Stabilizing the virtual response time in single-server processor sharing queues with slowly time-varying arrival rates

Yongkyu Cho1 · Young Myoung Ko2

Published online: 13 January 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
Motivated by the work of Whitt (Queueing Syst 81(4):341–378, 2015) and Ma and Whitt (in: Winter simulation conference, pp 2598–2609, 2015), that studied performance stabilization in a GI_t/GI_t/1 queue, this research study investigates the stabilization of the mean virtual response time in a single-server processor sharing (PS) queueing system with a slowly time-varying arrival rate and a service rate control (a GI_t/GI_t/1/PS queue). We propose and compare a modified square-root (SR) control and a difference-matching (DM) control to stabilize the mean virtual response time of a GI_t/GI_t/1/PS queue. Extensive simulation studies with various settings of arrival processes and service times show that the DM control outperforms the SR control for heavy-traffic conditions, and that the SR control performs better for light-traffic conditions.

Keywords Stabilizing performance · Nonstationary queues · Processor sharing · Service rate control · Queueing simulation

1 Introduction
Modern data centers consume tremendous amounts of energy to supply networking, computing, and storage services to global IT companies. Concerns about energy consumption have prompted researchers to explore operational methods that maximize energy efficiency and satisfy a certain level of quality of service (QoS) (Anselmi and Verloop 2011; Ko and Cho 2014; Liao et al. 2015). QoS can be achieved by adding constraints that impose upper bounds for response time-related metrics, e.g., the mean (virtual) response time and the tail probability of the response time. In general, these constraints are binding, because of the
conflict between the QoS-related metrics and energy consumption. Binding the QoS-related constraints implies that the metrics are maintained as a constant value, and suggests the need to investigate the stabilization of response times. Although some proposed methodologies (Anselmi and Verloop 2011; Ko and Cho 2014; Liao et al. 2015) assume the stationarity of data traffic arrival processes, nonstationary properties, such as time-varying arrival rates from real data (CAIDA 2016), make it difficult to analyze queueing system performance.

Due to such difficulties, stabilizing performance has been suggested so that it makes the systems look stationary in terms of the performance measures. The approach may give us a chance to design, operate, and manage large-scale stochastic systems in which statistical performance guarantees are an important concern (e.g., data center or telecommunications system).

The literature on this topic may be grouped into three different bodies according to the target system considered: single-server queues, many-server queues, and multiple parallel single server queues. First, Ma and Whitt (2015) and its sequels (Whitt 2015; Ma and Whitt 2018) studied how the different forms of continuously time-dependent service rates affect the stabilization in $GI_t/GI_t/1$ systems. Results showed that the rate-matching service rate (i.e., keeping the instantaneous traffic intensity a constant value) effectively stabilizes the mean queue length but none of the proposed service rates succeeded to stabilize the mean waiting time (we deal with this in Sect. 1.1). Recently, Ma and Whitt (2019) presented a method for stabilizing the mean waiting time by modifying the rate matching control with a periodic arrival rate. They added two parameters, namely a time lag and a damping factor, that horizontally shift and vertically scale the service rate function. Two parameters are obtained by solving a min-max problem—minimizing the maximum expected waiting time—using simulation optimization. Their new control scheme indeed stabilizes the mean waiting time, but as they mentioned, it is not perfect. Unlike previous studies, the new rate control requires an additional procedure for finding parameters that may not be computationally trivial. Those studies built both theoretical and experimental bases to handle nonstationary single server systems and opened up new research questions: What happens if the service policy is not the first-come-first-served (FCFS), for example, the processor sharing (PS)? Do the service rate functions have to be modified according to service disciplines while it is still work-conserving?

Another body of studies considered many-server queueing systems. In this type of system, the decision mainly concerns how to choose a time-varying number of servers (called server staffing) to achieve time-stable performance. A seminal paper on this topic initially proposed a staffing method for $G_t/G/s_t$ system to control the probability of a delay (before beginning service) toward hitting or falling below a target probability (Jennings et al. 1996). Feldman et al. (2008) extended the situation where there is abandonment. They developed a flexible simulation-based iterative staffing algorithm for the $M_t/G/s_t + G$ type queueing system with a nonstationary Poisson arrival process and generally distributed customer impatience times. For more references on this topic, see Horvath et al. (2007), Liu and Whitt (2012, 2014), He et al. (2016), Liu and Whitt (2017) and Liu (2018).

Kwon and Gautam (2016b) suggested a holistic control approach applying to parallel queues with multi-class and nonstationary job arrivals using assignment, sizing, and routing decisions. Under piecewise constant time-varying arrival rates, their suggested method showed effectiveness in stabilizing the queue length distribution. Using the fact that the response time distribution is also stabilized under the piecewise constant arrival rate setting, their sequel suggested a data center operating framework that achieves energy-efficiency with statistical performance guarantees on the response time (Kwon and Gautam 2016a).
1.1 Naive application of the FCFS-based service rate control schemes to PS systems

This section shows some results of simulation experiments applying the service rate controls proposed by Ma and Whitt (2016) and Whitt (2015) to PS queueing systems. The three controls considered in the literature are expressed as follows:

Rate-matching (RM): \( \mu_{RM}(t) \equiv \frac{\lambda(t)}{\rho} \), \hspace{1cm} (1)

The first square-root (SR1): \( \mu_{SR1}(t) \equiv \lambda(t) + \xi \sqrt{\lambda(t)} \), \hspace{1cm} (2)

The second square-root (SR2): \( \mu_{SR2}(t) \equiv \lambda(t) + \frac{\lambda(t)}{2} \left( \sqrt{1 + \frac{\zeta}{\lambda(t)}} - 1 \right) \), \hspace{1cm} (3)

for \( t \geq 0 \), where \( \rho \) is given target traffic intensity, and \( \xi \) and \( \zeta \) are also given positive parameters. The RM control is motivated by an experimental idea that keeps the instantaneous traffic intensity constant over time, which in turn showed the effectiveness in stabilizing the mean queue length. The other two controls are designed to stabilize the mean response time that is of our interest. For the detailed derivation of the controls, see Whitt (2015).

Figure 1 depicts some simulation results applying three service rate control schemes in Ma and Whitt (2015) and Whitt (2015) to a \( GI_t/GI_t/1/PS \) system with Erlang distribution for both service and arrival base distributions. Similar to the \( GI_t/GI_t/1 \) case in Whitt (2015), the RM control also stabilizes the mean queue length (Fig. 1a). For the SR1 control, we do not find any positive result (Fig. 1b) whereas the SR2 control shows some stabilizing performance in terms of the mean response time (Fig. 1c) although it is not perfect.

