*Adv. Appl. Prob.* **40,** 548–577 (2008)
*Printed in Northern Ireland*

# STEADY-STATE ANALYSIS OF A MULTISERVER QUEUE IN THE HALFIN–WHITT REGIME

DAVID GAMARNIK,* *Massachusetts Institute of Technology*
PETAR MOMČILOVIĆ,** *University of Michigan*

## Abstract

We consider a multiserver queue in the Halfin–Whitt regime: as the number of servers $n$ grows without a bound, the utilization approaches 1 from below at the rate $\Theta(1/\sqrt{n})$. Assuming that the service time distribution is lattice valued with a finite support, we characterize the limiting scaled stationary queue length distribution in terms of the stationary distribution of an explicitly constructed Markov chain. Furthermore, we obtain an explicit expression for the critical exponent for the moment generating function of a limiting stationary queue length. This exponent has a compact representation in terms of three parameters: the amount of spare capacity and the coefficients of variation of interarrival and service times. Interestingly, it matches an analogous exponent corresponding to a single-server queue in the conventional heavy-traffic regime.

*Keywords:* Multiserver queue; heavy-traffic approximation; Halfin–Whitt (QED) regime

2000 Mathematics Subject Classification: Primary 60K25
Secondary 90B22

## 1. Introduction

In their seminal paper Halfin and Whitt [22] formally introduced an unconventional heavy-traffic regime for queueing models (dubbed thereafter the Halfin–Whitt regime). Unlike in the traditional heavy-traffic approach, in their regime high utilization is achieved by simultaneously increasing the arrival rate *and* the number of servers. This regime is also referred to as the quality- and efficiency-driven (QED) regime, since it balances between the system utilization and quality of service perceived by customers. Moreover, the QED regime can be understood as critical with respect to the delay probability, i.e. the limiting delay probability is *strictly* in (0, 1) in QED systems (the delay probabilities 0 and 1 correspond to the quality-driven and efficiency-driven regimes, respectively). It should be noted that the QED regime was considered by Erlang [15] in the context of numerical steady-state analysis of M/M/$n$ and M/M/$n$/$n$ systems. An asymptotic analysis of the closely related Erlang loss function was carried out in [24]. A formal analysis of a queue with exponential service times in the QED regime was completed by Halfin and Whitt [22]. They established the criticality of the delay probability in terms of the square root spare capacity rule, both in steady-state and transient regimes.

Queueing models in the QED regime have found applications primarily in the area of large-scale call and customer contact centers [1], [18]. Hence, a number of related models have been considered in the literature. Models with customer impatience relevant to call center

Received 25 October 2007; revision received 15 March 2008.

* Postal address: Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email address: gamarnik@mit.edu
** Postal address: EECS Department, University of Michigan, Ann Arbor, MI 48109, USA.
Email address: petar@eecs.umich.edu

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

management were studied in [16], [19], and [41]. Approximations that take into account finiteness of buffers were introduced in [39] and [40]. Revenue maximization and constraint satisfaction were considered in [2], [3], [11], [30], and [32]. Optimal stochastic control of QED queues in various settings was examined in [5], [6], [23], and [37]; the problem of joint control and staffing was studied in [7] and [20]. Most of the aforementioned results assume exponential service times. This assumption significantly simplifies the analysis as we do not need to keep a track of residual service times. The literature on QED systems with nonexponential service time distribution is limited. Phase-type service time distribution in the transient regime was considered in [35]. The case of deterministic service times in the steady-state regime was examined in [25]. A more recent work [31] dealt with the transient distribution of the virtual waiting time in the case of discrete service times with a finite support. A process-level limit for the G/GI/$n$ QED queue with general service time distributions was recently obtained in [36].

In this paper we examine the stationary behavior of a GI/GI/$n$ system in the Halfin–Whitt regime when service times are lattice valued with a finite support. More specifically, we consider a sequence of first-come–first-served queues indexed by the number of servers $n \to \infty$. The utilization in the $n$th system is $1 - \beta/\sqrt{n} + o(1/\sqrt{n})$ for some parameter $\beta > 0$; equivalently, the number of servers $n$ is given by $R_n + \beta\sqrt{R_n} + o(\sqrt{R_n})$, where $R_n$ is the offered load of the $n$th system. The service distribution does not change with $n$. The stationary number of customers and waiting time in the $n$th system are denoted by $Q^n$ and $W^n$, respectively. The first main result of the paper states the existence of limiting random variables $\hat{Q}$ and $\hat{W}$ such that $Q^n/\sqrt{n} \xrightarrow{\mathrm{D}} \hat{Q}$ and $\sqrt{n}W^n \xrightarrow{\mathrm{D}} \hat{W}$ as $n \to \infty$, where '$\xrightarrow{\mathrm{D}}$' denotes convergence in distribution. The distribution of $\hat{Q}$ is shown to correspond to the unique stationary distribution of some underlying continuous-state Markov chain $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$, where $\{\hat{\boldsymbol{L}}_t,\ t \in \mathbb{Z}_+\}$ is a limiting process corresponding to the vector of customers in different stages of service. Our second main result identifies the exact exponential decay rate of the limiting variable $\hat{Q}$. Informally, we show that $\mathrm{P}[\hat{Q} > x] \approx \exp(-2\beta x/(c_a^2 + c_s^2))$ for large $x$, where $c_a$ is the (limiting) coefficient of variation of interarrival times and $c_s$ is the coefficient of variation of service times. Our analysis uses quadratic and geometric Lyapunov functions to establish the tightness of the sequences $\{Q^n/\sqrt{n},\ n \geq 1\}$ and $\{\sqrt{n}W^n,\ n \geq 1\}$.

Next we list some notational conventions used throughout the paper. For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with elements $x_i$ and $y_i$, respectively, $\boldsymbol{x} \cdot \boldsymbol{y}$ denotes the dot product $\sum_i x_i y_i$. All considered vectors are row vectors, and transposition of a vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}^\top$. Let $\boldsymbol{K} := (1, 2, \ldots, K)$. For $\boldsymbol{x} \in \mathbb{R}^m$, $\|\boldsymbol{x}\|$ denotes the $L_1$-norm, i.e. $\|\boldsymbol{x}\| = \sum_{i=1}^m |x_i|$. Denote by $\mathcal{T} : \mathbb{R}^K \to \mathbb{R}^K$ a linear operator defined by

$$\mathcal{T}\{(x_1, \ldots, x_K)\} = (x_2, \ldots, x_K, 0).$$

Given a random variable (RV) $X \in \mathbb{R}$, its moment generating function is $M_X(\theta) := \mathrm{E}[\mathrm{e}^{\theta X}]$. For every $\theta > 0$, we denote by $\mathcal{M}_\theta$ the family of sequences of RVs $\{X_n,\ n \geq 1\}$ such that $\limsup_{n\to\infty} M_{X_n}(\theta) < \infty$; let $\mathcal{M}_\infty = \bigcap_{\theta>0} \mathcal{M}_\theta$. Given an RV $X$, we write $X \in \mathcal{M}_\theta$ or $X \in \mathcal{M}_\infty$ if $\mathrm{E}[\mathrm{e}^{\theta X}] < \infty$ or $\mathrm{E}[\mathrm{e}^{\theta X}] < \infty$ for every $\theta > 0$. We denote by $\mathrm{E}_\pi[\cdot]$ the expectation operator with respect to a probability measure $\pi$; similarly, we use $\mathrm{P}_\pi[\cdot]$ when the probability measure $\pi$ is not clear from the context. For two reals $x$, $y$, we set $x \wedge y = \min\{x, y\}$, $x \vee y = \max\{x, y\}$, $x^+ = x \vee 0$, and $x^- = (-x)^+$; when the argument of a unary operation is a vector or matrix, it is understood that the operator is applied elementwise. The symbols $\mathbb{Z}_+$ and $\mathbb{R}_+$ denote nonnegative integers and reals, respectively.

The paper is organized as follows. In Section 2 we describe the considered model and formally introduce the Halfin–Whitt (QED) regime. Our main results are stated in Section 3.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

Section 4 contains preliminary results. The proofs of the main results can be found in Sections 5, 6, and 7.

## 2. Model

### 2.1. Queueing system description

We consider a sequence of first-come–first-served queues indexed by the number of servers $n$. Details of our model are as follows.

2.1.1. *Service times.* Service times are independent and identically distributed (i.i.d.) RVs, equal in distribution to an RV $S$ that does not depend on $n$ and takes values in a finite set $\{s_1, \dots, s_K\} \subset \mathbb{R}_+$. It is assumed that the set of service time values has a common divisor $s > 0$, i.e. $s_i = k_i s$ for some $k_i \in \mathbb{N}$, $1 \le i \le K$. Under this assumption, without loss of generality, we adopt $s = 1$ to be the largest common divisor of service time values. Let

$$p_i := \mathrm{P}[S = i], \qquad 0 \le i \le K,$$

where $K$ is the largest index such that $p_K > 0$. We assume that $p_0 = 0$, that is, no instantaneous service is possible. Then the expected service time is $\mu^{-1} := \mathrm{E}[S] = \sum_{i=1}^{K} i p_i$; the variance of $S$ is denoted by $\sigma_s$ and the coefficient of variation by $c_s = \mu\sigma_s$. The steady-state behavior of the system with deterministic service times ($S = 1$) has been characterized in [25] and, thus, we consider $\sigma_s > 0$. In this case there exist two values of the service time that are relatively prime, i.e. $p_i p_j > 0$ for some relatively prime $i \neq j$; otherwise a simple time change argument can be applied to rescale the service times. For convenience, let $\boldsymbol{p} = (p_1, \dots, p_K)$ and $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \dots, \tilde{p}_K)$, where $\tilde{p}_i = \mathrm{P}[S \ge i] = \sum_{j \ge i} p_j$ describes the tail of the service time distribution.

2.1.2. *Arrival times.* Customers arrive to the $n$th system according to a stationary renewal process with interarrival times equal in distribution to $\tau_n$. The arrival rate $\lambda_n := 1/\mathrm{E}[\tau_n]$ is such that $\lambda_n \to \infty$ as $n \to \infty$, while the coefficient of variation of interarrival times, $c_{a,n}$, satisfies $c_{a,n} \to c_a$ as $n \to \infty$ for some $0 \le c_a < \infty$. In view of the assumption that $S \in \mathbb{N}$ ($s = 1$), it is convenient to define $A_t^n$, $t \in \mathbb{R}$, as the number of arrivals in the time interval $(t - 1, t]$ in the $n$th system. In addition, let $a_t^n$ denote the backward recurrence time of the arrival process at time $t$, i.e. $a_t^n := \inf\{u > 0 \colon A_{t-u,t}^n > 0\}$, where $A_{s,t}^n$ is the number of arrivals in the time interval $(s, t]$ for two reals $s < t$. Our proving method is based on an analysis of a time-embedded process that has the Markov property. Hence, we require that the arrival process has limited dependency in its structure. To this end, it is assumed that the appropriately scaled number of arrivals, conditioned on the particular value of the backward recurrence time $a$, converges to a Gaussian distribution *uniformly* in $a$, i.e. for every $t \in \mathbb{R}$,

$$\sup_{a \ge 0} \left| \mathrm{P}\left[ \frac{A_t^n - \lambda_n}{\sqrt{\lambda_n}} \le x \;\middle|\; a_{t-1}^n = a \right] - \mathrm{P}[A \le x] \right| \to 0 \quad \text{as } n \to \infty, \tag{1}$$

where $A$ is normally distributed with zero mean and variance $c_a^2$. We assume additionally (since convergence in distribution does not necessarily imply the convergence of moments) that

$$\sup_{a \ge 0} \mathrm{E}\left[ \frac{A_t^n - \lambda_n}{\sqrt{\lambda_n}} \;\middle|\; a_{t-1}^n = a \right] \to 0 \quad \text{as } n \to \infty \tag{2}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

and

$$\limsup_{n\to\infty}\sup_{a\geq 0}\mathrm{E}\biggl[\biggl(\frac{A_t^n-\lambda_n}{\sqrt{\lambda_n}}\biggr)^2 \biggm| a_{t-1}^n=a\biggr]<\infty. \tag{3}$$

There exists a broad class of arrival processes that satisfy these assumptions. The simplest one is the class of renewal processes with interarrival times that have bounded conditional second moments uniformly in $a_t^n = a$. For example, let $\{\zeta_i,\ i\in\mathbb{Z}\}$ be an i.i.d. sequence of nonnegative RVs with unit mean and a finite second moment. By setting $\zeta_i/\lambda_n$ to be the $i$th interarrival time in the $n$th process we obtain a process that satisfies the aforementioned assumptions due to the central limit theorem for renewal processes [14, p. 114] when $\lambda_n\to\infty$ as $n\to\infty$.

Finally, since we consider multiserver queues in their steady states, the distribution of interarrival times should be such that the stationary distributions of all considered quantities exist and are unique (for all finite $n$). See the comments at the beginning of Section 3 and [4, Chapter XII] for details.

2.1.3. *Quantities of interest.* The number of customers *awaiting service* in the $n$th queue at time $t$ is denoted by $Q_t^n$, and the *total* number of customers in the system is denoted by $Y_t^n$. The fact that $Y_t^n = n + Q_t^n$ when all servers are busy, while $Q_t^n = 0$ when at least one server is idle renders

$$Q_t^n = (Y_t^n - n)^+ \tag{4}$$

for every time instant $t$. Let $L_{t,k}^n$, $k=1,\dots,K$, be the number of customers in service with remaining service times in the interval $(k-1,k]$ at time $t$. The notation $\boldsymbol{L}_t^n = (L_{t,1}^n,\dots,L_{t,K}^n)$ renders $\|\boldsymbol{L}_t^n\|\leq n$, with strict equality corresponding to the case when at least one server is idle. The following identity then holds for all $t\in\mathbb{R}_+$:

$$Q_t^n(n-\|\boldsymbol{L}_t^n\|)=0. \tag{5}$$

Let $J_{t,k}^n$, $k=1,\dots,K$, be the number of customers with service requirement $k$ that enter service during the time interval $(t-1,t]$; set $\boldsymbol{J}_t^n = (J_{t,1}^n,\dots,J_{t,K}^n)$. Thus, $\|\boldsymbol{J}_t^n\|$ is the total number of customers that enter service during the time interval $(t-1,t]$ and

$$Q_{t+1}^n = Q_t^n + A_{t+1}^n - \|\boldsymbol{J}_{t+1}^n\|. \tag{6}$$

## 2.2. QED regime and scaling

The offered load in the $n$th system is $\lambda_n/\mu$ and, hence, the utilization is given by $\rho_n := \lambda_n/n\mu$. In the Halfin–Whitt (QED) regime the relationship between the utilization and number of servers satisfies

$$\sqrt{n}(1-\rho_n)\to\beta \quad \text{as } n\to\infty \tag{7}$$

for some $\beta>0$, or, equivalently, $n=\lambda_n/\mu+\beta\sqrt{\lambda_n/\mu}+o(\sqrt{\lambda_n/\mu})$ as $n\to\infty$. For notational simplicity, we let $\beta_n$ be a quantity satisfying $n=\lambda_n/\mu+\beta_n\sqrt{n}$, i.e.

$$\beta_n := \frac{n-\lambda_n/\mu}{\sqrt{n}}\to\beta \quad \text{as } n\to\infty.$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

Under such a scaling, the following centered and scaled versions of RVs indexed by $t \in \mathbb{R}_+$ are of interest:

$$\hat{A}_t^n := \frac{A_t^n - \lambda_n}{\sqrt{n}},$$

$$\hat{Q}_t^n := \frac{Q_t^n}{\sqrt{n}},$$

$$\hat{\boldsymbol{L}}_t^n := \frac{\boldsymbol{L}_t^n - \lambda_n \tilde{\boldsymbol{p}}}{\sqrt{n}},$$

$$\hat{J}_t^n := \frac{\|\boldsymbol{J}_t^n\| - \lambda_n}{\sqrt{n}}, \tag{8}$$

$$\hat{\boldsymbol{J}}_t^n := \frac{\boldsymbol{J}_t^n - \|\boldsymbol{J}_t^n\| \boldsymbol{p}}{\sqrt{n}},$$

$$\hat{Y}_t^n := \frac{Y_t^n - \lambda_n/\mu}{\sqrt{n}}. \tag{9}$$

Given these definitions, the counterparts of (4), (5), and (6) are

$$\hat{Q}_t^n = (\hat{Y}_t^n - \beta_n)^+, \tag{10}$$

$$\hat{Q}_t^n \left( \beta_n - \sum_{k=1}^{K} \hat{L}_{t,k}^n \right) = 0,$$

and

$$\hat{Q}_{t+1}^n = \hat{Q}_t^n + \hat{A}_{t+1}^n - \hat{J}_{t+1}^n, \tag{11}$$

respectively.

