsymbol{v}_0A + \boldsymbol{w}D_1A - \boldsymbol{w}]H \\ &\quad = (L_1 - 1)\boldsymbol{w} + (\boldsymbol{v}_0 + \boldsymbol{w}D_1)AH \end{aligned} \tag{25}$$

and hence

$$\begin{aligned} \theta_2 &= \left[2\boldsymbol{v}_0A^{*(1)}(1^-) + 2V^{*(1)}(1^-)A^{*(1)}(1^-)\right] H\boldsymbol{\eta} \\ &\quad + \left[V^*(1^-)A^{*(2)}(1^-) - 2V^{*(1)}(1^-)\right] H\boldsymbol{\eta} \\ &= \left[2\boldsymbol{v}_0D_1A + 2V^{*(1)}(1^-)(D_1A - I) + \boldsymbol{w}D_2A\right] H\boldsymbol{\eta}. \end{aligned} \tag{26}$$

From (20), then we have

$$\begin{aligned} L_2 &= \frac{3(\theta_2 + \boldsymbol{v}_0\boldsymbol{\eta}_2) + 3V^{*(1)}(1^-)\boldsymbol{\eta}_2 + V^*(1^-)\boldsymbol{\eta}_3}{3(1-\rho)} \\ &= \frac{\theta_2 + \boldsymbol{v}_0D_2\boldsymbol{e} + V^{*(1)}(1^-)D_2\boldsymbol{e}}{1-\rho} + \frac{\boldsymbol{w}D_3\boldsymbol{e}}{3(1-\rho).} \end{aligned}$$

As a result, the variance of queue length is given by

$$\begin{aligned} Var(N) &= G_N^{(2)}(1^-) + G_N^{(1)}(1^-) - \left(G_N^{(1)}(1^-)\right)^2 \\ &= L_2 + L_1 - L_1^2 \\ &= \frac{\sum_{i=1}^{2}(\theta_i + \boldsymbol{v}_0D_i\boldsymbol{e}) + V^{*(1)}(1^-)D_2\boldsymbol{e}}{1-\rho} \\ &\quad + \sum_{i=2}^{3} \frac{\boldsymbol{w}D_i\boldsymbol{e}}{i(1-\rho)} - (\mathbb{E}[N])^2. \end{aligned}$$

The proof is completed. ■

In Theorem III.2, all variables are explicit, except the vector $\boldsymbol{v}_0$. To find $\boldsymbol{v}_0$, the matrix G must be numerically obtained (recall that $\boldsymbol{v}_0 = (1-\rho)\boldsymbol{g}$, where $\boldsymbol{g}G = \boldsymbol{g}$, $\boldsymbol{ge} = 1$). In general, (9) can be used to generate a sequence of matrices that converge to G, see [14] for an efficient algorithm. G also can be computed more quickly using more advanced algorithms (see, for example, the SMC Solver tool of [2]). In fact, in this particular case, it may be possible to exploit the structure of matrices A_i $(i \geq 0)$ to develop more efficient computing algorithms for G, which is beyond the focus of this paper.

D. Bounds on the Mean Queue Length

To study the bounds on the mean queue length, the definition and properties of stochastic orders are introduced first.

Definition III.1: Real random variable X is stochastically less than or equal to random variable Y, denoted as $X \leq_{st} Y$, if $P\{X > x\} \leq P\{Y > x\}, \forall x \in (-\infty, \infty)$.

The following properties hold when one random variable is stochastically less than or equal to the other [25].

1) $X \leq_{st} Y$ if and only if for all nondecreasing functions f, $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$.
2) If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is an increasing function (componentwise) and $\{X_i, 0 \leq i \leq n\}$ and $\{Y_i, 0 \leq i \leq n\}$ are independent sets of random variables with $X_i \leq_{st} Y_i$ for each i $(0 \leq i \leq n)$, then $f(X_1, \cdots, X_n) \leq_{st} f(Y_1, \cdots, Y_n)$.

Lemma III.3: For any fixed $\{y_1, y_2, \cdots\}$, let $f_1(x_1) = x_1$ and, for $n \geq 1$, let

$$\begin{aligned}&f_{n+1}(x_1, \cdots, x_{n+1}, y_1, \cdots, y_n)\\ &= x_{n+1} + \max\{0, f_n(x_1, \cdots, x_n, y_1, \cdots, y_{n-1}) - y_n\}. \quad (27)\end{aligned}$$

The function f_n, defined recursively above, is increasing in $\{x_1, x_2, \cdots, x_n\}$, for $n \geq 1$.

Note that if $\{x_1, x_2, \cdots\}$ and $\{y_1, y_2, \cdots\}$ are the sequences of service times and interarrival times of a queue, respectively, the waiting time (i.e., the sojourn time, including the waiting time in queue and the service time) of the nth job is equal to $f_n(x_1, \cdots, x_n, y_1, \cdots, y_{n-1})$, given that the system is empty at the beginning. The lemma thus states that: When the interarrival times are fixed, the waiting time of the nth job is an increasing function of the service times.

Let us go back to our single-server renewal queue. Denote by $\{\Lambda_1, \Lambda_2, \cdots\}$ the interarrival times. Let W_n be the waiting time of the nth job and assume the system is empty at the beginning. We have $W_n = f_n(Y_1, \cdots, Y_n, \Lambda_1, \cdots, \Lambda_{n-1})$, where Y_n is the service times of the nth job. Define $W = \lim_{n \to \infty} W_n$ as the limiting (stationary) waiting time for the server renewal queue. We can find bounds of the expected queue length $\mathbb{E}[N]$ if there is a stochastic order in the service times:

Lemma III.4: In a stable server renewal system, that is, $\rho < 1$, if there exists i such that $S_i \leq_{st} S_j$ $(0 \leq j \leq l)$, then we have $\mathbb{E}[N] \geq \rho_i + (1/2)\lambda^2\mathbb{E}[S_i^2](1 - \rho_i)^{-1}$. If there exists k such that $S_k \geq_{st} S_j$ $(0 \leq j \leq l)$ and $\rho_k < 1$, then we have $\mathbb{E}[N] \leq \rho_k + (1/2)\lambda^2\mathbb{E}[S_k^2](1 - \rho_k)^{-1}$.