As discovered by the preliminary experiments, the control schemes developed for \( GI_t/GI_t/1 \) queues do not show significant effectiveness in stabilizing the mean response time of \( GI_t/GI_t/1/PS \) systems. To this end, we study the service rate controls to stabilize the mean response time to a certain target value in a single-server processor sharing queue representing a computer server in a data center under time-varying arrival rates and controllable service rates.

Fig. 1 General performance measures of \( ER_t/ER_t/1/PS \) queues under the three former controls where \( \lambda(t) = 1 + 0.2 \sin(0.001t) \)
Our approach is similar to Ma and Whitt (2016) and Whitt (2015), who considered three different service rate controls, two of which were designed to stabilize the mean (virtual) waiting time in a $GI_t/GI_t/1/FCFS$ queue. The slowly time-varying traffic patterns of internet services CAIDA (2016) justify our use of pointwise stationary approximation (PSA) Green and Kolesar (1991). We adopt different heavy-traffic approximation results, because our objective is to stabilize the mean response time, which is one of our target performance measures. We propose two service rate control schemes:

$$\mu_{SR}(t; s) \equiv (s\lambda(t) + 1)\beta + \sqrt{(s\lambda(t) + 1)^2\beta^2 + 4s\lambda(t)\beta^2(VFCFS - 1)}}{2s}, \quad (4)$$

$$\mu_{DM}(t; s) \equiv \beta \left(\lambda(t) + \frac{VPS}{s}\right), \quad (5)$$

where $s$ is the desired response time, $\beta$ is the mean job size, $\lambda(t)$ is the arrival rate function, $VFCFS \equiv (C_a^2 + C_s^2)/2$, and $VPS \equiv (C_a^2 + C_s^2)/(1 + C_s^2)$, with $C_a^2$ and $C_s^2$ are the squared coefficient of variations (SCV) of the base interarrival and service time distributions. Equation (4) is a modification of the well-known square-root control (SR) suggested in Whitt (2015), and Eq. (5) is a new control scheme, which we call the difference-matching (DM) control, because it maintains the difference between $\mu(t)$ and $\beta\lambda(t)$ as a constant $\beta VPS/s$. The DM control is easy to implement thanks to its simplicity.

Figure 2 shows the general performance measures, i.e., mean queue length process $\mathbb{E}[Q(t)]$ (green line) and the mean response time process $\mathbb{E}[R(t)]$ (red line), of the simulated $GI_t/GI_t/1/PS$ queues with an Erlang base arrival distribution ($ER_t$) and a lognormal job size distribution ($LN_t$) with the SR control as in Eq. (4) and three different time-varying arrival rates. The dotted black lines are 95% confidence intervals and the dotted blue line plots the arrival rate function; its dedicated y-axis is on the right. The plots show that the response time is almost perfectly controlled by the SR control under the light-traffic condition.

Figure 3 depicts the general performance measures when the target response time is relatively long. While the stabilization looks imperfect for both controls, their relative amplitude—one of our performance metrics for the controls described in Sect. 4.2—is under 10%. Figure 3b depicts that the DM control achieves the target response time, which implies that using the DM control shows better accuracy—the other performance metric in Sect. 4.2—under a heavy-traffic condition (long response time).

This paper contributes to the published literature on queueing systems by studying the response time stabilizing controls for a $GI_t/GI_t/1/PS$ queue; proposing a new control scheme, i.e., the DM control, for heavy-traffic conditions; undertaking extensive simulations of the proposed control schemes; and gaining insights into their effectiveness for data centers.

The remainder of this paper is organized as follows. Section 2 introduces a single-server PS queueing model with a time-varying arrival rate and a controllable service rate. We explain some details for simulating a $GI_t/GI_t/1/PS$ queue, which is not straightforward, unlike its stationary counterpart. Section 3 explains the procedure to derive the two service rate controls, and some simple characteristics of the controls. Section 4 reports the results of the simulations including the interesting phenomena we find. Section 5 concludes and suggests some future research directions.
Fig. 2 General performance measures of $E R_t/L N_t/1/PS$ queues under the SR control where $\lambda(t) = 1 + 0.2 \sin (\gamma t)$ with target response time 0.1 (light-traffic)

(a) $\gamma = 0.1$
(b) $\gamma = 0.01$
(c) $\gamma = 0.001$

Fig. 3 General performance measures of $E R_t/E R_t/1/PS$ queues where $\lambda(t) = 1 + 0.2 \sin (0.001t)$ and the target response time is 10 (heavy-traffic)

(a) SR control
(b) DM control

2 The model

Section 2.1 introduces a single-server queueing model with nonstationary non-Poisson arrivals under the PS discipline and the service rate control. Section 2.2 explains the procedures to simulate such queueing systems. Throughout this paper, we use the following notations:

- $f(t)$: arbitrary periodic function with a period $T_f$
- $\bar{f}$: spatial scale average of $f$; $\bar{f} \equiv \int_t^{t+T_f} f(x)dx/T_f$ for any $t \in [0, \infty)$
- $\lambda(t)$: arrival rate function
- $\mu(t)$: service rate function
- $T_i$: base inter-arrival times between $i$th and $i-1$st job; i.i.d. random variables having a general distribution function $F(\cdot)$ with a mean $\tau \equiv \mathbb{E}[T_i] < \infty$ and an SCV $C_a^2 \equiv SCV(T_i) < \infty$
– $S_i$: service requirement that the $i$th job brings; i.i.d. random variables having a general distribution function $G(\cdot)$ with a mean $\beta \equiv \mathbb{E}[S_i] < \infty$ and an SCV $C_s^2 \equiv SCV(S_i) < \infty$
– $\rho(t)$: instantaneous traffic intensity; $\rho(t) \equiv \beta \lambda(t)/\mu(t)$
– $A_i$: time when the $i$th job arrives
– $D_i$: time when the $i$th job departs
– $A(t)$: arrival process; number of job arrivals during interval $(0, t]$
– $D(t)$: departure process; number of job departures during interval $(0, t]$
– $Q(t)$: queue length process; number of jobs in the system at time $t$
– $R(t)$: response time process; time that a virtual customer arriving at time $t$ would spend in the system

We note that a nonstationary system can be stable in the long-run although $\rho(t)$ is greater than or equal to 1 for some $t$ as long as the following condition (i.e., stability condition for the time-varying queue) holds:

$$\lim_{t \to \infty} \beta \int_0^t \lambda(s)ds \int_0^t \mu(s)ds < 1.$$

(6)

Throughout this paper, we explicitly keep the mean job size $\beta$ that is commonly omitted by letting $\beta = 1$ without loss of generality. We think omitting $\beta$ may give rise to confusion to practitioners who may operate systems where a job size is not usually 1 on average (e.g., network packet size).