## 3. Main results

A multiserver queue can be described by the standard Kiefer–Wolfowitz vector [26] of residual workloads; see, e.g. [4, Section 2.3] and [8, Chapter XII]. Provided that the stability condition $\rho_n = \lambda_n / n\mu < 1$ is satisfied and the arrival process is renewal, in [26] it was established that all relevant stationary measures exist when the system is observed just before arrivals, i.e. stationary measures exist for this particular time-embedded process. In order to ensure the existence of stationary probabilities for continuous-time processes $\{(Q_t^n, \boldsymbol{L}_t^n),\ t \in \mathbb{R}_+\}$ and $\{W_t^n,\ t \in \mathbb{R}_+\}$, additional conditions are needed [4, p. 348]. We assume that these stationary distributions exist and are unique. Let $\pi_n$ be the stationary distribution of $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n),\ t \in \mathbb{R}_+\}$, i.e. $\pi_n$ is time invariant with respect to $t$. We characterize the limit of $\pi_n$ as $n \to \infty$ in terms of the stationary probability of a certain *discrete-time* process $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$. Although the processes $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n),\ t \in \mathbb{R}_+\}$ are inherently continuous time, for the purposes of characterizing their stationary distributions, it is sufficient to consider their time-embedded versions ($t \in \mathbb{Z}_+$ due to the lattice-valued nature of service times, $S \in \mathbb{N}$). Such an approach has an advantage since these discrete-time processes have a tractable Markovian structure that is amenable to the Lyapunov function method [33].

Next we construct the Markov chain $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$ with state space $\mathbb{R}^{K+1}$. To this end, let $\{\hat{A}_t,\ t \in \mathbb{Z}_+\}$ be an i.i.d. sequence of zero mean normal RVs with variance $\mu c_a^2$. Also, let $\{\hat{\boldsymbol{J}}_t,\ t \in \mathbb{Z}_+\}$ be an i.i.d. sequence of normal random vectors with the zero mean and covariance

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

matrix $\mu\boldsymbol{\Sigma}$, where the elements of $\boldsymbol{\Sigma}$ are defined by

$$\Sigma_{ij} = \begin{cases} (1-p_i)p_i, & 1 \le i = j \le K, \\ -p_i p_j, & 1 \le i \ne j \le K; \end{cases} \tag{12}$$

the sequences $\{\hat{A}_t,\ t \in \mathbb{Z}\}$ and $\{\hat{\boldsymbol{J}}_t,\ t \in \mathbb{Z}\}$ are mutually independent. The process

$$\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$$

is defined by the following three recursions:

$$\hat{\boldsymbol{L}}_{t+1} = \mathcal{T}\{\hat{\boldsymbol{L}}_t\} + \hat{\boldsymbol{J}}_{t+1} + \hat{J}_{t+1}\boldsymbol{p}, \tag{13}$$

$$\hat{Q}_{t+1} = \left(\hat{Q}_t + \hat{A}_{t+1} + \sum_{k=2}^{K} \hat{L}_{t,k} - \beta\right)^+, \tag{14}$$

$$\hat{J}_{t+1} = (\hat{Q}_t + \hat{A}_{t+1}) \wedge \left(\beta - \sum_{k=2}^{K} \hat{L}_{t,k}\right), \tag{15}$$

and an initial condition $(\hat{Q}_0, \hat{\boldsymbol{L}}_0)$ that is independent of $\{\hat{A}_t,\ t \in \mathbb{Z}_+\}$ and $\{\hat{\boldsymbol{J}}_t,\ t \in \mathbb{Z}_+\}$; the random vector $(\hat{Q}_0, \hat{\boldsymbol{L}}_0)$ satisfies $\sum_{k=1}^{K} \hat{L}_{0,k} \le \beta$ and $\hat{Q}_0(\sum_{k=1}^{K} \hat{L}_{0,k} - \beta) = 0$ by definition. It is straightforward to verify that the preceding defines a continuous-state Markov chain due to the i.i.d. nature of $\{\hat{A}_t,\ t \in \mathbb{Z}_+\}$ and $\{\hat{\boldsymbol{J}}_t,\ t \in \mathbb{Z}_+\}$. Observe that (13), (14), and (15) imply that, for all $t \in \mathbb{N}$, $\sum_{k=1}^{K} \hat{L}_{t,k} \le \beta$ and

$$\hat{Q}_t\left(\sum_{k=1}^{K} \hat{L}_{t,k} - \beta\right) = 0.$$

We define a process $\{\hat{Y}_t,\ t \in \mathbb{Z}_+\}$ by $\hat{Y}_t = \sum_{k=1}^{K} \hat{L}_{t,k} + \hat{Q}_t$ and note that it satisfies $\hat{Q}_t = (\hat{Y}_t - \beta)^+$, $t \in \mathbb{Z}_+$; we refer to this process as the limiting number of customers in the queue.

Our first main result states the existence of a distributional limit of $(\hat{Q}^n, \hat{\boldsymbol{L}}^n)$ as $n \to \infty$, where the pair $(\hat{Q}^n, \hat{\boldsymbol{L}}^n)$ is distributed according to $\pi_n$. In particular, we relate the sequence of stationary distributions $\{\pi_n,\ n \ge 1\}$ of $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n),\ t \in \mathbb{R}_+\}$ to the stationary distribution of the discrete-time chain $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$. The proof is based on a tightness argument and can be found in Section 5.

**Theorem 1.** *We have* $\pi_n \xrightarrow{\mathrm{D}} \pi_*$ *as* $n \to \infty$, *where* $\pi_*$ *is the unique stationary distribution of the Markov chain* $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$.

*Outline of the proof.* The proof consists of three parts: (i) demonstrating that the sequence $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n),\ n \ge 1\}$ is tight with respect to the sequence of distributions $\{\pi_n,\ n \ge 1\}$ (as $n \to \infty$), (ii) showing that the stationary distribution of $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n),\ t \in \mathbb{R}_+\}$ converges to a stationary distribution of $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$ as $n \to \infty$, and (iii) proving that $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$ has a unique stationary distribution $\pi_*$. We briefly outline the main argument for part (i), as the proofs of parts (ii) and (iii) follow more or less a standard argument.

A polynomial function $\Psi_\theta(\boldsymbol{y}, \boldsymbol{z}) = (\tilde{\boldsymbol{p}} \cdot \boldsymbol{y} + \boldsymbol{\alpha} \cdot \boldsymbol{z})^\theta$ is defined, with $\boldsymbol{\alpha} \in \mathbb{R}^{K^2}$ being fixed (see Section 4.3); the function $\Psi_1$ can take negative values. For notational simplicity, let $\bar{\boldsymbol{Y}}_t^n = (\hat{Y}_t^n, \dots, \hat{Y}_{t-K+1}^n)$ and $\bar{\boldsymbol{Z}}_t^n = (\hat{\boldsymbol{Z}}_t^n, \dots, \hat{\boldsymbol{Z}}_{t-K+1}^n)$, where $\hat{\boldsymbol{Z}}_t^n = \hat{\boldsymbol{J}}_t^n + \boldsymbol{p}\hat{A}_t^n$. Based on

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

preliminary results (see Sections 4.1 and 4.2), the following is derived (see Section 4.3) for an explicitly constructed set $\mathcal{R}^n$:

$$\mathrm{E}[\Psi_2(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}^n_{t-1} \notin \mathcal{R}^n\} - \Psi_2(\bar{\boldsymbol{Y}}^n_{t-1}, \bar{\boldsymbol{Z}}^n_{t-1}) \mid \bar{\boldsymbol{Y}}^n_{t-1}, \bar{\boldsymbol{Z}}^n_{t-1}] \le -\delta\Psi_1(\bar{\boldsymbol{Y}}^n_{t-1}, \bar{\boldsymbol{Z}}^n_{t-1}) + \psi$$

for some $\delta > 0$, $\psi < \infty$, and all large enough $n$ (see Proposition 2, below), and

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Psi_2(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}^n_{t-1}, \bar{\boldsymbol{Z}}^n_{t-1} \in \mathcal{R}^n\}] < \infty$$

(see Lemma 4, below). These two relationships can be combined (see Theorem 4, below) to obtain

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Psi_1(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)] < \infty.$$

On the other hand, the expectation of the negative part of $\Psi_1(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)$ is also bounded in the limit (see Lemma 6, below):

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[-\Psi_1(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)\,\mathbf{1}\{\Psi_1(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t) < 0\}] < \infty.$$

Finally, the tightness of $\{\hat{Y}^n_t,\ n \ge 1\}$ (and hence of $\{\hat{Q}^n_t,\ n \ge 1\}$, since $\hat{Q}^n_t = (\hat{Y}^n_t - \beta_n)^+$) with respect to stationary $\{\pi_n,\ n \ge 1\}$ is due to $\mathrm{E}_{\pi_n}[\Psi_1(\bar{\boldsymbol{Y}}^n_t, \bar{\boldsymbol{Z}}^n_t)] = \mathrm{E}_{\pi_n}[\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}^n_t] = \mathrm{E}_{\pi_n}[\hat{Y}^n_t]/\mu$.

Let $(\hat{Q}, \hat{\boldsymbol{L}})$ be distributed according to $\pi_*$, i.e. if $Q^n$ is the stationary number of customers in the $n$th queue then $Q^n/\sqrt{n} \xrightarrow{\mathrm{D}} \hat{Q}$ as $n \to \infty$. It is immediate that $\mathrm{P}[\hat{Q} = 0] \in (0, 1)$, since the distribution of the Gaussian term $\hat{A}_t$ in (13), (14), and (15) has infinite support. The convergence $\pi_n \xrightarrow{\mathrm{D}} \pi_*$ implies that $\mathrm{P}[\hat{Q}^n = 0] \to \mathrm{P}[\hat{Q} = 0] \in (0, 1)$ as $n \to \infty$, and, thus, the system is indeed in the QED regime.

Our second result establishes the critical exponent for the moment generating function of $\hat{Q}$. The proof can be found in Section 6. The theorem is stated for the limiting queue length $\hat{Q}$. With additional conditions on the arrival processes, a weaker result can be obtained for the prelimit variables $\hat{Q}^n$ by slightly modifying the proof of Theorem 2.

**Theorem 2.** *Let* $\theta^* = 2\beta/(c_a^2 + c_s^2)$. *Then* $\mathrm{E}[\mathrm{e}^{\theta\hat{Q}}] < \infty$ *if* $\theta < \theta^*$ *and* $\mathrm{E}[\mathrm{e}^{\theta\hat{Q}}] = \infty$ *if* $\theta > \theta^*$.

*Outline of the proof.* Here we outline just the proof of the statement $\mathrm{E}[\mathrm{e}^{\theta\hat{Q}}] < \infty$ if $\theta < \theta^*$. The key idea is to define a geometric Lyapunov function $\Phi_\theta(\boldsymbol{y}, \boldsymbol{z}) = \exp(\theta\tilde{\boldsymbol{p}} \cdot \boldsymbol{y} + \theta\boldsymbol{\alpha} \cdot \boldsymbol{z})$ (see Section 4.4) with $\boldsymbol{\alpha} \in \mathbb{R}^{K^2}$ being fixed. Based on the rules according to which the number of customers in the system evolves (see Section 4.2), it is possible to define a set $\mathcal{R}$ (see Section 4.4) such that

$$\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}\} \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \le (1 - \delta)\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1})$$

for all $\theta < \theta^*/\mu$ and some $\delta < 1$ (see Proposition 3, below), and $\mathrm{E}_{\pi_*}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \in \mathcal{R}\}] < \infty$ for $\theta > 0$ (see Lemma 7, below). The preceding two inequalities are combined to conclude that $\mathrm{E}_{\pi_*}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)] < \infty$ for $\theta < \theta^*/\mu$ (see Theorem 3, below), and

$$\mathrm{E}_{\pi_*}[\exp(\theta\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t)] < \infty$$

follows since $\hat{\boldsymbol{J}}_t$ and $\hat{A}_t$ are normally distributed by definition. Finally, the proof is concluded by showing that

$$\mathrm{E}_{\pi_*}\biggl[\exp\biggl(\theta\biggl|\frac{\hat{Y}_t}{\mu} - \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t\biggr|\biggr)\biggr] < \infty \quad \text{for all } \theta > 0.$$

We remark that the criticality of the exponent $\theta^* = 2\beta/(c_a^2 + c_s^2)$ is consistent with the results obtained earlier in [22] and [25]. Namely, in the GI/M/*n* queue in the QED regime the conditional limited scaled steady-state number of customers is exponentially distributed [22]:

$$x^{-1} \log \mathrm{P}[\hat{Q} > x \mid \hat{Q} > 0] = -\frac{2\beta}{c_a^2 + 1}, \qquad x > 0,$$

while, for the QED GI/D/*n* queue [25], we have, as $x \to \infty$,

$$x^{-1} \log \mathrm{P}[\hat{Q} > x \mid \hat{Q} > 0] \to -\frac{2\beta}{c_a^2};$$

recall that in both cases $\mathrm{P}[\hat{Q} > 0] \in (0, 1)$ for each $\beta > 0$.

Furthermore, we point out that the same exponent $\theta^*$ appears in the Kingman approximation [27], [28] for a single-server queue in the conventional heavy-traffic regime. Moreover, the same exponent was established in analyses of queues with a fixed number of servers in the same heavy-traffic regime. In particular, consider a sequence of *single*-server queues indexed by $n$. The arrival rate to the $n$th system is $\lambda_n \to \infty$, with the arrival process being renewal, satisfying the central limit theorem and $c_{a,n} \to c_a$ as $n \to \infty$. The service times of customers are i.i.d. and equal in distribution to $S/n$ (equivalently, the service capacity grows linearly in $n$), and, thus, the utilization is given by $\rho_n = \lambda_n / n\mu$. Let $\tilde{Q}^n$ be the stationary number of customers *awaiting service* in the system indexed by $n$; $\tilde{Q}^n$ and the total number of customers in the system differ by at most one at any point in time. If $\sqrt{n}(1 - \rho_n) \to \beta > 0$ as $n \to \infty$ then $\tilde{Q}^n/\sqrt{n} \xrightarrow{\mathrm{D}} \tilde{Q}$ as $n \to \infty$, where $\tilde{Q}$ is exponentially distributed [38, Section 9.6] (see also [38, Section 5.7]):

$$x^{-1} \log \mathrm{P}[\tilde{Q} > x] = -\theta^*, \qquad x > 0,$$

where $\theta^*$ is as given in Theorem 2. The agreement of the critical exponent $\theta^*$ in the corresponding single- and $n$-server ($n \to \infty$) systems is interesting since the two evolve under different rules. Observe that, as $n \to \infty$, the total number of customers in the single-server system is $\Theta(\sqrt{n})$ as $n \to \infty$, while, for the $n$-server system, that quantity is $\Theta(n)$ as $n \to \infty$.

In conclusion of this section we state an analogue of Theorem 2 for stationary waiting times using the distributional Little's law [21] applied to the waiting room.

**Corollary 1.** *We have* $\sqrt{n} W^n \xrightarrow{\mathrm{D}} \hat{W} = \hat{Q}/\mu$ *as* $n \to \infty$. *Consequently,* $\mathrm{E}[\mathrm{e}^{\theta \hat{W}}] < \infty$ *if* $\theta < \mu\theta^*$ *and* $\mathrm{E}[\mathrm{e}^{\theta \hat{W}}] = \infty$ *if* $\theta > \mu\theta^*$.

*Proof.* See Section 7.

## 4. Preliminary results

This section contains four subsections. In the first subsection we consider a time-embedded version of $\{(\hat{Q}_t^n, \boldsymbol{L}_t^n),\ t \in \mathbb{R}_+\}$. The number of customers in the finite-$n$ and limiting systems is considered in the second subsection. Quadratic and geometric Lyapunov functions are introduced and analyzed in the last two subsections.