Proof: Recall that $\{S_0, S_1, \cdots, S_l\}$ are the service times for the $l + 1$ stages of the server. For the first part of the lemma, we can introduce an M/G/1 queue where the jobs' interarrival times are $\{\Lambda_1, \Lambda_2, \cdots\}$ and the service times are $\{\underline{Z}_1, \underline{Z}_2, \cdots\}$, which are i.i.d. random variables having the same distribution as S_i. Let $\underline{W}_n$ be the waiting time of the nth job in this queue, and let $\underline{W} = \lim_{n \to \infty} \underline{W}_n$, if the limit exists. Note that $W_n = f_n(Y_1, \cdots, Y_n, \Lambda_1, \cdots, \Lambda_{n-1})$ and $\underline{W}_n = f_n(\underline{Z}_1, \underline{Z}_2, \cdots, \underline{Z}_n, \Lambda_1, \cdots, \Lambda_{n-1})$. By the assumption on S_i, we have $\{\underline{Z}_1, \underline{Z}_2, \cdots, \underline{Z}_n\} \leq_{st} \{Y_1, Y_2, \cdots, Y_n\}$. By Lemma III.3 and the properties on the stochastic order, we have $W_n \geq_{st} \underline{W}_n$ for each n. Since $\rho < 1$, the original queue is stable. Because $\rho_i \leq \rho < 1$, the M/G/1 queue with service times $\{\underline{Z}_1, \underline{Z}_2, \cdots, \underline{Z}_n, \cdots\}$ is stable too. Therefore, by letting n to infinity, we obtain $W \geq_{st} \underline{W}$. Since the expectations $\mathbb{E}[W]$ and $\mathbb{E}[\underline{W}]$ are both finite, then we have $\mathbb{E}[W] \geq \mathbb{E}[\underline{W}] = 1/\mu_i + (1/2)\lambda\mathbb{E}[S_i^2](1 - \rho_i)^{-1}$. By Little's law, we have $\mathbb{E}[N] = \lambda\mathbb{E}[W] \geq \lambda\mathbb{E}[\underline{W}] = \rho_i + (1/2)\lambda^2\mathbb{E}[S_i^2](1 - \rho_i)^{-1}$. The other part of the lemma can be proved similarly. ■

Lemma III.4 provides easy-to-compute bounds on the average queue length. However, it requires the existence of a stochastic order in the service times of different stages. In the following section, another set of bounds is provided, based on an M/G/1 approximation.

IV. Optimal Control of Workload

In this section, we study the optimal workload control problem defined in Section II. Given the number of stages and their corresponding service times, the long-run average profit defined in (2) can be rewritten as

$$J(\lambda) = r\lambda - h\mathbb{E}[N] - c\mathbb{E}[L_C]^{-1}. \quad (28)$$

Denote by J^*, the maximum of the long-run average profit, which is given by $\max_{\lambda > 0} J(\lambda)$. Our objective is to find λ^* such that $J^* = J(\lambda^*)$. Since $\mathbb{E}[N]$ and $\mathbb{E}[L_C]^{-1}$ increase when λ increases, it is easy to see that $J(\lambda)$ is concave. Hence, the optimal solution λ^* is unique. For a given λ, all terms in (28) are explicit except the average queue length $\mathbb{E}[N]$, which can be calculated numerically by Theorem III.2. Therefore, we can find $J(\lambda)$ and obtain the optimal arrival rate λ^* and the maximal profit J^* numerically.

Besides the matrix analytic method, this paper also investigates an approximation on the optimal workload by introducing an $\mathrm{M/G_{app}/1}$ queue. The new $\mathrm{M/G_{app}/1}$ queue has the same Poisson arrival process as the server renewal queue and i.i.d. service times $\{\widetilde{Z}_1, \widetilde{Z}_2, \cdots\}$ with distribution function $\widetilde{B}(x) = \boldsymbol{w}\boldsymbol{B}(x)^T, \forall x \geq 0$, where $\boldsymbol{B}(x) = (B_0(x), \cdots, B_l(x))$. Let $\widetilde{Z}$ be the generic random variable of the service time. The LST of $\widetilde{Z}$ is $\widetilde{B}^*(s) = \boldsymbol{w}\boldsymbol{B}^*(s)^T$, where $\boldsymbol{B}^*(s) = (B_0^*(s), \cdots, B_l^*(s))$. The first and second moments of $\widetilde{Z}$ are $\mathbb{E}[\widetilde{Z}] = -\widetilde{B}^{*(1)}(0) = 1/\overline{\mu}$ and $\mathbb{E}[\widetilde{Z}^2] = \widetilde{B}^{*(2)}(0) = \sum_{j=0}^{l} w_j B_j^{*(2)}(0) = \boldsymbol{w}D_2\boldsymbol{e}$. Let $\widetilde{N}$ be the queue length, in steady state, of the $\mathrm{M/G_{app}/1}$ queue. We have the following theorem to show the approximation error on average queue length.