2.1 The $GI_t/GI_t/1/PS$ queue

We consider a single server processor sharing queueing system where arrivals follow an NSNP. We assume that the time-dependent arrival rate function $\lambda(\cdot)$ is continuous and bounded finitely both below and above. Under the assumption, the cumulative arrival function $\Lambda(t) \equiv \int_0^t \lambda(s)ds$ is well-defined for $t \geq 0$ and so is the inverse $\Lambda^{-1}(\cdot)$.

Each job has its own service requirement, e.g., job size, to be processed by a server. Assume that the job size is determined upon arrival in ICT service systems, e.g., packet size or file size. Let $S_i$ be the service requirement that the $i$th job brings, and assume that $S_i$’s are independent and identically distributed. Appropriate control schemes dynamically determine service rate function $\mu(\cdot)$. Assume that function $\mu(\cdot)$ is continuous and bounded so that it can be integrated on compact intervals to obtain a cumulative service function $M(t) \equiv \int_0^t \mu(s)ds$. The amount of service processed by the server during time interval $(t_1, t_2]$ is $M(t_2) - M(t_1) \equiv \int_{t_1}^{t_2} \mu(s)ds$.

The PS policy is a work-conserving service discipline which is commonly used to describe computer systems (especially CPUs) Gautam (2012). All jobs in the system evenly share the server or processor at any given time, e.g., if the processor runs at a processing speed of $\mu$ bits/s and there are $n$ jobs, then each job is processed by $\mu/n$ bits/s.

2.2 Simulating the $GI_t/GI_t/1/PS$ queue

Simulating a $GI_t/GI_t/1/PS$ queue is difficult and computationally expensive because of non-Poisson arrivals, time nonhomogeneity, processor sharing, and other factors. Therefore, we combine two algorithms (Gerhardt and Nelson 2009; Ma and Whitt 2016) for simulation. The first algorithm by Gerhardt and Nelson (2009) provides the supporting theory for gen-
erating an NSNP from its stationary counterpart, and the second algorithm by Ma and Whitt (2016) gives a numerical approximation method to relieve the computational burden when the rate function is periodic.

2.2.1 The arrival process

Let \( A(t) \) be the NSNP arrival process we want to simulate. Construct the process by applying the change of time to a stationary renewal process. Let \( N(t) \) be the stationary renewal process with i.i.d. interrenewal times \( \{T_i, i \geq 1\} \). Then,

\[
E[N(t)] = \frac{t}{\tau}, \quad (7)
\]

\[
Var[N(t)] = E[N(t)]SCV(T_i) + o(t), \quad (8)
\]

where \( \tau \equiv E[T_i] \). In particular, we call \( N(t) \) the standard equilibrium renewal process (SERP) when \( \tau = 1 \) and \( T_i \) is a random variable having the stationary excess distribution given by

\[
F_e(t) \equiv \frac{1}{E[T_i]} \int_0^t 1 - F(s)ds. \quad (9)
\]

By defining \( A(\cdot) \) to be the composition of \( N(\cdot) \) and \( \Lambda(\cdot) \), i.e., \( A(t) = N(\Lambda(t)) \) with \( E[A(t)] = \Lambda(t) \), we construct an NSNP.

Generating samples of an NSNP requires careful attention since the time-varying property of the process may complicate the simulation methods, which lets the generated set of samples have subtly different statistical properties (e.g., the variances are different while the means are identical) according to which method we use Gerhardt and Nelson (2009). In this study, we adopt the inversion method described in Gerhardt and Nelson (2009) (see Appendix 1 for the detailed algorithmic procedure). Constructing the arrival process \( A(t) \) by Algorithm 1 prompts the following remark.

**Remark 1** (Gerhardt and Nelson 2009) \( E[A(t)] = \Lambda(t) \), for \( t \geq 0 \) and \( Var[A(t)] \approx \Lambda(t)SCV(T_i) \), for large \( t \).

We note that NSNP is a generalization of the simple nonstationary Poisson process (NSPP), where \( T_i \) is exponentially distributed. It can be verified easily this by plugging 1 into \( SCV(T_i) \).

2.2.2 The service times

The service completion time is determined as soon as a job arrives when the FCFS discipline applies. Under the PS policy, however, it is not determined upon arrival, because future arrivals will affect the service times of of the jobs already existing in the system. Express the service completion time or the departure time \( D_i \) of the \( i \)th job that brings a random amount of service requirement \( S_i \) as:

\[
D_i = \inf \left\{ x \geq A_i : \int_{A_i}^x \frac{1}{Q(s)} \mu(s)ds \geq S_i \right\}, \quad (10)
\]

where \( A_i \) is the arrival time of the \( i \)th job and \( Q(s) \) is the number of customers in the system at time \( s \).
2.2.3 The response time process

Let $R(t; v)$ denote the entire time that a job spends in the system if it arrives at time $t$ and brings a $v$ amount of service requirement. Since $R(t; v)$ has a what-if characteristic, this is often called virtual response time (or virtual sojourn time) at time $t$. When we use $R(t)$ omitting $v$, we still assume a random service requirement. Our primary interest in the $GI_t/GI_t/1/PS$ queue is the mean response time process $E[R(t)]$ for $t \geq 0$. Note that the stochastic nature of $Q(t)$ in Eq. (10) means that $R(t)$ cannot be obtained conveniently as its FCFS counterpart where the Lindely’s recursion is applicable.

To obtain the response time process $\{R(t), t \geq 0\}$ in a $GI_t/GI_t/1/PS$ queue, we store the path of the queue for every replication of the simulation. The path contains the status of the system at each recording epoch. After a replication is terminated, re-run the simulations from each recording epoch during a replication length (say $t_1, t_2, \ldots$), given the stored status at time $t_k$, with a newly inserted job which is the virtual job. Each re-run of the simulation terminates when the virtual job is finished and results in a sample of a response time at $t_k$, i.e., $R(t_k)$. We obtain the expected process $\{E[R(t)], t \geq 0\}$ by averaging at least 10,000 replications.

3 Methods

As mentioned in Sect. 1, we combine the pointwise stationary approximation (PSA) and the heavy-traffic approximation, which were used by Whitt (2015) and Ma and Whitt (2015) to stabilize the waiting times (excluding service times) in $GI_t/GI_t/1/FCFS$ queues, and adjust the combined approximations to stabilize the response times (waiting time + service time) in $GI_t/GI_t/1/PS$ queues. Below, we explain our methods.