### 4.1. Time-embedded process

In this subsection we examine the triple $(Q_t^n, \boldsymbol{L}_t^n, a_t) \in \mathbb{Z}_+^{K+1} \times \mathbb{R}_+$ and the laws governing its evolution in time. The continuous-time process $\{(Q_t^n, \boldsymbol{L}_t^n, a_t^n),\ t \in \mathbb{R}_+\}$ is not Markovian due to the nonexponential nature of service times. Hence, in order to avoid enlarging the state space, we consider its time-embedded version $\{(Q_t^n, \boldsymbol{L}_t^n, a_t^n),\ t \in \mathbb{Z}_+\}$, i.e. the original process

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

observed at discrete-time instances $t \in \mathbb{Z}_+$ (recall from Section 2.2.1 that $S \in \mathbb{N}$). As seen in the following proposition, the evolution of the latter process is determined by the number of arrivals ($A_t^n$) and customers that enter service ($\boldsymbol{J}_t^n$) during a unit time interval.

**Proposition 1.** *The process* $\{(Q_t^n, \boldsymbol{L}_t^n, a_t^n),\ t \in \mathbb{Z}_+\}$ *is a Markov chain. For every* $t \in \mathbb{Z}_+$, *the value of* $(Q_{t+1}^n, \boldsymbol{L}_{t+1}^n)$ *satisfies*

$$\boldsymbol{L}_{t+1}^n = \mathcal{T}\{\boldsymbol{L}_t^n\} + \boldsymbol{J}_{t+1}^n, \tag{16}$$

$$Q_{t+1}^n = (Q_t^n + A_{t+1}^n + \|\boldsymbol{L}_t^n\| - n - L_{t,1}^n)^+, \tag{17}$$

*where* $\boldsymbol{J}_{t+1}^n \in \mathbb{Z}_+^K$ *is a multinomially distributed random vector that obeys*

$$\|\boldsymbol{J}_{t+1}^n\| = (Q_t^n + A_{t+1}^n) \wedge (n - \|L_{t,k}^n\| + L_{t,1}^n) \tag{18}$$

*and*

$$\mathrm{E}[\boldsymbol{J}_{t+1}^n \mid \|\boldsymbol{J}_{t+1}^n\|] = \|\boldsymbol{J}_{t+1}^n\|\boldsymbol{p}; \tag{19}$$

*given* $\|\boldsymbol{J}_{t+1}^n\|$, *vector* $\boldsymbol{J}_{t+1}^n$ *is conditionally independent of* $\{(Q_t^n, \boldsymbol{L}_t^n, a_t^n),\ t \in \mathbb{Z}_+\}$.

*Proof.* It is sufficient to demonstrate (16), (17), and (18); equality (19) is a straightforward consequence of the i.i.d. nature of service times. The Markov property follows from these relationships and the renewal structure of the arrival process.

Consider the number of customers that enter service in the time interval $(t, t+1]$. At time $t$ there are $\|\boldsymbol{L}_t^n\|$ customers in service, by the definition of $\boldsymbol{L}_t^n$. Out of these $\|\boldsymbol{L}_t^n\|$ customers, $L_{t,1}^n$ depart from the system not later than time $(t+1)$, since their residual service requirements at time $t$ are at most 1 (by the definition of $L_{t,1}^n$). This yields $n - \|\boldsymbol{L}_t^n\| + L_{t,1}^n$ customers that can potentially enter service in the time interval $(t, t+1]$; recall that $n - \|\boldsymbol{L}_t^n\|$ is the number of idle servers at time $t$. On the other hand, the number of customers that can enter service in the time interval $(t, t+1]$ is at most $Q_t^n + A_{t+1}^n$. Thus, the number of customers that do enter service in the time interval $(t, t+1]$ is

$$\|\boldsymbol{J}_{t+1}^n\| = (Q_t^n + A_{t+1}^n) \wedge (n - \|\boldsymbol{L}_t^n\| + L_{t,1}^n), \tag{20}$$

rendering (18). Now, customers in service at time $(t+1)$ with residual service requirements in $(i-1, i]$ are of two types: (i) customers already in service at time $t$ and (ii) customers that enter service in the time interval $(t, t+1]$. Thus, formally

$$L_{t+1,i}^n = \begin{cases} L_{t,i+1}^n + J_{t+1,i}^n, & i = 1, \dots, K-1, \\ J_{t+1,i}^n, & i = K. \end{cases} \tag{21}$$

The multinomial distribution of $\boldsymbol{J}_{t+1}^n$ follows from the assumption that customers' service requirements are i.i.d. RVs, independent from the arrival processes. Rewriting (21) in a vector form renders (16).

In order to establish the value of $Q_{t+1}^n$, it is sufficient to consider the difference between the number of customers that could start receiving service in the time interval $(t, t+1]$ and the actual number of customers that enter service, i.e. (6) and (20) yield

$$\begin{aligned} Q_{t+1}^n &= Q_t^n + A_{t+1}^n - \|\boldsymbol{J}_{t+1}^n\| \\ &= (Q_t^n + A_{t+1}^n + \|\boldsymbol{L}_t^n\| - L_{t,1}^n - n)^+, \end{aligned}$$

and (17) holds. This concludes the proof.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

An analogue of Proposition 1 for scaled processes is stated next.

**Corollary 2.** *The process* $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n, a_t^n),\ t \in \mathbb{Z}_+\}$ *is a Markov chain and it satisfies*

$$\hat{\boldsymbol{L}}_{t+1}^n = \mathcal{T}\{\hat{\boldsymbol{L}}_t^n\} + \hat{\boldsymbol{J}}_{t+1}^n + \hat{J}_{t+1}^n \boldsymbol{p}, \tag{22}$$

$$\hat{Q}_{t+1}^n = \bigg(\hat{Q}_t^n + \hat{A}_{t+1}^n + \sum_{i=2}^K \hat{L}_{t,i}^n - \beta_n\bigg)^+,$$

$$\hat{J}_{t+1}^n = (\hat{Q}_t^n + \hat{A}_{t+1}^n) \wedge \bigg(\beta_n - \sum_{i=2}^K \hat{L}_{t,i}^n\bigg),$$

*where* $\hat{\boldsymbol{J}}_{t+1}^n$*, conditional on* $\hat{J}_{t+1}^n$*, is independent of* $\{(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n, a_t^n),\ t \in \mathbb{Z}_+\}$.

*Proof.* The Markov property follows from Proposition 1 and the fact that there exists a one-to-one mapping between $(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n)$ and $(Q_t^n, \boldsymbol{L}_t^n)$. Now, (16) implies that

$$\begin{aligned}\boldsymbol{L}_{t+1}^n - \lambda_n \tilde{\boldsymbol{p}} &= \mathcal{T}\{\boldsymbol{L}_t^n\} + \boldsymbol{J}_{t+1}^n - \lambda_n \tilde{\boldsymbol{p}} \\ &= (\mathcal{T}\{\boldsymbol{L}_t^n\} - \lambda_n(\tilde{\boldsymbol{p}} - \boldsymbol{p})) + (\boldsymbol{J}_{t+1}^n - \|\boldsymbol{J}_{t+1}^n\| \boldsymbol{p}) + (\|\boldsymbol{J}_{t+1}^n\| - \lambda_n)\boldsymbol{p}.\end{aligned}$$

This equality and the observation that $\mathcal{T}\{\boldsymbol{L}_t^n - \lambda_n \tilde{\boldsymbol{p}}\} = \mathcal{T}\{\boldsymbol{L}_t^n\} - \lambda_n(\tilde{\boldsymbol{p}} - \boldsymbol{p})$ (owing to the definition of $\tilde{\boldsymbol{p}}$) yield (22). The remaining relationships are obtained similarly from their counterparts (17) and (18).

Properties of the vector $\hat{\boldsymbol{J}}_t^n$ are summarized in the following lemma.

**Lemma 1.** *The vector* $\hat{\boldsymbol{J}}_t^n$ *satisfies, for every* $k = 0, 1, \dots, n$ *and* $\theta > 0$,

$$\mathrm{E}[(\boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t^n)^2 \mid \|\boldsymbol{J}_t^n\| = k] = \frac{k\sigma_s^2}{n},$$

$$\mathrm{E}[(\hat{J}_{t,j}^n)^2 \mid \|\boldsymbol{J}_t^n\| = k] = \frac{kp_j(1-p_j)}{n},$$

$$\mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t^n) \mid \|\boldsymbol{J}_t^n\| = k] \le \bigg(\mathrm{E}\bigg[\exp\bigg(\theta \frac{S - 1/\mu}{\sqrt{n}}\bigg)\bigg]\bigg)^n.$$

*Proof.* Let $\{S_i\}_{i=1}^k$ be a sequence of i.i.d. RVs equal in distribution to $S$. Then, the definition of $\hat{\boldsymbol{J}}_t^n$ renders

$$\mathrm{E}[(\boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t^n)^2 \mid \|\boldsymbol{J}_t^n\| = k] = \mathrm{E}\bigg[\bigg(\sum_{i=1}^k \frac{S_i - 1/\mu}{\sqrt{n}}\bigg)^2\bigg] = \frac{k\sigma_s^2}{n}.$$

The other two equalities are obtained in a similar straightforward fashion. The last inequality is due to $\mathrm{E}[\mathrm{e}^{\theta(S-1/\mu)/\sqrt{n}}] \ge 1$. This follows from the convexity of $\mathrm{e}^{\theta(x-1/\mu)/\sqrt{n}}$ in $x$ and Jensen's inequality.

### 4.2. Number in system

This subsection is devoted to a detailed analysis of the rescaled number of customers in the system $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$ and its limiting counterpart. The dynamics of $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$ are related to a newly introduced process $\{\hat{\boldsymbol{Z}}_t^n,\ t \in \mathbb{Z}_+\}$,

$$\hat{\boldsymbol{Z}}_t^n = (\hat{Z}_{t,1}^n, \dots, \hat{Z}_{t,K}^n) := \hat{\boldsymbol{J}}_t^n + \boldsymbol{p}\hat{A}_t^n, \tag{23}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

and in particular to

$$\hat{V}_t^n := \sum_{i=1}^{K}\sum_{j=i}^{K}(\hat{J}^n_{t+1-i,j} + p_j\hat{A}^n_{t+1-i}) \tag{24}$$

$$= \sum_{i=1}^{K}\sum_{j=i}^{K}\hat{Z}^n_{t+1-i,j}, \tag{25}$$

as stated in the next lemma. Informally, for large $n$, the process $\{\hat{V}_t^n,\ t \in \mathbb{Z}_+\}$ serves as a proxy for a scaled infinite-server process. We remark that the lemma is a discrete-time analogue of Equation (1.1) of [36].

**Lemma 2.** *The process* $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$ *satisfies, for all* $t \geq K$,

$$\hat{Y}_t^n = \hat{V}_t^n + \sum_{i=1}^{K} p_i(\hat{Y}^n_{t-i} - \beta_n)^+.$$

*Proof.* Equalities (11) and (22) yield the following expression for the $K$th element of the vector $\hat{\boldsymbol{L}}^n_{t+1}$:

$$\hat{L}^n_{t+1,K} = \hat{J}^n_{t+1,K} + (\hat{A}^n_{t+1} + \hat{Q}^n_t - \hat{Q}^n_{t+1})p_K.$$

Furthermore, using (22) iteratively, it is straightforward to obtain the remaining elements of $\hat{\boldsymbol{L}}^n_{t+1}$. To this end, for $j = 0, \dots, K-1$,

$$\hat{L}^n_{t+1+j,K-j} = \sum_{i=0}^{j}(\hat{J}^n_{t+1+i,K-i} + (\hat{A}^n_{t+1+i} + \hat{Q}^n_{t+i} - \hat{Q}^n_{t+1+i})p_{K-i}),$$

which after a change of time indices renders, for $t \geq K$ and $j = 1, \dots, K$,

$$\hat{L}^n_{t,j} = \sum_{i=1}^{K+1-j}(\hat{J}^n_{t+1-i,j+i-1} + (\hat{A}^n_{t+1-i} + \hat{Q}^n_{t-i} - \hat{Q}^n_{t-i+1})p_{j+i-1}). \tag{26}$$

Summing both sides of (26) over $j = 1, \dots, K$ and using (24) results in

$$\sum_{j=1}^{K}\hat{L}^n_{t,j} = \sum_{i=1}^{K}\biggl(\sum_{j=i}^{K}\hat{J}^n_{t+1-i,j} + (\hat{A}^n_{t+1-i} + \hat{Q}^n_{t-i} - \hat{Q}^n_{t-i+1})\tilde{p}_i\biggr)$$

$$= \hat{V}_t^n - \hat{Q}_t^n + \sum_{i=1}^{K} p_i\hat{Q}^n_{t-i}.$$

The statement of the lemma follows from the preceding equality, (9), and (10).

The following corollary establishes a lower bound and upper bound on the value of $\hat{Y}_t^n$ in terms of the past values of $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$ and $\{\hat{V}_t^n,\ t \in \mathbb{Z}_+\}$.

**Corollary 3.** (i) *For every* $k \in \mathbb{Z}_+$ *and* $t \geq k + K$,

$$\hat{Y}_t^n \leq \sum_{i=0}^{k}(\hat{V}^n_{t-i})^+ + \sum_{i=k+1}^{k+K}\tilde{p}_{i-k}(\hat{Y}^n_{t-i} - \beta_n)^+.$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

(ii) *For* $1 \le i_1, \dots, i_k \le K$*, let* $p_{(k)} = \prod_{j=1}^{k} p_{i_j}$ *and* $s_{(k)} = \sum_{j=1}^{k} i_j$ *with* $s_{(0)} = 0$*. Then, for* $t \ge s_{(k)}$*,*

$$\hat{Y}_t^n \ge p_{(k)} \hat{Y}_{t-s_{(k)}}^n - \sum_{j=0}^{k-1} (\beta - \hat{V}_{t-s_{(j)}}^n)^+.$$

*Proof.* The proofs of the two parts follow by induction on $k$.

(i) The base of the induction ($k = 0$) is due to Lemma 2 and $\tilde{p}_i \ge p_i$ for $i = 1, \dots, K$. Then the bound follows from the inductive assumption, Lemma 2, and $\tilde{p}_i = p_i + \tilde{p}_{i+1} \le 1$:

$$\begin{aligned}\hat{Y}_t^n &\le \sum_{i=0}^{k} (\hat{V}_{t-i}^n)^+ + (\hat{Y}_{t-k-1}^n)^+ + \sum_{i=k+2}^{k+K} \tilde{p}_{i-k} (\hat{Y}_{t-i}^n - \beta_n)^+ \\ &\le \sum_{i=0}^{k+1} (\hat{V}_{t-i}^n)^+ + \sum_{i=k+2}^{k+1+K} \tilde{p}_{i-k-1} (\hat{Y}_{t-i}^n - \beta_n)^+.\end{aligned}$$

(ii) By Lemma 2 we have $\hat{Y}_t^n \ge \hat{V}_t^n + p_{i_1} (\hat{Y}_{t-i_1}^n - \beta)^+$, which implies that

$$\begin{aligned}\hat{Y}_t^n &\ge p_{i_1} \hat{Y}_{t-i_1}^n - (\beta - \hat{V}_t^n) \\ &\ge p_{i_1} \hat{Y}_{t-i_1}^n - (\beta - \hat{V}_t^n)^+. \end{aligned} \tag{27}$$

The preceding inequality provides the base of the induction ($k = 1$). Now suppose that the statement of the corollary holds for some $k \ge 1$. Then combining (27), the inductive assumption, and $p_i \le 1$ yields

$$\hat{Y}_t^n \ge p_{(k)} \hat{Y}_{t-s_{(k)}}^n - \sum_{j=0}^{k-1} (\beta - \hat{V}_{t-s_{(j)}})^+ \ge p_{(k+1)} \hat{Y}_{t-s_{(k+1)}}^n - \sum_{j=0}^{k} (\beta - \hat{V}_{t-s_{(j)}})^+,$$

where the second inequality is also due to $s_{(k+1)} = s_{(k)} + i_{k+1}$.