Theorem IV.1: The $\mathrm{M/G_{app}/1}$ approximation on average queue length has the absolute error given by

$$\mathbb{E}[N] - \mathbb{E}[\widetilde{N}] = \frac{[\boldsymbol{v}_0 + \boldsymbol{w}D_1A]H\boldsymbol{\eta} - \rho}{1 - \rho}. \quad (29)$$

Proof:

$$\begin{aligned}&\mathbb{E}[N] - \mathbb{E}[\widetilde{N}]\\ &\quad = \frac{(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2}{1 - \rho} + \frac{\boldsymbol{w}D_2\boldsymbol{e}}{2(1 - \rho)} - \rho - \frac{\lambda^2\mathbb{E}[\widetilde{Z}^2]}{2(1 - \rho)}\\ &\quad = \frac{(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2 - \rho + \rho^2}{1 - \rho}\\ &\quad = \frac{1}{1 - \rho}\left[(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho\right].\end{aligned}$$

The proof is completed. ■

By Theorem IV.1, several results on the difference between $\mathbb{E}[N]$ and $\mathbb{E}[\widetilde{N}]$ can be obtained.

1) If all of the service rates are equal, then $\boldsymbol{\eta} = \rho\boldsymbol{e}$. Since $H\boldsymbol{e} = \boldsymbol{e}$, $A\boldsymbol{e} = \boldsymbol{e}$, $D_1\boldsymbol{e} = \boldsymbol{\eta}$, $\boldsymbol{w\eta} = \rho$, and $\boldsymbol{v}_0\boldsymbol{e} = 1 - \rho$, Theorem 4.1 leads to $\mathbb{E}[N] = \mathbb{E}[\widetilde{N}]$. This result is interesting since the variances of the service times at different

stages can still be different even if all of the service rates are the same.

2) Let $\delta_{\min}$ be the smallest element in vector $H\boldsymbol{\eta}$ and $\delta_{\max}$ be the biggest element in vector $H\boldsymbol{\eta}$. Since $\boldsymbol{v}_0 + \boldsymbol{w}D_1A$ is non-negative, Theorem IV.1 leads to

$$\frac{\delta_{min} - \rho}{1-\rho} + \mathbb{E}[\widetilde{N}] \leq \mathbb{E}[N] \leq \frac{\delta_{\max} - \rho}{1-\rho} + \mathbb{E}[\widetilde{N}]. \tag{30}$$

It is clear that, if elements in vector $H\boldsymbol{\eta}$ are close to each other and ρ is not close to one, the approximation $\mathbb{E}[\widetilde{N}]$ works well (see the Appendix for explicit expressions of H, $\delta_{\max}$ and $\delta_{\min}$).

From Theorem IV.1, the *absolute* error may significantly increase as the expected queue length goes to infinity (i.e., $\rho \to 1$). We then define the relative error of the approximation as

$$\epsilon = \frac{\left|\mathbb{E}[N] - \mathbb{E}[\widetilde{N}]\right|}{\mathbb{E}[N]}. \tag{31}$$

It is easy to see that ϵ is bounded by

$$\begin{aligned}\epsilon &= \frac{\left|\mathbb{E}[N] - \mathbb{E}[\widetilde{N}]\right|}{|\mathbb{E}[N]|}\\ &= \frac{|(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho|}{\left|(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2 + \frac{1}{2}\boldsymbol{w}D_2\boldsymbol{e}\right|}\\ &\leq \frac{\max\{|\delta_{min} - \rho|, |\delta_{\max} - \rho|\}}{\left|(\boldsymbol{v}_0 + \boldsymbol{w}D_1A)H\boldsymbol{\eta} - \rho^2 + \frac{1}{2}\boldsymbol{w}D_2\boldsymbol{e}\right|}.\end{aligned}$$

To identify situations where the approximation may work well, we undertake numerous examples for which the parameters are selected in certain ranges: All of the service rates are randomly chosen within each of the five different segments: (0,10),(1,11),(2,12),(5,15), and (10,20). The lengths of these five segment are all 10, but the possible maximal and minimal service rates vary. We also consider different traffic intensities, for example, 0.85, 0.90, 0.95, and 0.99, as well as different numbers of service stages (from 2 to 20). The transition probabilities p_j's are selected from (0,1]. Given the traffic intensity ρ, the number of service stages $l+1$, the service stage transition probabilities p_j's, and the range of the service rates, we repeat 100 times for each case. For the purpose of illustration, we plot two cases in Figs. 2 and 3 with $\rho = 0.85$ and 0.99 respectively, all p_j's equal to 1, the service rates are randomly selected from (5,15), and the service times are exponentially distributed. In each figure, there are 1000 numerical examples. These examples are grouped by the number of the service stages. In each group, the largest 5% relative errors are plotted in different symbols. The rest are plotted under a line. In Fig. 2, where $\rho = 0.85$, 95% of the examples have relative errors of less than 5.08%, and 90% of the examples have relative errors less than 4.61%. In Fig. 3, where $\rho = 0.99$, 95% of the examples have relative errors less than 6.88%, and 90% of the examples have relative errors less than 6.06%.

Similar experiments for other ρ, μ_j's ranges and p_j's have been done without giving details. From the numerical experiments and Theorem IV.1, we have the following observations:

1) The relative approximation error ϵ slowly increases when traffic intensity ρ increases.

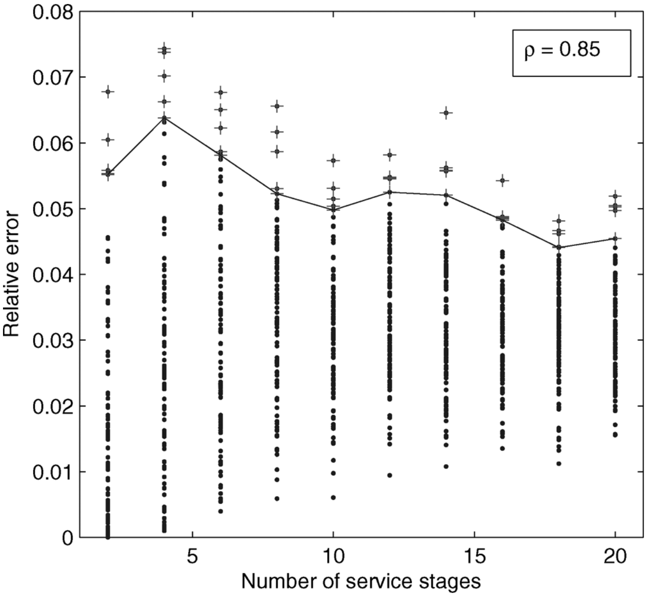


Fig. 2. Relative approximation errors when the service rates are selected within the range (5,15) and $\rho = 0.85$.