3.1 Pointwise stationary approximation with heavy-traffic limits

We briefly visit the idea of PSA. The PSA was firstly suggested by Green and Kolesar (1991) and further studied by Whitt (1991), which considers that the performance at different times in a nonstationary system is similar to the performance of the stationary counterpart with the instantaneous model parameters. It is known to be an appropriate approximation for when the arrival rate changes slowly relative to the average service time (Whitt 1991, 2015).

The heavy-traffic limit theory for $GI/GI/1/PS$ queues was initially developed by Grishechkin (1994) and further studied by Gromoll (2004) and Zhang and Zwart (2008). Zhang and Zwart (2008) provide the following approximate mean response time ($R$) in steady state for $GI/GI/1/PS$ queues:

$$E[R] \approx \frac{\beta}{\mu} \cdot \frac{1}{1 - \rho} \cdot V_{PS}. \quad (11)$$

3.2 Two service rate controls

Whitt (2015) derived the PSA-based service rate control to stabilize the waiting time. We take a similar approach, but our service rate control stabilizes the response time. We derive two service rate controls based on $GI/GI/1/FCFS$ and $GI/GI/1/PS$ heavy-traffic approx-
imations. Hereinafter, we use the subscripts FCFS and PS to indicate the discipline from which the result derives, e.g., variability factor $V_{FCFS}$ and $V_{PS}$.

### 3.2.1 The square-root (SR) control

In queueing systems, the workload processes are identical under any work-conserving disciplines. Thus, we derive a control based on a $GI/GI/1/FCFS$ queue as an experimental trial, which we later discover to be appropriate for $GI_t/GI_t/1/PS$ queues under light-traffic conditions (see Sect. 4.3.3 for the details).

The heavy-traffic approximation for the expected steady state response time in a $GI/GI/1/FCFS$ queue is:

$$E[R_{FCFS}] \approx \beta \frac{\mu}{\rho} \cdot \frac{1}{1-\rho} \cdot V_{FCFS},$$

(12)

where $\mu$ is the service rate, $\beta$ is the mean job size, $\rho$ is the traffic intensity, and $V_{FCFS} = (C_a^2 + C_s^2)/2$ is the variability parameter, given the SCVs for the arrival base and job size distributions (see Chen and Yao 2001 for the details). The expected response time at time $t$ in a $GI_t/GI_t/1/FCFS$ queue can be approximated based on the PSA by:

$$E[R_{FCFS}(t)] \approx \beta \frac{\mu(t)}{\rho(t)} \cdot \frac{1}{1-\rho(t)} \cdot V_{FCFS},$$

(13)

where $\rho(t) \equiv \lambda(t) \beta/\mu(t)$ is the instantaneous traffic intensity at time $t$. Fixing the LHS by a target response time $s$ and adjusting the terms gives:

$$s \mu(t)^2 - \beta (s \lambda(t) + 1) \mu(t) + \lambda(t) \beta^2 (1 - V_{FCFS}) = 0.$$  

(14)

Finally, the solution to the quadratic equation above can be obtained by:

$$\mu_{SR}(t; s) \equiv \frac{(s \lambda(t) + 1) \beta + \sqrt{(s \lambda(t) + 1)^2 \beta^2 + 4s \lambda(t) \beta^2 (V_{FCFS} - 1)}}{2s}.$$  

(15)

We call Eq. (15) the square-root (SR) control, which is the naming convention used in Whitt (2015).

### 3.2.2 The difference-matching (DM) control

Recall that Eq. (11) is the heavy-traffic approximation for the steady-state mean response time ($R_{PS}$) in a $GI/GI/1/PS$ queue:

$$E[R_{PS}] \approx \frac{\beta}{\mu} \cdot \frac{1}{1-\rho} \cdot V_{PS},$$

where $\mu$, $\beta$, and $\rho$ are defined as in Eq. (12), and $V_{PS} = (C_a^2 + C_s^2)/(1+C_s^2)$ is the variability parameter for the PS queue. The expected response time process can be obtained based on the PSA by:

$$E[R_{PS}(t)] \approx \frac{\beta}{\mu(t)} \cdot \frac{1}{1-\rho(t)} \cdot V_{PS}.$$  

(16)

Fixing the LHS by a certain constant $s$ and adjusting the terms gives a service rate control that is much simpler than $\mu_{SR}(t)$:
\[ \mu_{DM}(t; s) \equiv \beta \left( \lambda(t) + \frac{V_{PS}}{s} \right). \] (17)

As mentioned in Sect. 1, we call Eq. (17) the difference-matching (DM) control because \( \mu_{DM}(t; s) - \beta\lambda(t) \) is a constant \( \beta V_{PS}/s \).

### 3.2.3 Simple analysis on the two service rate controls

The two controls derived above result in different service rate functions except when both base distributions have SCVs 1. The most representative example is the \( M_t/M_t/1/PS \) queue. Applying \( V_{FCFS} = 1 \), the SR control (15) reduces to \( \beta (\lambda(t) + 1/s) \), which is the same as the DM control (17) with \( V_{PS} = 1 \). It prompts the following remark.

**Remark 2** For the time-varying queues having both the base distributions (arrival base and job size) of SCV 1, the two controls coincide.

Another simple but interesting phenomenon is that both controls become identical as we decrease or increase the target response time \( s \).

**Remark 3** The two controls coincide as \( s \to \infty \) (heavy-traffic) or \( s \to 0 \) (light-traffic).

**Proof** Both the SR control in Eq. (15) and the DM control in Eq. (17) converge to \( \beta\lambda(t) \) as \( s \to \infty \) and the traffic intensity converges to 1. On the other hand, taking \( s \to 0 \) results in \( \mu_{SR}(t; s) \to \infty \) and \( \mu_{DM}(t; s) \to \infty \), which implies that the traffic intensity becomes zero.

\[ \square \]

### 4 Simulation experiments

We investigate the performance of the two service rate controls through simulation experiments. Table 1 summarizes the simulation parameters.

#### 4.1 Simulation setting

We use the sinusoidal arrival rate function \( \lambda(t) = a + b \sin(\gamma t) \) with constants \( a = 1 \), \( b = 0.2 \), and \( \gamma = 0.001, 0.01, 0.1 \). Therefore, we have three functions of the same amplitude but of different periods. Two are the slowly time-varying functions (\( \gamma = 0.001, 0.01 \)) and the third one (\( \gamma = 0.1 \)) is not. We include the third, however, to observe how the controls work when the arrival rate is a quickly time-varying function.