In the rest of this subsection we state the limiting counterparts of the results derived for $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$. We start by introducing the limiting analogs of $\hat{\boldsymbol{Z}}_t^n$ and $\hat{V}_t^n$. Define

$$\hat{\boldsymbol{Z}}_t := \hat{\boldsymbol{J}}_t + \boldsymbol{p} \hat{A}_t$$

and

$$\hat{V}_t := \sum_{i=1}^{K} \sum_{j=i}^{K} (\hat{J}_{t+1-i,j} + p_j \hat{A}_{t+1-i}) = \sum_{i=1}^{K} \sum_{j=i}^{K} \hat{Z}_{t+1-i,j}, \tag{28}$$

where the $\hat{Z}_{t,i}$s are the elements of $\hat{\boldsymbol{Z}}_t$. Since $\hat{\boldsymbol{J}}_t$ and $\hat{A}_t$ are normal RVs by definition, $\hat{\boldsymbol{Z}}_t$ is normally distributed as well, and, for all $t$ and $i$, we have

$$|\hat{Z}_{t,i}| \in \mathcal{M}_\infty. \tag{29}$$

The properties of $\{\hat{Y}_t,\ t \in \mathbb{Z}_+\}$ are summarized in the following lemma, including limiting counterparts of Lemma 2 and Corollary 3.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

**Lemma 3.** (i) *The process* $\{\hat{Y}_t,\ t \in \mathbb{Z}_+\}$ *satisfies, for all* $t \geq K$,

$$\hat{Y}_t = \hat{V}_t + \sum_{i=1}^{K} p_i (\hat{Y}_{t-i} - \beta)^+.$$

(ii) *For every* $t \in \mathbb{Z}_+$,

$$(-\hat{Y}_t) \in \mathcal{M}_\infty.$$

(iii) *For every* $k \in \mathbb{Z}_+$ *and* $t \geq k + K$,

$$\hat{Y}_t \leq \sum_{i=0}^{k} (\hat{V}_{t-i})^+ + \sum_{i=k+1}^{k+K} \tilde{p}_{i-k} (\hat{Y}_{t-i} - \beta)^+.$$

(iv) *For* $1 \leq i_1, \dots, i_k \leq K$, *let* $p_{(k)} = \prod_{j=1}^{k} p_{i_j}$ *and* $s_{(k)} = \sum_{j=1}^{k} i_j$ *with* $s_{(0)} = 0$. *Then, for* $t \geq s_{(k)}$,

$$\hat{Y}_t \geq p_{(k)} \hat{Y}_{t-s_{(k)}} - \sum_{j=0}^{k-1} (\beta - \hat{V}_{t-s_{(j)}})^+.$$

*Proof.* (i) The proof is analogous to the proof of Lemma 2. Part (ii) follows from $\hat{Y}_t \geq \hat{V}_t$, part (i), and the fact that $\hat{A}_t$ and $\hat{\boldsymbol{J}}_t$ have normal distributions. The proofs of parts (iii) and (iv) are analogous to the proof of Corollary 3.

### 4.3. Quadratic Lyapunov function

Here we introduce a quadratic Lyapunov function and prove some of its properties. To this end, we define $K$ vectors $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K$, where elements of the vector $\boldsymbol{\alpha}_k = (\alpha_{k,1}, \dots, \alpha_{k,K})$ are defined by

$$\alpha_{k,j} = (j - k)^+; \tag{30}$$

let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_K) \in \mathbb{R}^{K^2}$. A function $\Psi_\theta(\boldsymbol{y}, \boldsymbol{z})\colon \mathbb{R}^{K+K^2} \to \mathbb{R}$ is defined by

$$\Psi_\theta(\boldsymbol{y}, \boldsymbol{z}) := (\tilde{\boldsymbol{p}} \cdot \boldsymbol{y} + \boldsymbol{\alpha} \cdot \boldsymbol{z})^\theta \tag{31}$$

and a set $\mathcal{R}_x$ by

$$\mathcal{R}_x := \{\boldsymbol{y} \in \mathbb{R}^K : y_i < x \text{ for some } i\}. \tag{32}$$

The case in which $\theta = 2$ is of particular importance since it corresponds to a quadratic Lyapunov function (see Appendix B for the definition) as established below. Finally, we introduce $\bar{\boldsymbol{Y}}_t^n := (\hat{Y}_t^n, \dots, \hat{Y}_{t-K+1}^n)$ and $\bar{\boldsymbol{Z}}_t^n := (\hat{\boldsymbol{Z}}_t^n, \dots, \hat{\boldsymbol{Z}}_{t-K+1}^n)$; the 'bar' notation in $\bar{\boldsymbol{Y}}_t^n$ and $\bar{\boldsymbol{Z}}_t^n$ indicates that elements of these vectors refer to different time indices.

**Proposition 2.** *There exist* $\delta > 0$, $\psi < \infty$, *and* $n_0$ *such that, for all* $n \geq n_0$,

$$\mathrm{E}[\Psi_2(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)\, \mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1}^n \notin \mathcal{R}_{\beta_n}\} - \Psi_2(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n) \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n] \leq -\delta \Psi_1(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n) + \psi.$$

*Proof.* On the event $\{\bar{\boldsymbol{Y}}_{t-1}^n \notin \mathcal{R}_{\beta_n}\}$, Lemma 2 renders in a vector form $\hat{Y}_t^n = \hat{V}_t^n - \beta_n + \boldsymbol{p} \cdot \bar{\boldsymbol{Y}}_{t-1}^n$, and, since $p_i + \tilde{p}_{i+1} = \tilde{p}_i$ by definition, it implies that $\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n = \hat{V}_t^n - \beta_n + \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_{t-1}^n$. Thus, the linear combination of $\bar{\boldsymbol{Y}}_t^n$ and $\bar{\boldsymbol{Z}}_t^n$ that appears in the definition of $\Psi_\theta$ can be expressed as

$$\begin{aligned}
\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t^n &= \hat{V}_t^n - \beta_n + \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_{t-1}^n + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t^n \\
&= \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_{t-1}^n + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_{t-1}^n - \beta_n + \boldsymbol{K} \cdot \hat{\boldsymbol{Z}}_t^n, \qquad (33)
\end{aligned}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

where the second equality follows from (25) and (30). Then, based on (33), we obtain

$$\begin{aligned}\mathrm{E}[\Psi_2(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1}^n \notin \mathcal{R}_{\beta_n}\} \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n] - \Psi_2(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n)\\ \le 2\Psi_1(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n)\,\mathrm{E}[-\beta_n + \boldsymbol{K}\cdot\hat{\boldsymbol{Z}}_t^n \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n]\\ + \mathrm{E}[(-\beta_n + \boldsymbol{K}\cdot\hat{\boldsymbol{Z}}_t^n)^2 \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n]. \end{aligned} \tag{34}$$

Now, by (23), the sum in (34) can be expressed in terms of $\hat{A}_t^n$ and $\hat{\boldsymbol{J}}_t^n$ as follows:

$$\begin{aligned}\mathrm{E}[\boldsymbol{K}\cdot\hat{\boldsymbol{Z}}_t^n \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n] &= \mathrm{E}\bigg[\frac{\hat{A}_t^n}{\mu} + \boldsymbol{K}\cdot\hat{\boldsymbol{J}}_t^n \,\bigg|\, \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n\bigg]\\ &= \mathrm{E}\bigg[\frac{\hat{A}_t^n}{\mu} \,\bigg|\, \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n\bigg]\\ &\le \sup_{a\ge 0}\mathrm{E}\bigg[\frac{\hat{A}_t^n}{\mu} \,\bigg|\, a_{t-1}^n = a\bigg], \end{aligned} \tag{35}$$

where the last inequality is due to the fact that $\hat{A}_t^n$ is conditionally independent of $\hat{\boldsymbol{J}}_t^n$ and $(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n)$ given $a_{t-1}$ (the arrival process is renewal). The second expectation on the right-hand side of (34) can be upper bounded by utilizing the same fact in addition to the observation that $\hat{\boldsymbol{J}}_t^n$ is conditionally independent of $\hat{A}_t^n$ and $(\bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n)$ given $\hat{J}_t^n$ (see Corollary 2). These two facts yield

$$\begin{aligned}&\mathrm{E}[(-\beta_n + \boldsymbol{K}\cdot\hat{\boldsymbol{Z}}_t^n)^2 \mid \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n]\\ &\quad= \mathrm{E}\bigg[\bigg(\frac{\hat{A}_t^n}{\mu} - \beta_n\bigg)^2 + (\boldsymbol{K}\cdot\hat{\boldsymbol{J}}_t^n)^2 \,\bigg|\, \bar{\boldsymbol{Y}}_{t-1}^n, \bar{\boldsymbol{Z}}_{t-1}^n\bigg]\\ &\quad\le \sup_{a\ge 0}\mathrm{E}\bigg[\bigg(\frac{\hat{A}_t^n}{\mu} - \beta_n\bigg)^2 \,\bigg|\, a_{t-1}^n = a\bigg] + \max_{0\le i\le n}\mathrm{E}[(\boldsymbol{K}\cdot\hat{\boldsymbol{J}}_t^n)^2 \mid \|\boldsymbol{J}_t^n\| = i]\\ &\quad\le \sup_{a\ge 0}\mathrm{E}\bigg[\bigg(\frac{\hat{A}_t^n}{\mu} - \beta_n\bigg)^2 \,\bigg|\, a_{t-1}^n = a\bigg] + \sigma_s^2, \end{aligned} \tag{36}$$

where the last inequality follows from Lemma 1 and $\|\boldsymbol{J}_t^n\| \le n$. The limit (as $n \to \infty$) of the right-hand side of the preceding inequality remains bounded owing to assumption (3) on the arrival process (see Section 2.1.2) and the fact that service times are bounded ($S \le K$).

Combining (34) with (2), (35), and (36) yields the statement of the proposition.

**Lemma 4.** *The following inequality holds:*

$$\limsup_{n\to\infty}\mathrm{E}_{\pi_n}[\Psi_2(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1}^n \in \mathcal{R}_{\beta_n}\}] < \infty.$$

In the proof of Lemma 4 the following number-theoretic fact will be utilized. For completeness, we provide its proof.

**Lemma 5.** *Let $p$ and $q$ be two relatively prime numbers. For any $K \in \mathbb{N}$, there exists $k \in \mathbb{N}$ such that any $l \in \{k+1, \dots, k+K\}$ can be represented as $l = i_l p + j_l q$ for some $i_l, j_l \in \mathbb{N}$.*

*Proof.* Since $p$ and $q$ are relatively prime, then any $m \in \{1, 2, \dots, K\}$ can be represented as $m = i'_n p + j'_n q$ for some possibly negative integers $i'_n$ and $j'_n$; see, e.g. [29, p. 104]. Let $t = \max_m\{i'_n, j'_n\} + 1$ and $k = tp + tq$. Then every $l \in \{k+1, \dots, k+K\}$ is given by $l = i_l p + j_l q$, where $i_l = (t + i'_{l-k})$ and $j_l = (t + j'_{l-k})$.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

*Proof of Lemma 4.* Let $\mathcal{R}_x^k = \{\boldsymbol{y} \in \mathbb{R}^K : y_1 \geq x, \dots, y_{k-1} \geq x,\ y_k < x\}$, $1 \leq k \leq K$. It is sufficient to prove the statement of the lemma with $\mathcal{R}_{\beta_n}$ replaced with $\mathcal{R}_{\beta_n}^k$ for an arbitrary $k \in \{1, \dots, K\}$ since $\mathcal{R}_{\beta_n} = \bigcup_k \mathcal{R}_{\beta_n}^k$. The proof is based on demonstrating the following bound for some positive integer $m$ and constants $\{c_i,\ i = 0, \dots, m+k+K\}$ and $\{d_i,\ i = 0, \dots, m+k+K\}$:

$$\Psi_2(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)\, \mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1}^n \in \mathcal{R}_{\beta_n}^k\} \leq \left(\sum_{i=0}^{m+k+K} (c_i + d_i |\hat{V}_{t-i}^n|)\right)^2 \quad \text{for all } n, . \tag{37}$$

Then the statement of the lemma follows from the definition of $\hat{V}_t^n$, Lemma 1, and (3) applied in the unconditioned case; thus, we focus on demonstrating (37).

On the event $\{\bar{\boldsymbol{Y}}_{t-1}^n \in \mathcal{R}_{\beta_n}^k\}$, applying Lemma 2 to $\hat{Y}_t^n$ yields

$$\begin{aligned} \hat{Y}_t^n &= \hat{V}_t^n + \sum_{i=1}^{k-1} p_i(\hat{Y}_{t-i}^n - \beta_n) + \sum_{i=k+1}^{K} p_i(\hat{Y}_{t-i}^n - \beta_n)^+ \\ &= \hat{V}_t^n + \sum_{i=1}^{k-1} g_i(\hat{V}_{t-i}^n - \beta_n) + \sum_{i=k+1}^{K+k-1} h_i(\hat{Y}_{t-i}^n - \beta_n)^+, \end{aligned} \tag{38}$$

where the constants $g_i$ and $h_i$ can be computed in a recursive fashion; i.e. $g_0 = 1$, $g_i = \sum_{j=0}^{i-1} g_j p_{i-j}$ for $i = 1, \dots, k-1$ and $h_i = \sum_{j=(i-K)^+}^{k-1} g_j p_{i-j}$ for $i = k+1, \dots, K+k-1$. Hence, based on (38), there exist finite $g$ and $h$ such that

$$\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n \leq g \sum_{i=0}^{k-1} (\hat{V}_{t-i}^n)^+ + h \sum_{i=k+1}^{K+k-1} (\hat{Y}_{t-i}^n)^+. \tag{39}$$

Next, on the event of interest, $\{\bar{\boldsymbol{Y}}_{t-1}^n \in \mathcal{R}_{\beta_n}^k\}$, we upper bound the second sum in (39) in two steps: (i) bound values of $\{\hat{Y}_t^n,\ t \in \mathbb{Z}_+\}$ on a time interval of length $K$ prior to time $(t-k)$ based on $\{\hat{Y}_{t-k}^n < \beta_n\}$ and (ii) obtain a desired bound based on (i). First, consider arbitrary $i_1, i_2 \leq K$ such that $p_{i_1} p_{i_2} > 0$; such a pair of indices exists since $\sigma_s > 0$ (see Section 2.1.1). By Lemma 5, there exists a sufficiently large $m$ such that every element of $\{m+1, m+2, \dots, m+K\}$ can be represented as $r_1 i_1 + r_2 i_2$ for some nonnegative integers $r_1$ and $r_2$. Invoking the second part of Corollary 3 and $\{\hat{Y}_{t-k}^n < \beta_n\}$ yields the existence of finite $r$, $q$, and $m \geq K$ such that

$$(\hat{Y}_{t-k-i}^n)^+ \leq r + q \sum_{j=k}^{m+k+K} |\hat{V}_{t-j}^n|$$

for all $i \in \{m+1, \dots, m+K\}$; we also used $|x+y| \leq |x| + |y|$ and the fact that the elements of the sum are nonnegative. The preceding inequality and the first part of Corollary 3 assure the existence of finite $r'$ and $q'$ such that

$$h \sum_{i=k+1}^{K+k-1} (\hat{Y}_{t-i}^n)^+ \leq r' + q' \sum_{i=k}^{m+k+K} |\hat{V}_{t-i}^n|, \tag{40}$$

since each summand on the left-hand side is upper bounded by an expression that appears on the right-hand side with $r'$ and $q'$ replaced by some other finite constants.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

Next, combining (39) and (40) provides the following bound on $\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n$ in terms of $\hat{V}_t^n, \dots, \hat{V}_{t-m-k-K}^n$:

$$\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n \le g \sum_{i=0}^{k-1} (\hat{V}_{t-i}^n)^+ + r' + q' \sum_{i=k}^{m+k+K} |\hat{V}_{t-i}^n| \le r' + g' \sum_{i=0}^{m+k+K} |\hat{V}_{t-i}^n|,$$

where $g'$ is finite. Finally, from (25) we know that the absolute value of $\boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t^n$ is upper bounded by a linear combination of $|\hat{V}_{t-i}^n|$s. Then, (37) follows from the preceding bound. This completes the proof.

**Lemma 6.** *The following inequality holds:*

$$\limsup_{n \to \infty} \mathrm{E}_{\pi_n}[-\Psi_1(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)\, \mathbf{1}\{\Psi_1(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n) < 0\}] < \infty.$$

*Proof.* Lemma 2 renders $\hat{Y}_t^n \ge \hat{V}_t^n$, which leads to $\Psi_1(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n) \ge \sum_{i=1}^K \tilde{p}_i \hat{V}_{t+1-i}^n + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t^n$. The statement follows from the preceding relationship, (3), (24), the Cauchy–Schwarz inequality, and Lemma 1.