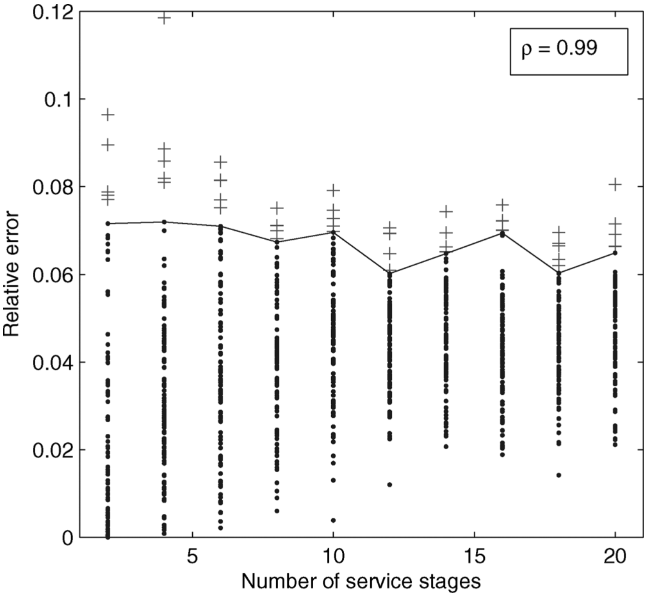


Fig. 3. Relative approximation errors when the service rates are selected within the range (5,15) and $\rho = 0.99$.

2) As the number of service stages increases, the range of the relative errors stabilizes.

3) The relative error increases significantly when most probabilities p_j's are small. On the other hand, if one stage dominates the others (e.g., $(1/p_i) > 9\sum_{j\neq i}(1/p_j)$) as in most manufacturing processes [12], the impact of small transition probabilities diminishes.

4) The approximation works well if the difference of maximal and minimal service rates is small.

In general, both of the parameters ρ and l have less of an impact on approximation error than the parameters p_j's and μ_j's. Therefore, using $\mathbb{E}[\widetilde{N}]$ to approximate $\mathbb{E}[N]$ may work well for situations where most p_j's are not very small and μ_j's are not significantly different (e.g., . $\max\{\rho_j\} < 3\min\{\rho_j\}$, $p_j \in (0.2, 1)$). More numerical results are shown in Section V.

By substituting $\mathbb{E}[\widetilde{N}]$ for the $\mathrm{M/G_{app}/1}$ queue into (28), we obtain an approximating profit function

$$\begin{aligned}\tilde{J}(\lambda) &= r\lambda - h\mathbb{E}[\widetilde{N}] - c\mathbb{E}[L_C]^{-1} \\ &= r\lambda - h\rho - \frac{h\lambda^2\mathbb{E}[\widetilde{Z}^2]}{2(1-\rho)} - c\alpha\rho. \end{aligned} \tag{32}$$

The first derivative of $\tilde{J}(\lambda)$ is given by

$$\tilde{J}'(\lambda) = r - h\overline{\mu}^{-1} - \frac{h\lambda\mathbb{E}[\widetilde{Z}^2](2-\rho)}{2(1-\rho)^2} - c\alpha\overline{\mu}^{-1}. \tag{33}$$

To maximize the long-run average profit, by setting $\tilde{J}'(\lambda) = 0$ and assuming $r\overline{\mu} - h - c\alpha > 0$, we obtain the approximating optimal arrival rate and traffic intensity as follows:

$$\widetilde{\lambda} = \overline{\mu}\left(1 - \sqrt{\frac{h\mathbb{E}[\widetilde{Z}^2]\overline{\mu}^2}{2(r\overline{\mu} - h - c\alpha) + h\mathbb{E}[\widetilde{Z}^2]\overline{\mu}^2}}\right), \tag{34}$$

$$\widetilde{\rho} = 1 - \sqrt{\frac{h\mathbb{E}[\widetilde{Z}^2]\overline{\mu}^2}{2(r\overline{\mu} - h - c\alpha) + h\mathbb{E}[\widetilde{Z}^2]\overline{\mu}^2}}. \tag{35}$$

The term $r\overline{\mu} - h - c\alpha$ can be interpreted as the profit if the machine is running all of the time. Thus, the optimization problem makes sense only if $r\overline{\mu} - h - c\alpha > 0$. If h is very small compared to r, then $\tilde{J}(\lambda) \approx J(\lambda)$, and $\widetilde{\lambda} \approx \lambda^*$. Denote the relative error of approximation on the optimal profit as ϵ_J, and

$$\epsilon_J = \frac{\left|J(\widetilde{\lambda}) - J(\lambda^*)\right|}{|J(\lambda^*)|}. \tag{36}$$

The performance of the approximate solution is examined in Section V, which is measured by the relative error ϵ_J.

V. Examples

In this section, we apply our methodologies to two typical systems and examine their effectiveness.