To observe the asymptotic behavior, we set the replication length to at least three cycles of the periods, e.g., we conduct simulations for period \( \gamma = 0.001 \) on a 20,000 U time and periods \( \gamma = 0.01, 0.1 \) on a 2000 U time considering the length of periods. For the target response time \( s \), we use two different values: 0.1 for the short and 10.0 for the long response times. Because the service rate controls are inversely proportional to \( s \), each value of \( s \) results in light-traffic and heavy-traffic, respectively. For each independent system, we conduct 10,000 replications to obtain the ensemble average of the performance measures.

We consider three different distributions for arrival base and job size distribution: Erlang distribution (ER); exponential distribution (EXP); and lognormal distribution (LN). The distributions have mean 1 and different SCVs. Specifically, we use ER with \( SCV = 0.5 \) and LN with \( SCV = 2 \). The SCV of EXP is always 1 by definition. We make five pairs of

[Springer]
Table 1  Simulation parameters

| System                  | GI_t/GI_t/1/PS |
|-------------------------|----------------|
| Arrival rate function   | \( \lambda(t) = 1 + 0.2 \sin(\gamma t) \) |
| Periodic coefficient    | \( \gamma = 0.001, 0.01, 0.1 \) |
| Replication length      | \( t = 20,000, 2000, 2000 \) |
| Service rate function   | \( \mu_{SR}(t), \mu_{DM}(t) \) |
| Target response time    | 0.1 (light-traffic), 10.0 (heavy-traffic) |
| Number of replication   | 10,000         |
| Base distribution       | Exponential (SCV = 1.0) |

Table 2  Variability factor for each distribution pair

| Distribution pair (arrival base/job size) | \( V_{FCFS} \) | \( V_{PS} \) |
|-----------------------------------------|--------------|--------------|
| Exponential/Exponential                 | 1            | 1            |
| Erlang/Erlang                           | 0.5          | 0.6667       |
| Lognormal/Lognormal                     | 2            | 1.3333       |
| Erlang/Lognormal                        | 1.25         | 0.8333       |
| Lognormal/Erlang                        | 1.25         | 1.6666       |

4.2 Two metrics to evaluate the effectiveness of the controls

We use two metrics to measure the performance of the two controls. First, we define the relative amplitude (RA) by

\[
\text{amplitude of } \frac{\mathbb{E}[R(t)]}{\text{spatial average of } \mathbb{E}[R(t)]} \times 100%,
\]

as a measure of stabilization. Second, we define the relative gap (RG) by

\[
\text{target response time} - \text{spatial average of } \frac{\mathbb{E}[R(t)]}{\text{spatial average of } \mathbb{E}[R(t)]} \times 100%,
\]

as a measure of accuracy. We obtain two metrics by numerically calculating the following:

\[
\text{RA } [\mathbb{E}[R(t)]] \equiv \frac{T_{E \circ R} \times \left[ \max_{t \in [x, x+T_{E \circ R}]} \{\mathbb{E}[R(t)]\} - \min_{t \in [x, x+T_{E \circ R}]} \{\mathbb{E}[R(t)]\} \right]}{2 \times \int_{x}^{x+T_{E \circ R}} \mathbb{E}[R(t)]dt} \times 100%, \tag{18}
\]

\[
\text{RG } [\mathbb{E}[R(t)]] \equiv \frac{s - \int_{x}^{x+T_{E \circ R}} \mathbb{E}[R(t)]dt}{s} \times 100%, \tag{19}
\]
where $T_{E\circ R}$ is the period of the expected response time process $\mathbb{E}[R(t)]$, $x$ is an arbitrary long time after the process has been stabilized, and $s$ is the target response time. For the values of $T_{E\circ R}$, we use the same values as the periods of the arrival rate functions since we observe that the periods are the same for both $\mathbb{E}[R(t)]$ and $\lambda(t)$.

The two measures above are favorable as they become closer to 0%. For RA, there is no negative value since the amplitude is a positive amount. Note, however, that RG allows a negative value such that the control overestimates the service rate which gives a smaller spatial average than our original target.

4.3 Results

Tables 4, 5, 6, 7 and 8, in Appendix B report both the absolute values (amplitude and spatial average) and the relative values (RA and RG). For the performance of the controls, we heuristically call them good if they control the response time with $RA \leq 10\%$ and $|RG| \leq 0.1\%$, and poor otherwise.

In the following plots, the green line corresponds to the mean queue length $\mathbb{E}[Q(t)]$ and the red line to the mean response time $\mathbb{E}[R(t)]$ of the simulated $GI_t/GI_t/1/PS$ queues under the various combinations of control and distribution. The dotted black lines are the 95% confidence intervals, and the dotted blue line plots the arrival rate function and has its dedicated y-axis on the right.

In the following subsections, we summarize the results of Tables 4, 5, 6, 7 and 8 by their traffic intensity. We obtain each traffic intensity by targeting the response time (short or long) we desire according to Remark 3. Specifically, the instantaneous traffic intensity is approximately $\rho(t) \approx 0.1$ when $s = 0.1$ and $\rho(t) \approx 0.9$ when $s = 10.0$, for the distribution pairs.

4.3.1 Control performances in light-traffic systems ($s = 0.1$)

Figures 4 and 5 depict the two expected processes $\mathbb{E}[Q(t)]$ and $\mathbb{E}[R(t)]$ in light-traffic systems under the two controls where the base distribution pair is Erlang/Erlang. The figures show universally good stabilizing performances ($|RA| \leq 5\%$), even under the quickly time-varying arrival rate ($\gamma = 0.1$), but, the accuracy of the DM control is poor. Specifically, the expected response time process stabilizes around 0.15 although the target is 0.1, which corresponds to about 0.5% of RG (Fig. 5). Intuitively, this poor performance stems from the inaccuracy of the heavy-traffic approximation in light-traffic systems. Meanwhile, the SR control results in only about 0.05% of RG (Fig. 4). Throughout the simulation experiments, we observe this tendency consistently from all of the distribution pairs (see Sect. 4.3.3 for the details).