### 4.4. Geometric Lyapunov function

In this subsection we introduce a family of Lyapunov functions parameterized by $\theta > 0$ and prove some of its properties. Given a parameter $\theta > 0$, consider a function

$$\Phi_\theta(\boldsymbol{y}, \boldsymbol{z}) \colon \mathbb{R}^{K+K^2} \to \mathbb{R}_+$$

defined by

$$\Phi_\theta(\boldsymbol{y}, \boldsymbol{z}) := \exp(\theta \tilde{\boldsymbol{p}} \cdot \boldsymbol{y} + \theta \boldsymbol{\alpha} \cdot \boldsymbol{z}). \tag{41}$$

We consider $\Phi_\theta$ as a function of the limiting pair $(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)$, where $\bar{\boldsymbol{Y}}_t := (\hat{Y}_t, \dots, \hat{Y}_{t-K+1})$ and $\bar{\boldsymbol{Z}}_t := (\hat{\boldsymbol{Z}}_t, \dots, \hat{\boldsymbol{Z}}_{t-K+1})$. Proposition 3, below, establishes a negative drift of the Lyapunov function $\Phi_\theta$ under an assumption that $\theta < \theta^*/\mu$. Moreover, $\theta^*/\mu$ is the critical exponent under which $\Phi_\theta$ is a geometric Lyapunov function (see Appendix B for the definition). Recall the definition of $\mathcal{R}_x$ from (32).

**Proposition 3.** *For every $\theta < \theta^*/\mu$, there exists $\delta = \delta_\theta > 0$ such that*

$$\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\, \mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta\} \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \le (1-\delta)\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}), \tag{42}$$

*and, for every $\theta > \theta^*/\mu$, there exists $\delta = \delta_\theta > 0$ such that*

$$\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t) \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \ge (1+\delta)\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}). \tag{43}$$

*Proof.* The proof is similar to that of Proposition 2. From Lemma 3(i) we have, on the event $\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta\}$,

$$\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t = \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_{t-1} + \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_{t-1} - \beta + \boldsymbol{K} \cdot \hat{\boldsymbol{Z}}_t.$$

This results in

$$\begin{aligned}
&\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\, \mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta\} \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \\
&\quad = \mathrm{e}^{-\theta\beta}\, \mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{Z}}_t)\, \mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta\} \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}]\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}) \\
&\quad \le \mathrm{e}^{-\theta\beta}\, \mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{Z}}_t) \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}]\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}) \\
&\quad = \mathrm{e}^{-\theta\beta}\, \mathrm{E}\biggl[\exp\biggl(\frac{\theta \hat{A}_t}{\mu}\biggr)\biggr] \mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t)]\Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}),
\end{aligned} \tag{44}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

where the last equality follows from the definition of $\hat{\boldsymbol{Z}}_t$, and the mutual independence of $\hat{A}_t$ and $\hat{\boldsymbol{J}}_t$ as well as their independence of $(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1})$. By definition, the RV $\hat{A}_t$ is normally distributed with zero mean and variance $\mu c_a^2$ and, hence,

$$\mathrm{E}\biggl[\exp\biggl(\frac{\theta \hat{A}_t}{\mu}\biggr)\biggr] = \exp\biggl(\frac{\theta^2 c_a^2}{2\mu}\biggr). \tag{45}$$

On the other hand, $\hat{\boldsymbol{J}}_t$ is normal with covariance matrix $\mu\boldsymbol{\Sigma} = \mu(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}^\top \boldsymbol{p})$ (see (12)), where $\mathrm{diag}(\boldsymbol{p})$ is the diagonal matrix defined by $\boldsymbol{p}$. Thus,

$$\mathrm{E}[(\boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t)^2] = \mu \boldsymbol{K}^\top (\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}^\top \boldsymbol{p})\boldsymbol{K} = \mu \sum_{j=1}^{K} j^2 p_j - \mu(\boldsymbol{K} \cdot \boldsymbol{p})^2 = \mu\sigma_s^2,$$

which in turn implies that $\mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{J}}_t)] = \exp(\theta^2 c_s^2 / 2\mu)$. This equality, (44), and (45) result in

$$\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta\} \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \le \mathrm{e}^{-\theta\beta} \exp\biggl(\frac{\theta^2 c_a^2}{2\mu}\biggr) \exp\biggl(\frac{\theta^2 c_s^2}{2\mu}\biggr) \Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}),$$

and then (42) follows provided that $-\theta\beta + \theta^2 c_a^2/2\mu + \theta^2 c_s^2/2\mu < 0$, or, equivalently, $\theta < \theta^*/\mu$. Therefore, the first part of the proposition is established.

The proof of (43) is very similar. We observe from Lemma 3(i) that $\hat{Y}_t \ge \hat{V}_t - \beta + \boldsymbol{p} \cdot \bar{\boldsymbol{Y}}_{t-1}$, regardless of whether $\bar{\boldsymbol{Y}}_{t-1} \in \mathcal{R}_\beta$ or $\bar{\boldsymbol{Y}}_{t-1} \notin \mathcal{R}_\beta$. Repeating the analysis for the previous case ($\theta < \theta^*/\mu$), we obtain

$$\begin{aligned}\mathrm{E}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t) \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] &\ge \mathrm{e}^{-\theta\beta}\, \mathrm{E}[\exp(\theta \boldsymbol{K} \cdot \hat{\boldsymbol{Z}}_t) \mid \bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}] \Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1}) \\ &= \exp\biggl(-\theta\beta + \frac{\theta^2 c_a^2}{2\mu} + \frac{\theta^2 c_s^2}{2\mu}\biggr) \Phi_\theta(\bar{\boldsymbol{Y}}_{t-1}, \bar{\boldsymbol{Z}}_{t-1});\end{aligned}$$

thus, (43) holds provided that $\theta > \theta^*/\mu$. This concludes the proof of the proposition.

The analogue of Lemma 4 for the geometric function applied to the limiting processes is stated next. The proof is very similar to that of Lemma 4, except that the fact that $\hat{A}_t$ and $\hat{\boldsymbol{J}}_t$ are normally distributed is utilized.

**Lemma 7.** *We have* $\mathrm{E}_{\pi_*}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \in \mathcal{R}_\beta\}] < \infty$ *for every* $\theta > 0$.

*Proof.* As in the proof of Lemma 4, it is sufficient to prove the statement of the lemma with $\mathcal{R}_\beta$ replaced with $\mathcal{R}_{\beta_n}^k$ for an arbitrary $k \in \{1, \dots, K\}$. Repeating the steps of the proof of Lemma 4 yields the existence of some positive integer $m$ and constants $c_i$ and $d_i$ such that

$$\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)\,\mathbf{1}\{\bar{\boldsymbol{Y}}_{t-1} \in \mathcal{R}_\beta^k\} \le \exp\biggl(\sum_{i=0}^{m+k+K} (c_i + d_i |\hat{V}_{t-i}|)\biggr).$$

Then the statement of the lemma follows from the definition of $\hat{V}_t$, the Gaussian distribution of its components, and Proposition 6 given in Appendix A.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

## 5. Proof of Theorem 1

The convergence in the statement of the theorem is established by proving the tightness of all relevant RVs. Recall that a sequence of RVs $\{X_n,\ n \geq 1\}$ is tight [14, p. 87] if, for all $\varepsilon > 0$, there exists an $x_\varepsilon$ such that

$$\sup_{n\to\infty} \mathrm{P}[X_n \notin (-x_\varepsilon, x_\varepsilon]] \leq \varepsilon.$$

**Proposition 4.** *For a fixed* $t \geq 0$*, the sequence* $\{\hat{Y}_t^n,\ n \geq 1\}$ *is tight with respect to the sequence of probability measures* $\{\pi_n,\ n \geq 1\}$.

*Proof.* Theorem 4, given in Appendix B, can be used to bound the sequence $\{\hat{Y}_t^n,\ n \geq 1\}$ away from $+\infty$. In order to obtain uniform boundedness away from $-\infty$, we utilize the fact that the negative part of $\hat{Y}_t^n$ can be upper bounded by $(\hat{V}_t^n)^-$ according to Lemma 3(i).

From Proposition 2, Lemma 4, Lemma 6, and Theorem 4, it follows that

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Psi_1(\bar{\boldsymbol{Y}}_t^n, \bar{\boldsymbol{Z}}_t^n)] < \infty.$$

Applying (3), (25), and Lemma 1, from (31) in the case in which $\theta = 1$, we obtain

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n] < \infty. \tag{46}$$

Next, Lemma 2 implies that $\hat{Y}_t^n \geq \hat{V}_t^n$, leading to $(\hat{Y}_t^n)^- \leq |\hat{V}_t^n| = \sqrt{(\hat{V}_t^n)^2}$, which, combined with (3), (24), and Lemma 1, yields

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[(\hat{Y}_t^n)^-] < \infty. \tag{47}$$

Now, in view of $\tilde{p}_1 = 1$ we have $\hat{Y}_t^n = \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n - \sum_{k=2}^K \tilde{p}_k \hat{Y}_{t-k}^n \leq \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t^n + \sum_{k=2}^K \tilde{p}_k (\hat{Y}_{t-k}^n)^-$, and it then follows, from (46) and (47), that

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\hat{Y}_t^n] < \infty.$$

This bound together with (47) and the Markov inequality implies the tightness of the sequence $\{\hat{Y}_t^n,\ n \geq 1\}$ with respect to the sequence of distributions $\{\pi_n,\ n \geq 1\}$.

For the purposes of the proof of Theorem 1, it is convenient to define the following sequence of *stationary* random processes:

$$\{\hat{\boldsymbol{\Upsilon}}_t^n = (\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n, \hat{A}_t^n, \hat{\boldsymbol{J}}_t^n, \hat{J}_t^n, \hat{Y}_t^n, a_t^n),\ t \in \mathbb{R}_+\}$$

indexed by $n$. Assume that $(\hat{Q}_t^n, \hat{\boldsymbol{L}}_t^n, a_t^n)$ (or, equivalently, the extended process $\hat{\boldsymbol{\Upsilon}}_t^n$) is distributed according to $\pi_n$ for all $t \in \mathbb{R}_+$ (see Section 3).

**Corollary 4.** *For a fixed* $t \geq 0$*, the sequence* $\{\hat{\boldsymbol{\Upsilon}}_t^n,\ n \geq 1\}$ *is tight with respect to the sequence of probability measures* $\{\pi_n,\ n \geq 1\}$.

*Proof.* The tightness of the RVs $\{\hat{A}_t^n,\ n \geq 1\}$ follows from (3) and the tightness of $\{\hat{Y}_t^n,\ n \geq 1\}$ is due to Proposition 4. The tightness of $\{\hat{Q}_t^n,\ n \geq 1\}$ then follows from (10), and, thus, (11) implies the tightness of $\{\hat{J}_t^n, n \geq 1\}$. The tightness of $\{\hat{J}_t^n,\ n \geq 1\}$ implies, via (8), that $\|\boldsymbol{J}_t^n\|/\mu n \to 1$ with probability 1. Recalling that $\hat{\boldsymbol{J}}_t^n$ conditional on $\hat{J}_t^n$ is independent of all the other RVs (see Corollary 2), we obtain the tightness of the sequence $\{\hat{\boldsymbol{J}}_t^n,\ n \geq 1\}$. Finally,

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

iteratively applying (22), we obtain the tightness of $\hat{L}^n_{t,K}$, $\hat{L}^n_{t,K-1}$, and $\hat{L}^n_{t,1}$. The tightness of $a^n_t$ follows from the equilibrium assumption of the arrival processes, which implies that $\mathrm{E}[a^n_t] = (c^2_{a,n} + 1)/2\lambda_n = O(1/n)$ as $n \to \infty$, owing to the assumption that $c_{a,n} \to c_a < \infty$ as $n \to \infty$. This completes the proof of the corollary.

The preceding result implies the weak convergence of $\pi_n$ along some subsequence $\{n_k,\ k \geq 1\}$ to some limiting probability measure $\pi_*$ [10, p. 59]. For now, let $\pi_*$ be any such limiting measure. Later in this section we establish the uniqueness of $\pi_*$. Observe that the tightness of $\{\boldsymbol{\Upsilon}^n_t,\ n \geq 1\}$ implies the tightness of $\{(\boldsymbol{\Upsilon}^n_t, \boldsymbol{\Upsilon}^n_{t+1}),\ n \geq 1\}$.

**Proposition 5.** *Let $\{\boldsymbol{\Upsilon}^n_t,\ t \in \mathbb{Z}_+\}$ be in stationarity, and suppose that*

$$(\boldsymbol{\Upsilon}^n_t, \boldsymbol{\Upsilon}^n_{t+1}) \xrightarrow{\mathrm{D}} (\check{\boldsymbol{\Upsilon}}_t, \check{\boldsymbol{\Upsilon}}_{t+1}) \quad \text{as } n \to \infty$$

*for some $(\check{\boldsymbol{\Upsilon}}_t, \check{\boldsymbol{\Upsilon}}_{t+1})$, where $\check{\boldsymbol{\Upsilon}}_t = (\check{Q}_t, \check{\boldsymbol{L}}_t, \check{A}_t, \check{\boldsymbol{J}}_t, \check{J}_t, \check{Y}_t, \check{a}_t)$. Then the RVs $\check{A}_{t+1}$ and $\check{\boldsymbol{\Upsilon}}_t$ are independent.*

*Proof.* By (1) and (7), it follows that $\check{A}_{t+1}$ is equal in distribution to $\hat{A}_{t+1}$. We need to show that, for every real $a$ and $\boldsymbol{b}$,

$$\mathrm{P}[\check{A}_{t+1} \leq a,\ \check{\boldsymbol{\Upsilon}}_t \leq \boldsymbol{b}] = \mathrm{P}[\check{A}_{t+1} \leq a]\,\mathrm{P}[\check{\boldsymbol{\Upsilon}}_t \leq \boldsymbol{b}], \tag{48}$$

where, for the vector case, '$\leq$' is interpreted coordinatewise. Given a multidimensional RV $\boldsymbol{X}$, recall that a vector $\boldsymbol{x}$ is defined to be a continuity point if $\mathrm{P}[X_i = x_i] = 0$ for every coordinate $i$; it is known that the set of continuity points is a dense uncountable set (see [14, Section 2.9]). Since distribution functions are right continuous, it suffices to establish the identity (48) for the case when $a$ and $\boldsymbol{b}$ are continuity points of $\check{A}_{t+1}$ and $\check{\boldsymbol{\Upsilon}}_t$, respectively, as in this case, by the density property, we can find a sequence of continuity points $(a_n, \boldsymbol{b}_n) \downarrow (a, \boldsymbol{b})$ as $n \to \infty$. Thus, we need to establish (48) with $a$ and $\boldsymbol{b}$ being continuity points.

The key to the proof is the observation that, conditional on the backward recurrence time $a^n_t$, the RVs $\hat{A}^n_{t+1}$ and $\hat{\boldsymbol{\Upsilon}}^n_t$ are independent, i.e.

$$\mathrm{P}_{\pi_n}[\hat{A}^n_{t+1} \leq a,\ \hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b}] = \int_0^\infty \mathrm{P}[\hat{A}^n_{t+1} \leq a \mid a^n_t = z]\,\mathrm{P}_{\pi_n}[\hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b} \mid a^n_t = z]\,\mathrm{dP}[a^n_t \leq z].$$

By assumption (1) we have

$$\sup_{z \geq 0} |\,\mathrm{P}[\hat{A}^n_{t+1} \leq a \mid a^n_t = z] - \mathrm{P}[\hat{A}_{t+1} \leq a]| \leq \varepsilon$$

for all sufficiently large $n$. Therefore, for all such $n$, the following holds:

$$\begin{aligned}\mathrm{P}_{\pi_n}[\hat{A}^n_{t+1} \leq a,\ \hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b}] &\leq (\mathrm{P}[\check{A}_{t+1} \leq a] + \varepsilon) \int_0^\infty \mathrm{P}_{\pi_n}[\hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b} \mid a^n_t = z]\,\mathrm{dP}[a^n_t \leq z] \\ &\leq \mathrm{P}[\check{A}_{t+1} \leq a]\,\mathrm{P}_{\pi_n}[\hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b}] + \varepsilon.\end{aligned}$$

Recall that $\boldsymbol{b}$ is a continuity point of $\hat{\boldsymbol{\Upsilon}}^n_t$. Then the weak convergence, $\hat{\boldsymbol{\Upsilon}}^n_t \xrightarrow{\mathrm{D}} \check{\boldsymbol{\Upsilon}}_t$, implies that $\mathrm{P}_{\pi_n}[\hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b}] \to \mathrm{P}[\check{\boldsymbol{\Upsilon}}_t \leq \boldsymbol{b}]$ as $n \to \infty$, resulting in

$$\limsup_{n \to \infty} \mathrm{P}_{\pi_n}[\hat{A}^n_{t+1} \leq a,\ \hat{\boldsymbol{\Upsilon}}^n_t \leq \boldsymbol{b}] \leq \mathrm{P}[\check{A}_{t+1} \leq a]\,\mathrm{P}[\check{\boldsymbol{\Upsilon}}_t \leq \boldsymbol{b}] + \varepsilon.$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

Similarly, we establish

$$\liminf_{n\to\infty} \mathrm{P}_{\pi_n}[\hat{A}_{t+1}^n \le a,\, \hat{\boldsymbol{\Upsilon}}_t^n \le \boldsymbol{b}] \ge \mathrm{P}[\check{A}_{t+1} \le a]\,\mathrm{P}[\check{\boldsymbol{\Upsilon}}_t \le \boldsymbol{b}] - \varepsilon.$$

On the other hand, by the assumed weak convergence we have

$$\mathrm{P}_{\pi_n}[\hat{A}_{t+1}^n \le a,\, \hat{\boldsymbol{\Upsilon}}_t^n \le \boldsymbol{b}] \to \mathrm{P}[\check{A}_{t+1} \le a,\, \check{\boldsymbol{\Upsilon}}_t \le \boldsymbol{b}]$$

as $n \to \infty$ since $(a, \boldsymbol{b})$ is a continuity point of the vector $(\check{A}_{t+1}, \check{\boldsymbol{\Upsilon}}_t)$. Parameter $\varepsilon$ is arbitrary and, hence, the assertion of the proposition follows.