A. Learning System

Learning is acquiring new skills, knowledge, or values, which is a significant ability possessed by human beings. If learning is strictly defined as an increase in performance over a specified time interval, then even simple nonfeedback machines can exhibit learning behaviors. Learning techniques are of interest to control engineers on the design of artificial learning systems [27]. It is well known that the learning process over time tends to follow a learning curve, a graphical representation of the changing rates of learning for a given activity. As individuals get more experienced at a task, they usually become more efficient at it, following a progression of learning first getting faster and then slower. Typically, the increase in learning speed is the sharpest at the initial phase, and then gradually flattens out, meaning that less and less is learned after each repetition. Generally, a learning curve has an "S" shape which may have a different appearance.

The learning curve theory states that as the production quantity doubles, the required direct labour hours decrease at a predictable rate. This predictable rate can be described by the following equation which is the basis for the so-called *unit curve* [5]:

$$Y_x = Kx^{\log_2 b} \tag{37}$$

where Y_x is the number of direct labor hours to produce the xth unit, K is the number of direct labor hours to produce the first unit, x is the unit number, and b is the learning percentage.

Equation (37) is known as the Henderson's Law in the context of learning theory, from which we develop the learning process in a stochastic environment. In (37), the time spent in each product is constant. We make the time for each task random, and the mean service time is a nonincreasing function of the number of jobs. The stochastic learning process discussed in this section is a special case of the server renewal process, where the service rate at a higher stage is larger, that is, $\mu_0 \le \mu_1 \le \cdots \le \mu_l$.

Our cyclic server renewal model proposed in Section II can generalize the deterministic learning process to a stochastic one accordingly. When $p_j = 1$ and the $B_j(\cdot)$ is a one-step function for all j, our stochastic learning process becomes deterministic. Especially when $\mu_x = \mu_0 x^{\log_2^b}(x \ge 1)$, the deterministic learning process is exactly the one described in (37).

In the following text, we conduct a numerical study to show the effectiveness of the approximate solution provided in Section IV. We calculate the optimal traffic intensity ρ^* numerically according to the matrix analytic method introduced in Section III, and the approximate traffic intensity $\tilde{\rho}$ by using (34). We also calculate and compare the maximal profit and the profit associated with the approximate traffic intensity $\tilde{\rho}$.

We design various numerical examples in order to illustrate the effectiveness of the approximation solution: Some examples have service times with small variance, while others have large variances, or mixed. It is well known that an exponential random variable has a coefficient of variation 1, which is considered a medium variation. Distributions with $c_v < 1$ (such as an Erlang distribution) are considered to have a low variation, while those with $c_v > 1$ (such as a hyper-exponential distribution) are considered to have a high variation. In our study, we consider three types of service time distributions: exponential, hyper-exponential, and Erlang-2, as shown in Table I.

The only approximation in the profit function is the average queue length $\mathbb{E}[N]$. We first set $r = 100$, $h = 10$, $c = 100$. The numerical results, in fact, are similar for other r, h, and c. In Table II, the server has two learning stages: 0 and 1. The service rate at stage 0 is 3, and 6 at stage 1. The stage transition probability is selected from the set $\{0.1, 0.5, 0.9\}$. The results show that the approximate solutions are very close to the optimal solutions in most scenarios. Although in a few cases, the gap between $\widetilde{\rho}$ and ρ^* can reach 3%, the difference of corresponding average profit is only about 1% (see examples 9, 18, and 27 in Table II). In these examples, the maximal service rate is twice the minimal service rate. The minimal transition probability is 0.1. According to our results in Section IV, the approximation

TABLE I
CONSTRUCTION OF DIFFERENT DISTRIBUTIONS WITH THE SAME MEAN VALUE

Distribution	Mean	Variance	c_v	Memo
Exp.	μ^{-1}	μ^{-2}	1	
Hyper-Exp.	μ^{-1}	$2\sum_{i=1}^{n} q_i x_i^2 - \mu^{-2}$	$\sqrt{2\mu^2 \sum_{i=1}^{n} q_i x_i^2 - 1}$	$\sum_{i=1}^{n} q_i x_i = \mu^{-1}$
E_k	μ^{-1}	μ^{-2}/k	$1/\sqrt{k}$	

Note: In our numerical examples, we set $n = 2, q_1 = q_2 = 0.5, x_1 = \mu^{-1}/3, x_2 = 5\mu^{-1}/3$ for the hyper-exponential distribution, hence $c_v = \sqrt{17}/3$.

TABLE II
TWO-STAGE LEARNING SYSTEM WITH DIFFERENT SERVICE TIME DISTRIBUTIONS ($r = 100, h = 10, c = 100$)