4.3.2 Control performances in heavy-traffic systems ($s = 10$)

Figure 6 compares the effects of two controls in terms of mean response time under a heavy-traffic condition with three pairs of base distributions (EXP/EXP, ER/ER, LN/LN). The purple and orange colors respectively represent expected response times under the SR and DM controls. Notice that both controls show the same performance when the arrival base/job size distribution pair is EXP/EXP (lines in Fig. 6a–c overlap so that they look like only one line) as Remark 2 tackles since the SCVs are both 1. Comparing with light-traffic systems, we do not observe perfectly controlled results. Specifically for the quickly time-varying arrival rate ($\gamma = 0.1$, Fig. 6a, d, g), poor control performance is obvious since the PSA is not
appropriate. We observe positive results for the slowest time-varying arrival rate \((\gamma = 0.001, \text{ Fig. } 6c, f, i)\) despite the imperfect stabilization. Overall, the consistently better accuracy of the DM control (i.e., the RG under the DM control is smaller than that under SR for all the cases) justifies its use in heavy-traffic systems.

### 4.3.3 Why does DM fail to meet the target response time in light-traffic?

In light-traffic systems, the probability that two or more jobs will present simultaneously becomes smaller, i.e., it is rare that multiple jobs will share the same processor. We recall the following heavy-traffic based PSAs of the expected response time processes [Eqs. (13) and (16) in Sect. 3].
Fig. 6 Comparison between two controls in terms of the mean response time under a heavy-traffic condition ($s = 10$)

(a) EXP/EXP, $\gamma = 0.1$  (b) EXP/EXP, $\gamma = 0.01$  (c) EXP/EXP, $\gamma = 0.001$

(d) ER/ER, $\gamma = 0.1$  (e) ER/ER, $\gamma = 0.01$  (f) ER/ER, $\gamma = 0.001$

(g) LN/LN, $\gamma = 0.1$  (h) LN/LN, $\gamma = 0.01$  (i) LN/LN, $\gamma = 0.001$
Letting $\rho(t) \to 0$, the two approximations above converge, respectively, to

$$\mathbb{E}[R_{FCFS}(t)] \approx \frac{\beta}{\mu(t)},$$

(20)

$$\mathbb{E}[R_{PS}(t)] \approx \frac{\beta}{\mu(t)} \cdot \frac{1}{1 - \rho(t)} \cdot V_{PS},$$

(21)

and then the two controls reduce to the constants:

$$\mu_{SR}(t; s) \equiv \frac{\beta}{s},$$

(22)

$$\mu_{DM}(t; s) \equiv \frac{\beta V_{PS}}{s}.$$

(23)

Throughout the simulation experiments, we use the base distributions having mean $\beta = 1$ and the target mean response time $s = 0.1$ for light-traffic so $\mu_{SR}(t) = 10$ regardless of the distributions. In comparison, $\mu_{DM}(t)$ varies depending on both the base arrival and job size distributions because of variability factor $V_{PS}$.

In a $GI_t/GI_t/1/PS$ queue with a service rate function $\mu(\cdot)$, the response time of a job with random size $S$ and arrival time $t$ denoted by $R(t; S, \mu)$, is expressed as

$$R(t; S, \mu) = \inf \left\{ y > 0 : \int_t^y \frac{\mu(s)}{Q(s)} \, ds \geq S \right\} - t.$$

(24)

Approximating $Q(t) \approx 1$ under the light-traffic condition and letting $\mu(\cdot) \leftarrow \mu_{DM}(\cdot)$, the above expression reduces to the analytic form:

$$R(t; S, \mu_{DM}) = \frac{S}{\mu_{DM}(t)}.$$

(25)

Replacing $S$ by the mean job size $\beta = 1$ obtains the numerical values shown in Table 3. We observe that the simulation results and the approximately calculated values coincide, e.g., $\mathbb{E}[R(t)]$ in Fig. 5c and $R(t; \beta, \mu_{DM})$ for Erlang/Erlang in Table 3 are 0.15. In comparison, we observe that $R(t; S, \mu_{SR})$ is consistently 0.1 regardless of the distributions based on the reasoning we use to obtain the numerical values in Table 3.

We gain two insights into light-traffic systems. First, the PS queue exhibits behavior similar to the FCFS queue. Second, the two service rate controls do not require time dependency. Thus, we conclude that the SR control is appropriate for stabilizing the response times in $GI_t/GI_t/1/PS$ queues as the target response time shortens.

### 4.3.4 Heavy-traffic behavior of the two controls

Figure 7 plots the the result of two controls under different distribution pairs (LN/ER and ER/LN). As we calculated in Sect. 4.3.3, those controls cause significantly different behavior of the system when the target response time is short; see the huge gap between the SR control (red surface) and the DM control (blue surface) where the target response time ($s$) is around zero. The gap diminishes as the target response time becomes larger. However, the difference
Table 3  Approximately calculated expected response times ($s = 0.1, \beta = 1$)

| Distribution pair           | $\mu_{DM}(t)$ | $R(t; \beta, \mu_{DM})$ |
|-----------------------------|---------------|------------------------|
| Exponential/Exponential     | 10            | 0.1                    |
| Erlang/Erlang               | 6.667         | 0.15                   |
| Lognormal/Lognormal         | 13.333        | 0.07                   |
| Erlang/Lognormal            | 8.333         | 0.12                   |
| Lognormal/Erlang            | 16.666        | 0.06                   |

Fig. 7  Comparison of the two controls

(a) LN/ER  
(b) ER/LN

in convergence speeds causes the two controls to perform differently in non-asymptotic heavy-traffic systems, e.g., traffic intensities are around 0.9 (not definitely 1) throughout our heavy-traffic systems.

5 Conclusion

This paper studied the service rate functions that control the mean (virtual) response time required to obtain stabilization in $GI_t/GI_t/1/PS$ queues with slowly time-varying arrival rates. Modifying Whitt’s method for analyzing PS queues resulted in a modified square-root (SR) service rate control and we introduced a new difference-matching (DM) service rate control that appears practically advantageous due to its ease of use and simplicity. Extensive simulation experiments were performed to investigate the performance of two controls. The SR control was effective under a light-traffic condition with a short target response time relative to the inter-arrival times. Neither control, however, perfectly stabilized the response time under a heavy-traffic condition. The DM control outperformed the SR control in terms of meeting the target response time.

We suggest several research directions based on the results presented in this paper. Limit theorems, e.g., fluid and diffusion limits, should be derived for $GI_t/GI_t/1/PS$ queues with periodically time-varying arrival rate functions. We believe that such supporting theories should provide important clues for achieving perfect stabilization of the response time process. Light-traffic behaviors in queueing systems also deserve more analysis, since studies of time-varying queues are rare to the best of our knowledge. Conceivably, interpolating the
two controls could extend the coverage of the target response time beyond short and long. Of course, practical applications in ICT infrastructures should be accompanied.