We have now developed the necessary tools for proving Theorem 1. In the first part of the proof we show the existence of a weak subsequential limit of $\hat{\boldsymbol{\Upsilon}}_t^n$, as $n \to \infty$, that must correspond to the stationary distribution of the Markov chain defined by (13), (14), and (15) (see Section 3). In the second part of the proof we argue that a stationary distribution of this Markov chain is unique.

*Part 1.* By Corollary 4, there exists a subsequence $\{n_k,\ k \ge 1\}$ along which a weak convergence $(\hat{\boldsymbol{\Upsilon}}_t^{n_k}, \hat{\boldsymbol{\Upsilon}}_{t+1}^{n_k}) \xrightarrow{\mathrm{D}} (\check{\boldsymbol{\Upsilon}}_t, \check{\boldsymbol{\Upsilon}}_{t+1})$ as $k \to \infty$ takes place [14, Section 2.2] for a fixed $t$ and a pair of random vectors $\check{\boldsymbol{\Upsilon}}_{\cdot} = (\check{Q}_{\cdot}, \check{\boldsymbol{L}}_{\cdot}, \check{A}_{\cdot}, \check{\boldsymbol{J}}_{\cdot}, \check{J}_{\cdot}, \check{Y}_{\cdot}, \check{a}_{\cdot})$. The continuous mapping theorem [10, Section 2] yields the following two weak limits along $\{n_k,\ k \ge 1\}$:

$$\Big(\hat{Q}_t^{n_k} + \hat{A}_{t+1}^{n_k} + \sum_{j=2}^{K} \hat{L}_{t,j}^{n_k} - \beta_{n_k}\Big)^+ \xrightarrow{\mathrm{D}} \Big(\check{Q}_t + \check{A}_{t+1} + \sum_{j=2}^{K} \check{L}_{t,j} - \beta\Big)^+,$$

$$(\hat{Q}_t^{n_k} + \hat{A}_{t+1}^{n_k}) \wedge \Big(\beta_{n_k} - \sum_{j=2}^{K} \hat{L}_{t,j}^{n_k}\Big) \xrightarrow{\mathrm{D}} (\check{Q}_t + \check{A}_{t+1}) \wedge \Big(\beta - \sum_{j=2}^{K} \check{L}_{t,j}\Big).$$

Then, from the preceding and Corollary 2, the following relations follow for the elements of $\check{\boldsymbol{\Upsilon}}_t$ and $\check{\boldsymbol{\Upsilon}}_{t+1}$:

$$\check{\boldsymbol{L}}_{t+1} = \mathcal{T}\{\check{\boldsymbol{L}}_t\} + \check{\boldsymbol{J}}_{t+1} + \check{J}_{t+1}\boldsymbol{p},$$

$$\check{Q}_{t+1} = \Big(\check{Q}_t + \check{A}_{t+1} + \sum_{j=2}^{K} \check{L}_{t,j} - \beta\Big)^+,$$

$$\check{J}_{t+1} = (\check{Q}_t + \check{A}_{t+1}) \wedge \Big(\beta - \sum_{j=2}^{K} \check{L}_{t,j}\Big).$$

Now, note that $\hat{A}_t^n \xrightarrow{\mathrm{D}} \hat{A}_t$ and $\hat{\boldsymbol{J}}_t^n \xrightarrow{\mathrm{D}} \hat{\boldsymbol{J}}_t$ as $n \to \infty$ for every $t \in \mathbb{Z}_+$. These weak limits are due to central limit theorems for renewal processes [10, p. 154] and vectors in $\mathbb{R}^K$ [9, p. 385], respectively. Moreover, $\hat{A}_t$ and $\hat{\boldsymbol{J}}_t$ are mutually independent in addition to the independence of $\check{A}_{t+1}$ and $\check{\boldsymbol{\Upsilon}}_t$ (see Proposition 5). Since $\pi_n$ is the stationary distribution of $\{\hat{\boldsymbol{\Upsilon}}_t^n,\ t \in \mathbb{Z}_+\}$, we find that the distribution of $\check{\boldsymbol{\Upsilon}}_t$ coincides with a stationary distribution of the Markov chain specified by (13), (14), and (15).

*Part 2.* We established in part 1 that every weak subsequential limit $\check{\boldsymbol{\Upsilon}}_t$ of $\hat{\boldsymbol{\Upsilon}}_t^n$ is a stationary distribution of the Markov chain $\{\hat{\boldsymbol{\Upsilon}}_t,\ t \in \mathbb{Z}_+\}$ defined by (13), (14), and (15). It remains to establish the uniqueness of the stationary measure $\pi_*$ of $\{\hat{\boldsymbol{\Upsilon}}_t,\ t \in \mathbb{Z}_+\}$. The uniqueness of this measure also implies the convergence $\pi_n \xrightarrow{\mathrm{D}} \pi_*$, using standard results of weak convergence theory [10, p. 59].

The proof of uniqueness uses the framework of Harris chains and Harris recurrence. All of the definitions and results are adopted from [14, Chapter 5]. Recall that the Markov chain $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$ is a Harris chain if there exist two (measurable) sets $\mathcal{A}$, $\mathcal{B} \subset \mathbb{R}^{K+1}$ and a probability measure $\nu$ concentrated on $\mathcal{B}$ such that, for every $\boldsymbol{x} \in \mathbb{R}^{K+1}$,

$$\sum_{t \geq 0} \mathrm{P}[(\hat{Q}_t, \hat{\boldsymbol{L}}_t) \in \mathcal{A} \mid (\hat{Q}_0, \hat{\boldsymbol{L}}_0) = \boldsymbol{x}] > 0,$$

and there exists $\varepsilon > 0$ such that, for every $\mathcal{C} \subset \mathcal{B}$,

$$\inf_{x \in \mathcal{A}} \mathrm{P}[(\hat{Q}_{t+1}, \hat{\boldsymbol{L}}_{t+1}) \in \mathcal{C} \mid (\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}] \geq \varepsilon \nu(\mathcal{C}). \tag{49}$$

Moreover, if these conditions hold for some $\mathcal{B} = \mathcal{A}$, and the Markov chain admits a stationary distribution, then the Markov chain is also mixing and, as a result, the stationary distribution is unique (see [14, Theorem 6.8] and the comment on aperiodicity just preceding it). Note that if $\pi$ is a stationary distribution of $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t), t \in \mathbb{Z}_+\}$ then $\pi$ is also a stationary distribution of $\{(\hat{Q}_{2Kt}, \hat{\boldsymbol{L}}_{2Kt}),\ t \in \mathbb{Z}_+\}$. In view of this, (49) can be replaced by

$$\inf_{x \in \mathcal{A}} \mathrm{P}[(\hat{Q}_{t+2K}, \hat{\boldsymbol{L}}_{t+2K}) \in \mathcal{C} \mid (\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}] \geq \varepsilon \nu(\mathcal{C}). \tag{50}$$

Thus, our task of proving the uniqueness of the stationary distribution $\pi_*$ is reduced to constructing the set $\mathcal{A} = \mathcal{B}$ satisfying the assumptions above. For this purpose, we set

$$\mathcal{A} = \biggl\{\boldsymbol{x} \in \mathbb{R}^{K+1} \colon x_1 = 0,\ |x_j| < \frac{\beta}{K^2},\ j = 2, \dots, K+1\biggr\}.$$

Namely, $(\hat{Q}_t, \hat{\boldsymbol{L}}_t) \in \mathcal{A}$ implies that the queue length $\hat{Q}_t$ is equal to 0 and each $\hat{L}_{t,j}$, $1 \leq j \leq K$, is upper bounded by $\beta/K^2$ in absolute value. We set $\mathcal{B} = \mathcal{A}$ and claim that $\mathcal{A}$ satisfies the requirements when $\nu$ is the uniform distribution on $\mathcal{A}$. For a pair of positive constants $c$ and $C$, define an event $\mathcal{U}$ by

$$\mathcal{U} = \{\hat{A}_{t+i} < -C,\ i = 1, \dots, K\} \cap \{|\hat{A}_{t+i}| < c,\ i = K+1, \dots, 2K\},$$

and note that $\mathrm{P}[\mathcal{U}] > 0$, owing to the Gaussian nature of $\hat{A}_t$.

First, we show that $\mathrm{P}[(\hat{Q}_{t+2K}, \hat{\boldsymbol{L}}_{t+2K}) \in \mathcal{A} \mid (\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}] > 0$ for every $\boldsymbol{x}$. To this end, given (13), (14), (15), and $(\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}$, there exists large enough $C$ so that

$$\mathrm{P}\biggl[\hat{Q}_{t+K} = 0, \bigvee_{i=1}^{K} \hat{L}_{t+K,i} < -\beta \biggm| \mathcal{U}\biggr] > 0. \tag{51}$$

To verify this claim, note that (13) implies that

$$\hat{L}_{t+K,i} = \sum_{j=i}^{K} (\hat{J}_{t+K+i-j,j} + p_j \hat{J}_{t+K+i-j}) \tag{52}$$

and that $\mathrm{P}[\bigvee_{i,j=1}^{K} |\hat{J}_{t+i,j}| \leq \varepsilon] > 0$ for any $\varepsilon > 0$, owing to the normal distribution. Then, by selecting $C > (\hat{Q}_t + \|\hat{L}_t\| + \beta + \varepsilon K)/p_K$ and small enough $\varepsilon$, recursions (14) and (15) render $\hat{Q}_{t+1} = 0$ and $\hat{J}_{t+1} = \hat{Q}_t + \hat{A}_{t+1} \leq -(\beta + \varepsilon K)/p_K$ on the event $\mathcal{U} \cap \{\bigvee_{i,j=1}^{K} |\hat{J}_{t+i,j}| \leq \varepsilon\}$;

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

this leads to $\hat{L}_{t+1,K} \leq \varepsilon - \beta - \varepsilon K$ by (52). Next, on the same event $\hat{Q}_{t+2} = 0$, $\hat{J}_{t+2} = \hat{A}_{t+2} \leq -(\beta + \varepsilon K)/p_K$, $\hat{L}_{t+2,K} \leq -\beta - \varepsilon(K-1)$, and $\hat{L}_{t+2,K-1} \leq -\beta - \varepsilon(K-2)$. Further iteration over the time index and (52) yield (51).

In addition, for small enough $c$ in the definition of $\mathcal{U}$, on the event $\{\hat{Q}_{t+K} = 0,\ \bigvee_i \hat{L}_{t+K,i} < -\beta\}$, we have $Q_{t+K+i} = 0$ and $\hat{J}_{t+K+i} = \hat{A}_{t+K+i}$ for $i = 1, \ldots, K$ by a similar argument as above. Then the components $2, \ldots, K+1$ of $(\hat{Q}_{t+2K}, \hat{\boldsymbol{L}}_{t+2K})$ are bounded in absolute value by $\beta/K^2$ provided that

$$\hat{J}'_i = \sum_{j=i}^{K} \hat{J}_{t+2K+i-j,j} \in \left\{\left[-\frac{\beta}{K^2}, \frac{\beta}{K^2}\right] - \sum_{j=i}^{K} p_j \hat{A}_{t+2K+i-j}\right\} \quad \text{for } i = 1, \ldots, K. \tag{53}$$

We denote by $\mathcal{E}$ the conjunction of $\mathcal{U}$ and the event described by (53). Recall that $\{\hat{\boldsymbol{J}}_t,\ t \in \mathbb{Z}_+\}$ is an i.i.d. sequence of multivariate Gaussian random vectors, independent from all other RVs, with the covariance matrix $\mu\boldsymbol{\Sigma}$ (see (12)). Thus, $(\hat{J}'_1, \ldots, \hat{J}'_K)$ is a zero-mean multivariate Gaussian vector with $\mathrm{E}[\hat{J}'^2_i] = \mu \sum_{j=i}^{K} (1 - p_j) p_j$ and $\mathrm{E}[\hat{J}'_i \hat{J}'_j] = -\mu \sum_{k=j}^{K} p_{k+i-j} p_k$, $i < j$. In particular, it has a continuous positive density everywhere on $\mathbb{R}^K$. To this end, assume that $(\hat{J}'_i, \hat{J}'_{i+1}, \ldots, \hat{J}'_K)$ has a continuous positive density everywhere on $\mathbb{R}^{K+1-i}$; this assumption holds for $i = K$ because $p_K > 0$. Then, $(\hat{J}'_{i-1}, \hat{J}'_i, \ldots, \hat{J}'_K)$ has a continuous density everywhere on $\mathbb{R}^{K+2-i}$ since

$$\hat{J}'_{i-1} = \hat{J}_{t+K+i-1,K} + \sum_{j=i-1}^{K-1} \hat{J}_{t+2K+i-1-j,j},$$

$\hat{J}_{t+K+i-1,K}$ is independent of $\{\hat{\boldsymbol{J}}_{t+K+j},\ j = i, \ldots, K\}$ and $(\hat{J}'_i, \hat{J}'_{i+1}, \ldots, \hat{J}'_K)$ is a deterministic function of $\{\hat{\boldsymbol{J}}_{t+K+j},\ j = i, \ldots, K\}$. Then, clearly, $\mathrm{P}[\mathcal{E} \mid (\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}] > 0$ for every $\boldsymbol{x}$.

Second, as in the preceding, by continuity and strict positivity of the density of $\hat{A}_t$ and $\hat{J}'_i$, there exists $\alpha > 0$ such that, for every set $\mathcal{C} \subset \mathcal{A}$,

$$\inf_{\boldsymbol{x} \in \mathcal{A}} \mathrm{P}[(\hat{Q}_{t+2K}, \hat{\boldsymbol{L}}_{t+2K}) \in \mathcal{C} \mid (\hat{Q}_t, \hat{\boldsymbol{L}}_t) = \boldsymbol{x}] \geq \alpha\nu(\mathcal{C}),$$

and requirement (50) holds. Thus, $\{(\hat{Q}_t, \hat{\boldsymbol{L}}_t),\ t \in \mathbb{Z}_+\}$ is indeed a Harris chain that admits a unique stationary distribution. This completes the proof.

## 6. Proof of Theorem 2

This section is devoted to proving our second main result, Theorem 2. The approach is based on the results of Section 4.4 for the limiting Markov chain $\{\hat{\boldsymbol{\Upsilon}}_t,\ t \in \mathbb{Z}_+\}$ in steady state. The proof utilizes the following preparatory lemma. The operators '$\leq$' and '$\geq$' are interpreted elementwise.