					Traffic intensity		Profit		
Distribution	#	$\mu_0(p_0)$	$\mu_1(p_1)$	$\overline{\mu}$	$\widetilde{\rho}$	ρ^*	$J(\widetilde{\lambda})$	$J(\lambda^*)$	$\epsilon_J(\%)$
Exp.	1	3(.9)	6(.9)	4.0000	.7758	.7808	134.3049	134.3264	0.02
	2	3(.9)	6(.5)	4.4211	.8068	.8090	197.1861	197.1987	0.01
	3	3(.9)	6(.1)	5.4545	.8530	.8530	361.6894	361.6894	0.00
	4	3(.5)	6(.9)	3.6522	.7911	.7925	156.2928	156.2978	0.00
	5	3(.5)	6(.5)	4.0000	.8079	.8079	196.5385	196.5385	0.00
	6	3(.5)	6(.1)	5.1429	.8473	.8434	336.2375	336.2979	0.02
	7	3(.1)	6(.9)	3.1579	.8112	.8112	189.2782	189.2782	0.00
	8	3(.1)	6(.5)	3.2727	.8138	.8108	198.1585	198.1655	0.00
	9	3(.1)	6(.1)	4.0000	.8293	.8018	248.4020	250.1623	0.70
Hyper-Exp.	10	3(.9)	6(.9)	4.0000	.7335	.7360	122.5283	122.5388	0.01
	11	3(.9)	6(.5)	4.4211	.7697	.7719	182.2516	182.2589	0.00
	12	3(.9)	6(.1)	5.4545	.8242	.8242	340.8949	340.8949	0.00
	13	3(.5)	6(.9)	3.6522	.7513	.7527	143.5917	143.5947	0.00
	14	3(.5)	6(.5)	4.0000	.7710	.7710	181.9288	181.9288	0.00
	15	3(.5)	6(.1)	5.1429	.8174	.8135	316.7011	316.7355	0.01
	16	3(.1)	6(.9)	3.1579	.7749	.7749	175.4406	175.4406	0.00
	17	3(.1)	6(.5)	3.2727	.7779	.7749	184.0186	184.0207	0.00
	18	3(.1)	6(.1)	4.0000	.7961	.7736	234.7526	235.7365	0.42
E_2	19	3(.9)	6(.9)	4.0000	.8046	.8096	142.8761	142.9103	0.02
	20	3(.9)	6(.5)	4.4211	.8319	.8364	207.8928	207.9117	0.01
	21	3(.9)	6(.1)	5.4545	.8723	.8723	376.2419	376.2419	0.00
	22	3(.5)	6(.9)	3.6522	.8181	.8195	165.4406	165.4477	0.00
	23	3(.5)	6(.5)	4.0000	.8329	.8329	206.9448	206.9448	0.00
	24	3(.5)	6(.1)	5.1429	.8673	.8615	349.7760	349.8697	0.03
	25	3(.1)	6(.9)	3.1579	.8358	.8358	199.1215	199.1215	0.00
	26	3(.1)	6(.5)	3.2727	.8381	.8350	208.1544	208.1669	0.01
	27	3(.1)	6(.1)	4.0000	.8516	.8191	257.1033	259.8009	1.04

would generally work well. The worst case is when both transition probabilities reach the smallest value 0.1. The results are consistent with our previous findings. Besides these, we can see that a higher variance in service times produces a lower profit. All of the examples in Table II with hyper-exponential service times generate lower average profits than those with exponential service times. The examples with Erlang-2 service times (with the lowest variances) produce the highest profits.

Other examples also show similar results. For instance, we study 1000 examples with μ_j selected from [5], [15] and $p_j \in (0.2, 1]$. In this case, we have found that 95% of the examples have relative errors on the optimal profit of less than 2.28%.

B. Deteriorating System

Deteriorating servers are well observed in manufacturing systems. Machines deteriorate over time, and need to be replaced when their performance fails to meet the requirement. There are tremendous works dealing with the related problems (see [4], [15], [19], and [30], and the references therein). In this section, we discuss the application of our methodologies to a simple deteriorating system. It is assumed that the status of a server may deteriorate after each service completion. This server deteriorating process is another special case of a server renewal process, where the service rate at a higher stage is smaller, that is, $\mu_0 \geq \mu_1 \geq \cdots \geq \mu_l$. Most of the results we obtain from the learning system hold for the deteriorating system. For the optimal workload control problem of the server deteriorating system, the approximate solution works well, which is illustrated in Table III and verified by various numerical examples. In Table III, the service rates are (6, 5, 4, 3), while the transition probabilities are randomly generated. We consider three types of service time distributions: exponential, hyper-exponential, and Erlang distribution. In addition, we have studied 1000 examples with μ_j selected from [5] and [15] and $p_j \in (0.2, 1]$. In summary, we have found that 95% of the examples have relative approximation errors on the optimal profit of less than 1.75%.

It is worth noticing that a deteriorating system is *not* a simple reverse of the learning system. For example, let $\rho = 0.9$, $\mu = (60,5,4,3)$, and $p = (0.03,0.04,0.05,0.06)$. The average queue length is 53.22, while it is 56.69 for its reverse learning system. One possible reason may be the difference in service processes. For instance, the service times sequence of a learning system

TABLE III
SEVERE DETERIORATING SYSTEMS WITH FOUR SERVICE STAGES ($r = 100$, $h = 10$, $c = 100$)

						Traffic intensity		profit	
#	$\mu_0(p_0)$	$\mu_1(p_1)$	$\mu_2(p_2)$	$\mu_3(p_3)$	$\overline{\mu}$	$\widetilde{\rho}$	ρ^*	$J(\widetilde{\lambda})$	$J(\lambda^*)$
Exp.									
1	6(.6315)	5(.3676)	4(.0576)	3(.5869)	4.0882	.8385	.8385	275.2312	275.2312
2	6(.4544)	5(.0841)	4(.6927)	3(.7176)	4.7415	.8475	.8475	320.6466	320.6466
3	6(.6756)	5(.1536)	4(.3533)	3(.4418)	4.3457	.8381	.8358	281.3433	281.3491
Hyper-Exp.									
4	6(.6315)	5(.3676)	4(.0576)	3(.5869)	4.0882	.8071	.8071	257.9198	257.9198
5	6(.4544)	5(.0841)	4(.6927)	3(.7176)	4.7415	.8177	.8177	301.5551	301.5551
6	6(.6756)	5(.1536)	4(.3533)	3(.4418)	4.3457	.8065	.8042	263.7481	263.7502
E_2									
7	6(.6315)	5(.3676)	4(.0576)	3(.5869)	4.0882	.8597	.8597	287.4112	287.4112
8	6(.4544)	5(.0841)	4(.6927)	3(.7176)	4.7415	.8676	.8676	334.0341	334.0341
9	6(.6756)	5(.1536)	4(.3533)	3(.4418)	4.3457	.8593	.8570	293.6782	293.6883

is $\{S_{1,0}, S_{1,1}, S_{1,2}, S_{2,0}, S_{2,1}, S_{2,2}, S_{3,0}, \ldots\}$, while for the deteriorating system, it is $\{S_{1,2}, S_{1,1}, S_{1,0}, S_{2,2}, S_{2,1}, S_{2,0}, \ldots\}$, where $\{S_{i,j}, i = 1, 2, \ldots\}$ are i.i.d. r.v.s for $j = 0,\ 1,\ 2$. Regardless of the starting point, these two sequences are different.