Acknowledgements  This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1B04933453).

Appendix A: Inversion method to simulate a nonstationary non-Poisson process

For all the simulation experiments conducted in this study, we generate samples from the arrival process \( A(\cdot) \) using the inversion method described in Gerhardt and Nelson (2009). The following Algorithm 1 summarizes the procedure.

Result: An 1-dimensional array \( A \) such that \( A[i] \) is the \( i \)th arrival time

begin
\( T_{\text{run}} \leftarrow \) simulation running time
\( A \leftarrow \) empty 1-dimensional array
\( A[1] \leftarrow \) random number generated by the equilibrium pdf: \( f_{e}(t) = 1 - \frac{F(t)}{E[T]} \)
\( n \leftarrow 1 \)
while \( A[n] < T_{\text{run}} \) do
\( n \leftarrow n + 1 \)
\( x \leftarrow \) random number generated by the df: \( F(\cdot) \) // stationary inter-renewal time
\( A[n] \leftarrow \Lambda^{-1}(x; A[n-1]) \) // \( \Lambda^{-1}(x; a) \equiv \inf \left\{ y \geq a : \int_{a}^{y} \lambda(s)ds \geq x \right\} \)
end
return \( A \)

Algorithm 1: The inversion method to generate an NSNP from a stationary renewal process Gerhardt and Nelson (2009)

Appendix B: Performances of the suggested controls

B.1 Numerical data

See Tables 4, 5, 6, 7 and 8.

Table 4  Control performance of \( \mu_{SR} \) and \( \mu_{DM} : M_t/M_t/1/PS \)

| \( s \) | Window (\%) | \( \mu_{SR} \) Amplitude (RA) | Spatial average (TG) | \( \mu_{DM} \) Amplitude (RA) | Spatial average (TG) |
|---|---|---|---|---|---|
| 0.1 | 0.001 | 0.0044 (4.37%) | 0.1 (0.0%) | 0.0037 (3.7%) | 0.1001 (0.0%) |
| | 0.01 | 0.0035 (3.45%) | 0.1001 (0.0%) | 0.0039 (3.9%) | 0.1001 (0.0%) |
| | 0.1 | 0.003 (2.95%) | 0.1015 (0.02%) | 0.0025 (2.5%) | 0.1012 (0.01%) |
| 10.0 | 0.001 | 0.7555 (7.44%) | 10.1493 (0.01%) | 0.6808 (6.72%) | 10.1321 (0.01%) |
| | 0.01 | 1.7516 (17.07%) | 10.264 (0.03%) | 1.8011 (17.58%) | 10.2452 (0.02%) |
| | 0.1 | 1.1104 (10.82%) | 10.2665 (0.03%) | 1.0656 (10.57%) | 10.0773 (0.01%) |
| $s$ | $\gamma$ | $\mu_{SR}$ | $\mu_{DM}$ |
|-----|------|-------------|-------------|
|     |      | Amplitude (RA) | Spatial average (TG) | Amplitude (RA) | Spatial average (TG) |
| 0.1 | 0.001 | 0.0037 (3.52%) | 0.1047 (0.05%) | 0.005 (3.37%) | 0.1487 (0.49%) |
|     | 0.01  | 0.0039 (3.72%) | 0.1045 (0.04%) | 0.0044 (2.93%) | 0.1485 (0.49%) |
|     | 0.1   | 0.0026 (2.46%) | 0.1064 (0.06%) | 0.0027 (1.78%) | 0.1513 (0.51%) |
| 10.0| 0.001 | 1.1872 (8.64%) | 13.7467 (0.37%) | 0.7828 (7.20%) | 10.8678 (0.09%) |
|     | 0.01  | 2.5553 (18.56%) | 13.765 (0.38%) | 1.8884 (17.19%) | 10.9824 (0.10%) |
|     | 0.1   | 1.4838 (10.74%) | 13.822 (0.38%) | 1.3446 (12.04%) | 11.1719 (0.12%) |

| $s$ | $\gamma$ | $\mu_{SR}$ | $\mu_{DM}$ |
|-----|------|-------------|-------------|
|     |      | Amplitude (RA) | Spatial average (TG) | Amplitude (RA) | Spatial average (TG) |
| 0.1 | 0.001 | 0.0049 (5.27%) | 0.0922 (−0.08%) | 0.004 (5.28%) | 0.0751 (−0.25%) |
|     | 0.01  | 0.0057 (6.14%) | 0.0921 (−0.08%) | 0.0044 (5.88%) | 0.075 (−0.25%) |
|     | 0.1   | 0.0041 (4.38%) | 0.0936 (−0.06%) | 0.0041 (5.42%) | 0.0761 (−0.24%) |
| 10.0| 0.001 | 0.4776 (7.34%) | 6.5104 (−0.35%) | 0.8366 (9.05%) | 9.2406 (−0.08%) |
|     | 0.01  | 0.9325 (14.33%) | 6.5061 (−0.35%) | 1.6268 (17.64%) | 9.2206 (−0.08%) |
|     | 0.1   | 0.8908 (13.47%) | 6.6132 (−0.34%) | 0.9665 (10.3%) | 9.3818 (−0.06%) |

| $s$ | $\gamma$ | $\mu_{SR}$ | $\mu_{DM}$ |
|-----|------|-------------|-------------|
|     |      | Amplitude (RA) | Spatial average (TG) | Amplitude (RA) | Spatial average (TG) |
| 0.1 | 0.001 | 0.0049 (5.07%) | 0.0976 (−0.02%) | 0.0055 (4.56%) | 0.1197 (0.2%) |
|     | 0.01  | 0.0057 (6.14%) | 0.0974 (−0.03%) | 0.0045 (4.22%) | 0.1193 (0.19%) |
|     | 0.1   | 0.0031 (3.08%) | 0.0995 (−0.01%) | 0.0046 (3.8%) | 0.1209 (0.21%) |
| 10.0| 0.001 | 0.5891 (8.26%) | 7.1291 (−0.29%) | 1.0496 (10.06%) | 10.4293 (0.04%) |
|     | 0.01  | 1.2382 (17.11%) | 7.2366 (−0.28%) | 2.0119 (19.39%) | 10.375 (0.04%) |
|     | 0.1   | 0.8216 (11.29%) | 7.2763 (−0.27%) | 1.2382 (11.84%) | 10.456 (0.05%) |