**Lemma 8.** *Let*

$$\boldsymbol{\Gamma}^\top = \begin{bmatrix} & \boldsymbol{p} \\ \boldsymbol{I} & \boldsymbol{0}^\top \end{bmatrix},$$

*where* $\boldsymbol{I}$ *is the* $(K-1) \times (K-1)$ *identity matrix and* $\mathbf{0}$ *is a* $(K-1)$*-dimensional vector of* 0s*. Then, for* $t \geq K-1$ *and* $k \geq 0$*,*

$$-\bar{\boldsymbol{V}}_{t+k} - \beta\boldsymbol{B}_k \leq \bar{\boldsymbol{Y}}_{t+k} - (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^k \leq \bar{\boldsymbol{V}}_{t+k},$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

*where* $\boldsymbol{B}_k = (k, (k-1)^+, \dots, (k-K+1)^+)$ *and*

$$\bar{\boldsymbol{V}}_t = \Bigg(\sum_{i=t-K+1}^{t} |\hat{V}_i|, \sum_{i=t-K+1}^{t-1} |\hat{V}_i|, \dots, \sum_{i=t-K+1}^{t-K+1} |\hat{V}_i|\Bigg).$$

**Remark 1.** Note that $\boldsymbol{\Gamma}^\top$ is an irreducible, aperiodic stochastic matrix since $\|\boldsymbol{p}\| = 1$, $p_K > 0$, and there exist relatively prime $i$ and $j$ such that $p_i p_j > 0$ (see Section 2.1.1). Therefore, $\boldsymbol{\Gamma}^k \to (\boldsymbol{\psi}^\top, \dots, \boldsymbol{\psi}^\top)$ as $k \to \infty$ for some unique probability vector $\boldsymbol{\psi}$.

*Proof of Lemma 8.* The proof is by induction over $k$. First, we claim that the statement holds for $k = 0$:

$$-\bar{\boldsymbol{V}}_t - \beta \boldsymbol{B}_0 \le \bar{\boldsymbol{Y}}_t - (\bar{\boldsymbol{Y}}_t)^+ \le \bar{\boldsymbol{V}}_t,$$

or in the scalar form

$$-\sum_{i=t-K+1}^{t-j} |\hat{V}_i| \le \hat{Y}_{t-j} - (\hat{Y}_{t-j})^+ \le \sum_{i=t-K+1}^{t-j} |\hat{V}_i|,$$

where $j = 0, 1, \dots, K-1$. The upper bound is trivial owing to the nonnegativity of $|\hat{V}_i|$ for all $i$; the same holds for the lower bound when $\hat{Y}_{t-j} \ge 0$. The case in which $\hat{Y}_{t-j} < 0$ is covered by Lemma 3(i) since it implies that $\hat{Y}_{t-j} \ge \hat{V}_{t-j}$. Now, assume that the statement holds for some $k$ and note that

$$(\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma} = \Bigg(\sum_{i=1}^{K} p_i \hat{Y}^+_{t+1-i}, \hat{Y}^+_t, \hat{Y}^+_{t-1}, \dots, \hat{Y}^+_{t-K+2}\Bigg),$$

$$\bar{\boldsymbol{V}}_t \boldsymbol{\Gamma} \le \Bigg(\sum_{i=t-K+1}^{t} |\hat{V}_i|, \sum_{i=t-K+1}^{t} |\hat{V}_i|, \sum_{i=t-K+1}^{t-1} |\hat{V}_i|, \dots, \sum_{i=t-K+1}^{t-K+2} |\hat{V}_i|\Bigg).$$

Consider the upper bound first. The preceding two relationships, Lemma 3(i), and the inductive assumption yield

$$\begin{aligned}
\bar{\boldsymbol{Y}}_{t+k+1} &\le (\bar{\boldsymbol{Y}}_{t+k})^+ \boldsymbol{\Gamma} + (|\hat{V}_{t+k+1}|, 0, \dots, 0) \\
&\le (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^{k+1} + \bar{\boldsymbol{V}}_{t+k} \boldsymbol{\Gamma} + (|\hat{V}_{t+k+1}|, 0, \dots, 0) \\
&\le (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^{k+1} + \bar{\boldsymbol{V}}_{t+k+1},
\end{aligned}$$

where $(x-\beta)^+ \le x^+$ is also used. As far as the lower bound is concerned, the same arguments and $(x-\beta)^+ \ge x^+ - \beta$ result in

$$\begin{aligned}
\bar{\boldsymbol{Y}}_{t+k+1} &\ge (\bar{\boldsymbol{Y}}_{t+k})^+ \boldsymbol{\Gamma} - (|\hat{V}_{t+k+1}| + \beta, 0, \dots, 0) \\
&\ge (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^{k+1} - \bar{\boldsymbol{V}}_{t+k} \boldsymbol{\Gamma} - \beta \boldsymbol{B}_k \boldsymbol{\Gamma} - (|\hat{V}_{t+k+1}| + \beta, 0, \dots, 0) \\
&\ge (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^{k+1} - \bar{\boldsymbol{V}}_{t+k+1} - \beta \boldsymbol{B}_{k+1}.
\end{aligned}$$

We now proceed with the proof of Theorem 2.

Proposition 3, Lemma 7, and Theorem 3 of Appendix B (where each $\Xi^n$ is identified with $\{\hat{\boldsymbol{\Upsilon}}_t, t \in \mathbb{Z}\}$, $\pi_n = \pi_*$, and $\mathcal{R}_{\beta_n} = \mathcal{R}_\beta$) yield

$$\mathrm{E}_{\pi_*}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)] < \infty \quad \text{for every } \theta < \theta^*/\mu. \tag{54}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

On the other hand, taking the expectation (with respect to $\pi_*$) of both sides of (43) implies that

$$\mathrm{E}_{\pi_*}[\Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t)] = \infty \quad \text{for every } \theta > \theta^*/\mu. \tag{55}$$

Next, definition (41) of $\Phi_\theta$ renders $\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t = \theta^{-1} \log \Phi_\theta(\bar{\boldsymbol{Y}}_t, \bar{\boldsymbol{Z}}_t) - \boldsymbol{\alpha} \cdot \bar{\boldsymbol{Z}}_t$. This equality, (54), (55), the normal distribution of $\bar{\boldsymbol{Z}}_t$, and Proposition 7 of Appendix A result in

$$\mathrm{E}_{\pi_*}[\exp(\theta \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t)] < \infty \quad \text{for every } \theta < \theta^*/\mu, \tag{56}$$

while

$$\mathrm{E}_{\pi_*}[\exp(\theta \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t)] = \infty \quad \text{for every } \theta > \theta^*/\mu. \tag{57}$$

Given (56) and (57), in order to complete the proof of the theorem it is sufficient to prove that, for every $\theta > 0$,

$$\mathrm{E}_{\pi_*}[\exp(\theta|\mu^{-1}\hat{Y}_t - \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t|)] < \infty, \tag{58}$$

or, equivalently, $|\mu^{-1}\hat{Y}_t - \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t| \in \mathcal{M}_\infty$, assuming the stationarity of $\{\hat{Y}_t,\ t \in \mathbb{Z}\}$. Informally, (58) implies that the stationary RVs $\mu^{-1}\hat{Y}_t$ and $\tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t$ have the same exponential decay rate.

The rest of the proof is devoted to establishing (58). Given that $\mu^{-1}\hat{Y}_t = \sum_k \tilde{p}_k \hat{Y}_t$, by Proposition 6 of Appendix A it suffices to show that $|\hat{Y}_t - \hat{Y}_{t-k}| \in \mathcal{M}_\infty$ for every $k = 1, \dots, K-1$ and stationary $\{\hat{Y}_t,\ t \in \mathbb{Z}\}$. Consider an arbitrary such $k$ and note that Lemma 8 renders, for $j \geq 1$ and $t \geq K-1$,

$$-\bar{\boldsymbol{V}}_{t+j} - \beta \boldsymbol{B}_j \leq \bar{\boldsymbol{Y}}_{t+j} - (\bar{\boldsymbol{Y}}_t)^+ \boldsymbol{\Gamma}^j \leq \bar{\boldsymbol{V}}_{t+j}.$$

Rewriting the preceding relationship in a scalar form renders

$$-\sum_{i=t-K+1}^{t+j-k} |\hat{V}_i| - (j-k)^+\beta \leq \hat{Y}_{t+j-k} - \sum_{i=0}^{K-1} (\Gamma^j)_{i+1,k+1} \hat{Y}^+_{t-i} \leq \sum_{i=t-K+1}^{t+j-k} |\hat{V}_i|,$$

and, hence,

$$|\hat{Y}_{t+j} - \hat{Y}_{t+j-k}| \leq \sum_{i=0}^{K-1} |(\Gamma^j)_{i+1,1} - (\Gamma^j)_{i+1,k+1}| \hat{Y}^+_{t-i} + 2 \sum_{i=t-K+1}^{t+j} |\hat{V}_i| + 2(j+K+1)\beta. \tag{59}$$

In view of Remark 1, the speed of convergence of $\boldsymbol{\Gamma}^k$ in $k$ is exponential [12, p. 211], i.e. there exist constants $C$ and $\gamma < 1$ such that

$$\sup_{1 \leq i \leq K} |(\Gamma^j)_{i,1} - (\Gamma^j)_{i,k+1}| \leq C\gamma^j.$$

Then, (59) and the preceding inequality yield

$$|\hat{Y}_{t+j} - \hat{Y}_{t+j-k}| \leq C\gamma^j \sum_{i=0}^{K-1} \hat{Y}^+_{t-i} + 2 \sum_{i=t-K+1}^{t+j} |\hat{V}_i| + 2(j+K+1)\beta. \tag{60}$$

Now, observe that the last two terms on the right-hand side of the preceding inequality are elements of $\mathcal{M}_\infty$ owing to (28), (29), and Proposition 6 of Appendix A. In addition, from $\hat{Y}_t = \tilde{\boldsymbol{p}} \cdot \bar{\boldsymbol{Y}}_t - \sum_{i=2}^{K} \tilde{p}_i \hat{Y}_{t-i+1}$ (Lemma 3(i)), (56), Lemma 3(ii), and Proposition 6, it follows that $\hat{Y}_t \in \mathcal{M}_{\theta'}$ for some sufficiently small $\theta' > 0$. By stationarity of $\{\hat{Y}_t,\ t \in \mathbb{Z}_+\}$, this applies to every term in the first sum on the right-hand side of (60). It then follows that $|\hat{Y}_t - \hat{Y}_{t-k}| \in \mathcal{M}_{\theta''}$ with $\theta'' = \gamma^{-j}\theta'/CK$. Since $j$ is arbitrary, by taking it sufficiently large, we establish that $|\hat{Y}_t - \hat{Y}_{t-k}| \in \mathcal{M}_\infty$. This concludes the proof of (58) and the proof of the theorem.

## 7. Proof of Corollary 1

First, we note that, for $x \geq 0$, as $n \to \infty$,

$$\frac{A^n_{0,x/\sqrt{n}}}{\sqrt{n}} \to \mu x \tag{61}$$

in probability. Let $\{\tau_{n,i},\ i \geq 1\}$ be interarrival times in the $n$th system, with $\tau_{n,1}$ being the time of the first arrival after time $t = 0$. The limit is based on the following: (i) $\{A_{0,t} \geq k\} = \{\sum_{i=1}^k \tau_{n,i} \leq t\}$ for $t \geq 0$ and $k \geq 1$; (ii) for large enough $n$, Markov's inequality yields, for $\varepsilon > 0$,

$$\begin{aligned} \mathrm{P}\Bigg[\sum_{i=2}^{\lfloor(\mu x+\varepsilon)\sqrt{n}\rfloor} \tau_{n,i} \leq \frac{x}{\sqrt{n}}\Bigg] &\leq \mathrm{P}\Bigg[\sum_{i=2}^{\lfloor(\mu x+\varepsilon)\sqrt{n}\rfloor}\bigg(\tau_{n,i} - \frac{1}{\lambda_n}\bigg) \leq -\frac{2\varepsilon}{\mu\sqrt{n}}\Bigg] \\ &\leq \mu^2(\mu x + \varepsilon)\varepsilon^{-2} n^{3/2}\operatorname{var}(\tau_{n,2}) \to 0 \quad \text{as } n \to \infty, \end{aligned}$$

and, similarly,

$$\mathrm{P}\Bigg[\sum_{i=2}^{\lceil(\mu x-\varepsilon)\sqrt{n}\rceil} \tau_{n,i} > \frac{x}{\sqrt{n}}\Bigg] \to 0 \quad \text{as } n \to \infty;$$

and (iii) the arrival processes are in stationarity and, thus, $\tau_{n,1}$ has the equilibrium distribution and does not impact (61).

Second, from the distributional Little's law [21] it follows that $Q^n$ equals in distribution the number of arrivals in a renewal process $A^n_t$ during the time interval of length $W^n$ (recall that $\{A^n_t,\ t \in \mathbb{R}\}$ is in stationarity), i.e. $Q^n = A^n_{0,W^n}$ in distribution. Then, for every $x > 0$, the event $\{W^n \leq x\}$ implies that $\{Q^n \leq A^n_{0,x}\}$ and, therefore,

$$\mathrm{P}_{\pi_n}[\sqrt{n}W^n \leq x] \leq \mathrm{P}_{\pi_n}\Bigg[\frac{Q^n}{\sqrt{n}} \leq \frac{A^n_{0,x/\sqrt{n}}}{\sqrt{n}}\Bigg].$$

The distribution of $\hat{Q}$ is continuous everywhere on $(0, \infty)$, as seen from the presence of $\hat{A}_{t+1}$ in the expression for $\hat{Q}_{t+1}$ in (14). Letting $n \to \infty$ in the preceding inequality and applying (61) yields

$$\limsup_{n\to\infty} \mathrm{P}_{\pi_n}[\sqrt{n}W^n \leq x] \leq \mathrm{P}_{\pi_*}[\hat{Q} \leq \mu x].$$

Similarly, for every $x > 0$, the event $\{W^n > x\}$ implies that $\{Q^n \geq A^n_{0,x}\}$, leading to

$$\mathrm{P}_{\pi_n}[\sqrt{n}W^n > x] \leq \mathrm{P}_{\pi_n}\Bigg[\frac{Q^n}{\sqrt{n}} \geq \frac{A^n_{0,x/\sqrt{n}}}{\sqrt{n}}\Bigg]$$

and

$$\liminf_{n\to\infty} \mathrm{P}_{\pi_n}[\sqrt{n}W^n \leq x] \geq \mathrm{P}_{\pi_*}[\hat{Q} \leq \mu x].$$

The preceding establishes $\mathrm{P}_{\pi_n}[\sqrt{n}W^n \leq x] \to \mathrm{P}_{\pi_*}[\hat{Q} \leq \mu x]$ as $n \to \infty$ for every $x > 0$. The assertion then follows.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

## 8. Conclusions

We analyzed a stationary multiserver queue in the Halfin–Whitt (QED) regime when the service times have a lattice-valued distribution with a finite support. The steady-state distribution of the appropriately scaled queue length was described in terms of the steady-state distribution of a continuous-state Markov chain; we can estimate the steady-state distribution of this chain either numerically or by simulations. We also established that the large-deviations rate of the limiting queue length in steady state is given by $\theta^* = 2\beta/(c_a^2 + c_s^2)$, where $\beta$ is the extra capacity parameter of the model and $c_a$ and $c_s$ are the coefficients of variation of interarrival and service times, respectively. We conjecture that the expression for $\theta^*$ remains valid for a broad class of service time distributions.

## Appendix A. Moment generating functions

Here we state some basic properties of moment generating functions. The proofs of these facts are obvious.

**Proposition 6.** *Any affine combination of (not necessarily independent) nonnegative elements of $\mathcal{M}_\infty$ is an element of $\mathcal{M}_\infty$.*

**Proposition 7.** *Suppose that $\{X_n,\ n \geq 1\} \in \mathcal{M}_\theta$ for some $\theta > 0$ and $\{Y_n,\ n \geq 1\} \in \mathcal{M}_\infty$. Then $\{X_n + Y_n,\ n \geq 1\} \in \mathcal{M}_{\theta'}$ for every $\theta' < \theta$.*

## Appendix B. Lyapunov functions

The following definitions play a key role in the proofs of our main results.