VI. CONCLUSION

In this paper, we develop two methods to study systems where the server's performance changes cyclically. We apply matrix-analytic methods and obtain analytical results for system performance measures. To solve the optimal workload control problem, we use an M/G/1 queue approximation to help design an explicit solution. The developed methodology is illustrated by numerical examples. In future research, a more efficient computing algorithm for the matrix analytic solution can be explored for particular cyclic change patterns, which can be used when the M/G/1 approximation yields an unsatisfactory error. It might also be interesting to see whether any bounds can be provided for the variance of the queue length using an M/G/1 approximation.

APPENDIX
EXPRESSION OF H

We show that H has the following explicit expression:

$$H = \left(1 + \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\right)\boldsymbol{ew} - \left(\sum_{i=0}^{l} p_i^{-1}\right)^{-1} B \tag{38}$$

where

$$B(i,j) = \begin{cases} p_{j-1}^{-1}\left(\sum_{k=i-1}^{j-1} p_k^{-1}\right), & i \le j; \\ p_{j-1}^{-1}\left(\sum_{k=i-1}^{l} p_k^{-1} + \sum_{k=0}^{j-1} p_k^{-1}\right), & i > j. \end{cases} \tag{39}$$

It can be verified with routine algebra that

$$\boldsymbol{w}B = \sum_{i\le j}(p_ip_j)^{-1}\boldsymbol{w}, \quad B\boldsymbol{e} = \sum_{i\le j}(p_ip_j)^{-1}\boldsymbol{e} \tag{40}$$

and

$$\begin{aligned} &(I - A + \boldsymbol{ew})H \\ &= \left(1 + \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\right)\boldsymbol{ew} - \frac{(I - A + \boldsymbol{ew})B}{\sum_{i=0}^{l} p_i^{-1}} \\ &= \left(1 + \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\right)\boldsymbol{ew} - \frac{(I-A)B}{\sum_{i=0}^{l} p_i^{-1}} - \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\boldsymbol{ew} \\ &= \boldsymbol{ew} - \frac{(I - A)B}{\sum_{i=0}^{l} p_i^{-1}} \\ &= \boldsymbol{ew} - \frac{\left(\sum_{i=0}^{l} p_i^{-1}\right)(\boldsymbol{ew} - I)}{\sum_{i=0}^{l} p_i^{-1}} \\ &= I. \end{aligned}$$

Thus, the matrix H defined in (38) is indeed the inverse of $I - A + \boldsymbol{ew}$. The expressions for $\delta_{\max}$ and δ_{min} are given as follows.

$$\delta_{\max} = \rho + \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\rho - \frac{\min_k\left\{\sum_j B(k,j)\rho_j\right\}}{\sum_{i=0}^{l} p_i^{-1}}, \tag{41}$$

$$\delta_{min} = \rho + \frac{\sum_{i\le j}(p_ip_j)^{-1}}{\sum_{i=0}^{l} p_i^{-1}}\rho - \frac{\max_k\left\{\sum_j B(k,j)\rho_j\right\}}{\sum_{i=0}^{l} p_i^{-1}}. \tag{42}$$

REFERENCES

[1] E. Berk and K. Moinzadeh, "Analysis of maintenance policies for M machines with deteriorating performance," *IIE Trans.*, vol. 32, pp. 433–444, 2000.

[2] D. Bini, B. Meini, S. Steff, J. F. Prez, and B. Van Houdt, "SMCSolver and Q-MAM: Tools for matrix-analytic methods," *SIGMETRICS Performance Eval. Rev.*, vol. 39, pp. 46–46, 2012.

[3] O. J. Boxma and I. A. Kurkova, "The M/G/1 Queue with two service speeds," *Adv. Appl. Probability*, vol. 33, pp. 520–540, 2001.

[4] R. E. Barlow, F. Proschan, and L. C. Hunter, *Mathematical Theory of Reliability*. New York: Wiley, 1965.

[5] R. B. Chase, *Operations Management for Competitive Advantage*, International ed. New York, USA: McGraw-Hill/Irwin, 2001.

[6] S.-K. Cheung, R. J. Boucherie, and R. Núñez Queija, "Quasi-stationary analysis for queues with temporary overload," in *Proc. ITC 22*, S. C. Borst, M. R. H. Mandjes, and M. S. Squillante, Eds., Amsterdam, the Netherlands, 2010, pp. 1–8.

[7] J. W. Cohen*, The Single Server Queue*. Amsterdam, the Netherlands: North-Holland, 1982.

[8] J. H. Dshalalow and L. Tadj, "A queueing system with random server capacity and multiple control," *Queueing Syst.*, vol. 14, pp. 369–384, 1993.

[9] A. Dudin*, Queue M/G/1 in a Semi-Markovian Random Environment*. Minsk, Belarus, Russia: Lab. Appl. Probabilistic Anal., Belarus State Univ., 2001.

[10] A. Dudin and A. Markov, "Calculation of the characteristics of the queue in the semi-Markovian cyclic RE," in *Proc. 15th All-Union Workshop Comput. Netw., Part 2*, Moscow, Russia, 1990, pp. 241–246.

[11] S. Halfin, "Steady-state distribution for the buffer content of an M/G/1 queue with varying service rate," *SIAM J. Appl. Math.*, vol. 23, pp. 356–363, 1972.

[12] W. Kuo and T. Kim, "An overview of manufacturing yield and reliability modelling for semiconductor products," *Proc. IEEE*, vol. 87, no. 8, pp. 1329–1344, Aug. 1999.