| $s$ | $\gamma$ | $\mu_{SR}$ | $\mu_{DM}$ |
|-----|------|-------------|-------------|
|     |      | Amplitude (RA) | Spatial average (TG) | Amplitude (RA) | Spatial average (TG) |
| 0.1 | 0.001 | 0.0034 (3.51%) | 0.0978 (−0.02%) | 0.003 (4.91%) | 0.0601 (−0.4%) |
|     | 0.01  | 0.0039 (3.99%) | 0.0977 (−0.02%) | 0.003 (4.97%) | 0.0602 (−0.4%) |
|     | 0.1   | 0.0026 (2.64%) | 0.0997 (−0.00%) | 0.0023 (3.82%) | 0.0609 (−0.39%) |
| 10.0| 0.001 | 0.6518 (5.63%) | 11.5748 (0.16%) | 0.4418 (5.43%) | 8.1343 (−0.19%) |
|     | 0.01  | 1.8191 (15.4%) | 11.8151 (0.18%) | 1.0204 (12.49%) | 8.1706 (−0.18%) |
|     | 0.1   | 1.223 (10.41%) | 11.7491 (0.17%) | 0.9637 (11.56%) | 8.3389 (−0.17%) |
B.2 Plots

See Figs. 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17.

$GI_2GI_1/1/PS$, Target response time = 0.1, Arrival=Exponential, Service=Exponential

Fig. 8 General performance measures of $M_t/M_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 0.1 ($s = 0.1$)

---

$E[Q(t)]$ and $E[R(t)]$ are plotted against time for both square-root control and difference-matching control. The plots show the expected queue length and response time over time for two different control strategies, with $\mu_{SR}$ and $\mu_{DM}$.

---

Note: The actual figures are not included in the text representation, but the descriptions and expected behaviors of the systems are clear.
Fig. 9 General performance measures of $Mt/Mt/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 10.0 ($s = 10.0$)
Fig. 10  General performance measures of $ER_t/ER_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 0.1 ($s = 0.1$)
Fig. 11  General performance measures of $E R_t / E R_t / 1 / PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 10.0 ($s = 10.0$)
$GI_1/GI_1/1/PS$, Target response time = 0.1, Arrival=Lognormal, Service=Lognormal

Fig. 12  General performance measures of $LN_t/LN_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 0.1 ($s = 0.1$)
Fig. 13  General performance measures of $LN_t/LN_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 10.0 ($s = 10.0$)
**Fig. 14** General performance measures of $ER_t/\ln_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time $0.1$ ($s = 0.1$)
$G_t/G_t/1/PS$, Target response time = 10.0, Arrival=Erlang, Service=Lognormal

Fig. 15  General performance measures of $ER_t/LN_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 10.0 ($s = 10.0$)
$G_l/G_l/1/PS$, Target response time = 0.1, Arrival=Lognormal, Service=Erlang

**Fig. 16** General performance measures of $LN_l/ER_l/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 0.1 ($s = 0.1$)
Fig. 17  General performance measures of $LN_t/ER_t/1/PS$ queues under $\mu_{SR}$ and $\mu_{DM}$ with target response time 10.0 ($s = 10.0$)

References

Anselmi, J., & Verloop, M. (2011). Energy-aware capacity scaling in virtualized environments with performance guarantees. *Performance Evaluation, 68*(11), 1207–1221.

CAIDA. (2016). *CAIDA internet data: Realtime monitors*. Retrieved August 24, 2018 from http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirB

Chen, H., & Yao, D. D. (2001). *Fundamentals of queueing networks*. New York, NY: Springer.
Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science, 54*(2), 324–338.

Gautam, N. (2012). *Analysis of queues: Methods and applications*. Boca Raton, FL: Taylor and Francis.

Gerhardt, I., & Nelson, B. L. (2009). Transforming renewal processes for simulation of nonstationary arrival processes. *INFORMS Journal on Computing, 21*(4), 630–640.

Green, L., & Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science, 37*(1), 84–97.

Grishechkin, S. (1994). GI/G/1 processor sharing queue in heavy traffic. *Advances in Applied Probability, 26*(2), 539–555.

Gromoll, H. C. (2004). Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability, 14*(2), 555–611.

He, B., Liu, Y., & Whitt, W. (2016). Staffing a service system with non-Poisson non-stationary arrivals. *Probability in the Engineering and Informational Sciences, 30*(4), 593–621.

Horvath, T., Abdelzaher, T., Skadron, K., & Liu, X. (2007). Dynamic voltage scaling in multitier web servers with end-to-end delay control. *IEEE Transactions on Computers, 56*(4), 444–458.

Jennings, O. B., Mandelbaum, A., Massey, W. A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science, 42*(10), 1383–1394.

Ko, Y. M., & Cho, Y. (2014). A distributed speed scaling and load balancing algorithm for energy efficient data centers. *Performance Evaluation, 79*, 120–133.

Kwon, S., & Gautam, N. (2016a). Guaranteeing performance based on time-stability for energy-efficient data centers. *IIE Transactions, 48*(9), 812–825.

Kwon, S., & Gautam, N. (2016b). Time-stable performance in parallel queues with non-homogeneous and multi-class workloads. *IEEE/ACM Transactions on Networking, 24*(3), 1322–1335.

Liao, D., Li, K., Sun, G., Anand, V., Gong, Y., & Tan, Z. (2015). Energy and performance management in large data centers: A queuing theory perspective. In *2015 international conference on computing, networking and communications (ICNC)* (pp. 287–291).

Liu, Y. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research, 66*(2), 514–534.

Liu, Y., & Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research, 60*(6), 1551–1564.

Liu, Y., & Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences, 28*(4), 419–449.

Liu, Y., & Whitt, W. (2017). Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research, 256*(2), 473–486.

Ma, N., & Whitt, W. (2015). Using simulation to study service-rate controls to stabilize performance in a single-server queue with time-varying arrival rate. In *2015 winter simulation conference (WSC)* (pp. 2598–2609).

Ma, N., & Whitt, W. (2016). Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters, 109*, 202–207.

Ma, N., & Whitt, W. (2018). A rare-event simulation algorithm for periodic single-server queues. *INFORMS Journal on Computing, 30*(1), 71–89.

Ma, N., & Whitt, W. (2019). Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems, 9*(3), 261–290. [https://doi.org/10.1287/stsy.2018.0027](https://doi.org/10.1287/stsy.2018.0027).

Whitt, W. (1991). The pointwise stationary approximation for \( M_t / M_t / s \) queues is asymptotically correct as the rates increase. *Management Science, 37*(3), 307–314.

Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems, 81*(4), 341–378.

Zhang, J., & Zwart, B. (2008). Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems, 60*(3), 227–246.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.