[13] G. Levitin*, The Universal Generating Function in Reliability Analysis and Optimization*. London, U.K.: Springer, 2005.

[14] D. M. Lucantoni and M. F. Neuts, "Some steady-state distributions for the MAP/SM/1 queue," *Stochastic Models*, vol. 10, pp. 575–598, 1994.

[15] J. J. McCall, "Maintenance policies for stochastically failing equipment: A survey," *Manage. Sci.*, vol. 11, pp. 493–524, 1965.

[16] M. F. Neuts, "A queue subject to extraneous phase changes," *Adv. Appl. Prob.*, vol. 3, pp. 78–119, 1971.

[17] M. F. Neuts*, Structured Stochastic Matrices of M/G/1 Type and Their Applications*. New York: Marcel Dekker, 1989.

[18] R. Núñez Queija, "A queueing model with varying service rate for ABR," in *Computer Performance Evaluation—Modelling Techniques and Tools. Proc. TOOLS (Mallorca, Spain)*, R. Puigjaner, N. N. Savino, and B. Serra, Eds. Berlin, Germany: Springer Verlag, 1998, pp. 93–104.

[19] W. P. Pierskalla and J. A. Voelker, "A survey of maintenance models: The control and surveillance of deteriorating systems," *Naval Res. Logist. Quart.*, vol. 23, pp. 353–388, 1976.

[20] G. Pestalozzi, "A queue with Markov-dependent service times," *J. Appl. Probab.*, vol. 5, pp. 461–466, 1968.

[21] H. Pham and H. Z. Wang, "Imperfect maintenance," *Eur. J. Oper. Res.*, vol. 94, pp. 425–438, 1996.

[22] V. Ramaswami, "A stable recursion for the steady state vector in Markov chains of M/G/1 type," *Stochastic Models*, vol. 4, pp. 183–188, 1988.

[23] G. J. K. Regterschot and J. H. A. De Smit, "The queue M/G/1 with Markov modulated arrivals and service," *Math. Oper. Res.*, vol. 11, pp. 465–483, 1986.

[24] S. M. Ross, J. G. Shanthikumar, and X. Zhang, "Some pitfalls of black box queue inference: The case of state-dependent server queues," *Probab. Eng. Inf. Sci.*, vol. 7, pp. 149–157, 1993.

[25] M. Shaked and J. G. Shanthikumar*, Stochastic Orders and their Applications*. New York, USA: Associated Press, 1994.

[26] J. G. Shanthikumar, "On a single-server queue with state-dependent service," *Naval Res. Logist. Quart.*, vol. 26, no. 2, pp. 305–309, 1979.

[27] J. Sklansky, "Learning systems for automatic control," *IEEE Trans. Autom. Control*, vol. AC-11, no. 1, pp. 6–19, Jan. 1966.

[28] T. Takine, "A new recursion for the queue length distribution in the stationary BMAP/G/1 queue," *Stochastic Models*, vol. 16, pp. 335–341, 2000.

[29] T. Takine and T. Hasegawa, "The workload in the MAP/G/1 queue with state-dependent services: Its application to a queue with preemptive resume priority," *Stochastic Models*, vol. 10, no. 1, pp. 183–204, 1994.

[30] C. Valdez-Flores and R. M. Feldman, "A survey of preventive maintenance models for stochastically deteriorating single-unit systems," *Naval Res. Logistics*, vol. 36, pp. 419–446, 1989.

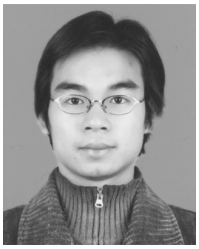
[31] R. W. Wolf, "Poisson arrivals see time averages," *Oper. Res.*, vol. 30, pp. 223–231, 1982.

[32] U. Yechiali, "A queueing-type birth-and-death process defined on a continuous-time Markov chain," *Oper. Res.*, vol. 21, pp. 604–609, 1973.

[33] U. Yechiali and P. Naor, "Queueing problems with heterogeneous arrival and service," *Oper. Res.*, vol. 19, pp. 722–734, 1971.

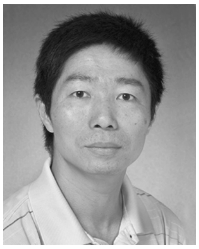
Boray Huang received the M.S. degree in industrial engineering at National Tsinghua University, Taiwan, and the Ph.D. degree in industrial engineering and management sciences at Northwestern University, Chicago, IL, USA, in 2004.

Currently, he is an Assistant Professor in the Department of Industrial and Systems Engineering, National University of Singapore, Singapore. His research interests include production and inventory controls in manufacturing systems, service operations, and supply chain management. His most recent research is working on stochastic job sequencing problems and their applications in manufacturing and service systems.



Jingui Xie received the Ph.D. degree in management science and engineering from Tsinghua University, Beijing, China, in 2010.

Currently, he is an Associate Professor in the School of Management, University of Science and Technology of China, Hefei, Anhui, China. Prior to that, he was a Research Fellow with NUS Business School, and the Department of Industrial and Systems Engineering, National University of Singapore, Singapore. He was a *Chazen* Visiting Scholar in the Graduate School of Business, Columbia University, New York. His research interests focus on stochastic modeling and control, and their applications in manufacturing and service operations. He is now working on queueing models and their applications in health-care systems.



Qi-Ming He received the Ph.D. degree in operations research from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China, in 1989 and the Ph.D. degree in management science from the University of Waterloo, Waterloo, ON, Canada, in 1996.

Currently, he is a Professor in the Department of Management Sciences, University of Waterloo. In investigating various stochastic models, his favorite methods are matrix-analytic methods. Recently, he has been 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. His main research areas are algorithmic methods in applied probability, queueing theory, inventory control, and production management.