**Definition 1.** (*Geometric Lyapunov function.*) Let $\Xi = \{\Xi_t,\ t \in \mathbb{Z}_+\}$ be a discrete-time Markov chain defined on a state space $\mathcal{X}$, equipped with a $\sigma$-algebra $\mathcal{F}$. A function $\Phi\colon \mathcal{X} \to \mathbb{R}_+$ is defined to be a geometric Lyapunov function for $\Xi$ with a geometric drift size $0 < \delta < 1$ and exception set $\mathcal{R} \subset \mathcal{X}$ if, for every $x \in \mathcal{X} \setminus \mathcal{R}$,

$$\mathrm{E}[\Phi(\Xi_1) \mid \Xi_0 = x] \leq (1 - \delta)\Phi(x).$$

**Definition 2.** (*Quadratic Lyapunov function.*) Under the same setting as Definition 1, a function $\Psi\colon \mathcal{X} \to \mathbb{R}$ is defined to be a quadratic Lyapunov function for $\Xi$ with exception set $\mathcal{R} \subset \mathcal{X}$ and parameters $\delta > 0$ and $0 \leq \psi < \infty$ if, for every $x \in \mathcal{X} \setminus \mathcal{R}$,

$$\mathrm{E}[\Psi^2(\Xi_1) \mid \Xi_0 = x] - \Psi^2(x) \leq -\delta\Psi(x) + \psi.$$

Informally, the following result shows that if a sequence of Markov chains admits the same geometric Lyapunov function that is uniformly bounded in expectation in the exception region, then this function is uniformly bounded in expectation in general. Our definition of a geometric Lyapunov function as well as the following result is fairly standard [17], [33].

**Theorem 3.** *Let $\{\Xi^n,\ n \geq 1\}$ be a sequence of discrete-time Markov chains with $\mathcal{X}_n$ and $\pi_n$ being the state space and a stationary distribution of $\Xi^n$, respectively. Suppose that, for every $n \geq 1$, the function $\Phi\colon \cup\, \mathcal{X}_n \to \mathbb{R}_+$ is a geometric Lyapunov function for $\Xi^n$ with drift $\delta$ and exception set $\mathcal{R}_n \subset \mathcal{X}_n$. If*

$$C_{\mathcal{R}} := \limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Phi(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\}] < \infty \tag{62}$$

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

*then*

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Phi(\Xi_1^n)] \leq \frac{C_{\mathcal{R}}}{\delta}.$$

**Remark 2.** Note that the uniqueness of a stationary distribution $\pi_n$ is not assumed. Theorem 3 holds for *every* sequence of stationary distributions.

**Remark 3.** Our treatment of the geometric Lyapunov function is unconventional. Typically it is assumed that in the exception region the jumps $\Phi(\Xi_1^n) - \Phi(\Xi_0^n)$ are deterministically bounded; see, e.g. [33]. The intuition behind our result is as follows. The expected value of the Lyapunov function is uniformly bounded (in $n$) since (i) when the chain is in the exception region, $\Phi$ is bounded by assumption (in the next time step), and (ii) when the chain is outside of the exception region, there is a downward uniform drift decreasing the expected value of $\Phi$.

*Proof of Theorem 3.* The proof is similar to the approach taken in [17], and it is based on the monotone convergence theorem. Assumption (62) implies the existence of $n_0$ such that $\mathrm{E}_{\pi_n}[\Phi(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\}] < \infty$ for all $n > n_0$. Fix an arbitrary such $n$, introduce the following two conditional expectations:

$$G^b(x) := \mathrm{E}[\Phi(\Xi_1^n) \wedge b \mid \Xi_0^n = x],$$
$$H(x) := \mathrm{E}[\Phi(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\} \mid \Xi_0^n = x],$$

and let $G(x) = G^\infty(x)$ for notational simplicity. Then, by the Lyapunov nature of $\Phi$, the difference of $G(x)$ and $\Phi(x)$ for $x \in \mathcal{X}_n$ can be bounded as

$$G(x) - \Phi(x) \leq \begin{cases} -\delta\Phi(x), & x \in \mathcal{X}_n \setminus \mathcal{R}_n, \\ H(x) - \Phi(x), & x \in \mathcal{R}_n, \end{cases}$$

the second case being in fact the equality. Owing to the nonnegativity of $H(\cdot)$ and $\Phi(\cdot)$, the two cases in the preceding inequality can be combined into

$$G(x) - \Phi(x) \leq -\delta\Phi(x) + H(x) \quad \text{for all } x \in \mathcal{X}_n; \tag{63}$$

recall that $0 < \delta < 1$ by Definition 1. Furthermore, the preceding inequality, $G^b(x) \leq b$ (by definition) and the nonnegativity of $H(\cdot)$ yield

$$G^b(x) - \Phi(x) \wedge b \leq H(x), \qquad x \in \mathcal{X}_n; \tag{64}$$

the validity of the inequality can be verified by considering separately the cases $\Phi(x) < b$ and $\Phi(x) \geq b$. Then, (64) implies that

$$\mathrm{E}_{\pi_n}[G^b(\Xi_0^n) - \Phi(\Xi_0^n) \wedge b] \leq \mathrm{E}_{\pi_n}[H(\Xi_0^n)] < \infty, \tag{65}$$

where the strict inequality is due to the choice of $n > n_0$.

Now, the monotone convergence theorem renders $\{G^b(x) - \Phi(x) \wedge b\} \to \{G(x) - \Phi(x)\}$ as $b \to \infty$ for every $x \in \mathcal{X}_n$. Using Fatou's lemma, applicable due to (65) (see also [13, p. 44]), we obtain

$$\begin{aligned} \mathrm{E}_{\pi_n}[G(\Xi_0^n) - \Phi(\Xi_0^n)] &= \mathrm{E}_{\pi_n}\Big[\lim_{b\to\infty} \{G^b(\Xi_0^n) - \Phi(\Xi_0^n) \wedge b\}\Big] \\ &\geq \limsup_{b\to\infty} \mathrm{E}_{\pi_n}[G^b(\Xi_0^n) - \Phi(\Xi_0^n) \wedge b] \\ &= 0, \end{aligned} \tag{66}$$

where the last equality follows from the stationary nature of the distribution $\pi_n$.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

Finally, (63) and (66) result in $-\delta\,\mathrm{E}_{\pi_n}[\Phi(\Xi_0^n)] + \mathrm{E}_{\pi_n}[H(\Xi_0^n)] \geq 0$, and the conclusion of the theorem follows since this inequality holds for every $n > n_0$.

**Theorem 4.** *Let $\{\Xi^n,\ n \geq 1\}$ be a sequence of discrete-time Markov chains with $\mathcal{X}_n$ and $\pi_n$ being the state space and a stationary distribution of $\Xi^n$, respectively. Suppose that, for every $n \geq 1$, the function $\Psi\colon \cup\,\mathcal{X}_n \to \mathbb{R}$ satisfies*

$$\mathrm{E}[\Psi^2(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \notin \mathcal{R}_n\} - \Psi^2(\Xi_0^n) \mid \Xi_0^n] \leq -\delta\Psi(\Xi_0^n) + \psi \tag{67}$$

*for some $\delta > 0$, $0 \leq \psi < \infty$, and $\mathcal{R}_n \subset \mathcal{X}_n$. If*

$$C_{\mathcal{R}} := \limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Psi^2(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\}] < \infty \tag{68}$$

*and*

$$C_0 := \limsup_{n\to\infty} \mathrm{E}_{\pi_n}[-\Psi(\Xi_0^n)\,\mathbf{1}\{\Psi(\Xi_0^n) < 0\}] < \infty, \tag{69}$$

*then*

$$\limsup_{n\to\infty} \mathrm{E}_{\pi_n}[\Psi(\Xi_1^n)] \leq \frac{C_{\mathcal{R}} + C_0 + \psi}{\delta}.$$

**Remark 4.** A nonstandard part of our definition of the quadratic Lyapunov function is allowing $\Psi$ to be negative. Our second result in this section shows that if a sequence of Markov chains admits the same quadratic Lyapunov function that is uniformly bounded in expectation in the exception region, then the (linear part of this) function is uniformly bounded away from $+\infty$.

*Proof of Theorem 4.* The proofs of Theorems 3 and 4 are similar. Assumptions (68) and (69) imply the existence of $n_0$ such that $\mathrm{E}_{\pi_n}[\Psi^2(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\} - \Psi(\Xi_0^n)\,\mathbf{1}\{\Psi(\Xi_0^n) < 0\}] < \infty$ for all $n > n_0$. Fix an arbitrary such $n$, introduce the following two conditional expectations:

$$G^b(x) := \mathrm{E}[\Psi^2(\Xi_1^n) \wedge b \mid \Xi_0^n = x],$$
$$H(x) := \mathrm{E}[\Psi^2(\Xi_1^n)\,\mathbf{1}\{\Xi_0^n \in \mathcal{R}_n\} \mid \Xi_0^n = x],$$

and let $G(x) = G^\infty(x)$ for notational simplicity. Then, by (67), the difference of $G(x)$ and $\Psi^2(x)$ for $x \in \mathcal{X}_n$ can be bounded as

$$G(x) - \Psi^2(x) \leq -\delta\Psi(x) + \psi + H(x). \tag{70}$$

Furthermore, the preceding inequality, $G^b(x) \leq b$ (by definition), and the nonnegativity of $H(\cdot)$ yield

$$G^b(x) - \Psi^2(x) \wedge b \leq -\delta\Psi(x)\,\mathbf{1}\{\Psi(x) < 0\} + \psi + H(x), \qquad x \in \mathcal{X}_n; \tag{71}$$

the validity of the inequality can be verified by considering separately the cases $\Psi^2(x) < b$ and $\Psi^2(x) \geq b$. Then, (71) implies that

$$\mathrm{E}_{\pi_n}[G^b(\Xi_0^n) - \Phi(\Xi_0^n) \wedge b] \leq \delta\,\mathrm{E}_{\pi_n}[-\Psi(\Xi_0^n)\,\mathbf{1}\{\Psi(\Xi_0^n) < 0\}] + \psi + \mathrm{E}_{\pi_n}[H(\Xi_0^n)] < \infty, \tag{72}$$

where the strict inequality is due to the choice of $n > n_0$.

Now, the monotone convergence theorem renders $\{G^b(x) - \Psi^2(x) \wedge b\} \to \{G(x) - \Psi^2(x)\}$ as $b \to \infty$ for every $x \in \mathcal{X}_n$. Using Fatou's lemma, applicable due to (72), we obtain

$$\begin{aligned} \mathrm{E}_{\pi_n}[G(\Xi_0^n) - \Psi^2(\Xi_0^n)] &= \mathrm{E}_{\pi_n}\Big[\lim_{b\to\infty}\{G^b(\Xi_0^n) - \Psi^2(\Xi_0^n) \wedge b\}\Big] \\ &\geq \limsup_{b\to\infty} \mathrm{E}_{\pi_n}[G^b(\Xi_0^n) - \Psi^2(\Xi_0^n) \wedge b] \\ &= 0, \end{aligned} \tag{73}$$

where the last equality follows from the stationary nature of the distribution $\pi_n$.

Finally, (70) and (73) result in $-\delta\, \mathrm{E}_{\pi_n}[\Psi(\Xi_0^n)] + \psi + \mathrm{E}_{\pi_n}[H(\Xi_0^n)] \geq 0$, and the conclusion of the theorem follows since this inequality holds for every $n > n_0$.

## Acknowledgements

PM thanks Itay Gurvich and Avishai Mandelbaum for discussions on the QED regime. DG gratefully acknowledges support from the NSF, grant number CMMI-0726733.

## References

[1] Aksin, Z., Armony, M. and Mehrotra, V. (2007). The modern call center: a multi-disciplinary perspective on operations management research. *Production Operat. Manag.* **16,** 665–668.

[2] Armony, M. and Maglaras, C. (2004). Contact centers with a call-back option and real-time delay information. *Operat. Res.* **52,** 527–545.

[3] Armony, M. and Maglaras, C. (2004). On customer contact centers with a call-back option: customer decisions, sequencing rules and system design. *Operat. Res.* **52,** 271–292.

[4] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.

[5] Atar, R. (2005). A diffusion model of scheduling control in queueing systems with many servers. *Ann. Appl. Prob.* **15,** 820–852.

[6] Atar, R. (2005). Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. *Ann. Appl. Prob.* **15,** 2606–2650.

[7] Atar, R., Mandelbaum, A. and Reiman, M. (2004). Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *Ann. Appl. Prob.* **14,** 1084–1134.

[8] Baccelli, F. and Brémaud, P. (2003). *Elements of Queueing Theory*, 2nd edn. Springer, Berlin.

[9] Billingsley, P. (1995). *Probability and Measure*, 3rd edn. John Wiley, New York.

[10] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd edn. Wiley, New York.

[11] Borst, S., Mandelbaum, A. and Reiman, M. (2004). Dimensioning of large call centers. *Operat. Res.* **52,** 17–34.

[12] Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, New York.

[13] Chung, K. L. (1974). *A Course in Probability Theory*, 2nd edn. Academic Press, New York.

[14] Durrett, R. (2005). *Probability: Theory and Examples*, 3rd edn. Thomson, Belmont, CA.

[15] Erlang, A. K. (1948). On the rational determination of the number of circuits. In *The Life and Works of A. K. Erlang*, eds E. Brockmeyer *et al.*, The Copenhagen Telephone Company, pp. 216–221.

[16] Fleming, P., Stolyar, A. and Simon, B. (1994). Heavy traffic limit for a mobile phone system loss model. In *Proc. 2nd Internat. Conf. Telecommun. Syst. Model. Analysis* (Nashville, TN).

[17] Gamarnik, D. and Zeevi, A. (2006). Validity of heavy traffic steady-state approximations in open queueing networks. *Ann. Appl. Prob.* **16,** 56–90.

[18] Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone call centers: tutorial, review and research prospects. *Manufacturing Service Operat. Manag.* **5,** 79–141.

[19] Garnett, O., Mandelbaum, A. and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing Service Operat. Manag.* **4,** 208–227.

[20] Gurvich, I., Armony, M. and Mandelbaum, A. (2008). Service level differentiation in call centers with fully flexible servers. *Manag. Sci.* **54,** 279–294.

[21] Haji, R. and Newell, G. (1971). A relationship between stationary queue and waiting time distributions. *J. Appl. Prob.* **8,** 617–620.

[22] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operat. Res.* **29,** 567–588.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press

[23] Harrison, J. M. and Zeevi, A. (2004). Dynamic scheduling of a multiclass queue in the Halfin–Whitt heavy traffic regime. *Operat. Res.* **52,** 243–257.

[24] Jagerman, D. (1974). Some properties of the Erlang loss function. *Bell System Tech. J.* **53,** 525–551.

[25] Jelenković, P., Mandelbaum, A. and Momčilović, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems Theory Appl.* **47,** 53–69.

[26] Kiefer, J. and Wolfowitz, J. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc.* **78,** 1–18.

[27] Kingman, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Camb. Philos. Soc.* **57,** 902–904.

[28] Kingman, J. F. C. (1964). The heavy traffic approximation in the theory of queues. In *Proc. Symp. Congestion Theory*, eds W. L. Smith and R. I. Wilkinson, University of North Carolina Press, Chapel Hill, pp. 137–169.

[29] Lovasz, P., Pelikan, J. and Vesztergombi, K. (2003). *Discrete Mathematics: Elementary and Beyond*. Springer, New York.

[30] Maglaras, C. and Zeevi, A. (2003). Pricing and capacity sizing for systems with shared resources: approximate solutions and scaling relations. *Manag. Sci.* **49,** 1018–1038.

[31] Mandelbaum, A. and Momčilović, P. (2008). Queues with many servers: the virtual waiting-time process in the QED regime. To appear in *Math. Operat. Res.*

[32] Mandelbaum, A. and Zeltyn, S. (2006). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Preprint, Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology.

[33] Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

[34] Puhalskii, A. (1994). On the invariance principle for the first passage time. *Math. Operat. Res.* **19,** 946–954.

[35] Puhalskii, A. and Reiman, M. (2000). The multiclass GI/PH/N queue in the Halfin–Whitt regime. *Adv. Appl. Prob.* **32,** 564–595.

[36] Reed, J. (2007). The G/GI/N queue in the Halfin–Whitt regime. Preprint, Stern School of Business, New York University.

[37] Tezcan, T. (2008). Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Math. Operat. Res.* **33,** 51–90.

[38] Whitt, W. (2002). *Stochastic-Process Limits*. Springer, New York.

[39] Whitt, W. (2004). A diffusion approximation for the G/GI/$n$/$m$ queue. *Operat. Res.* **52,** 922–941.

[40] Whitt, W. (2005). Heavy-traffic limits for the G/$\mathrm{H}_2^*$/$n$/$m$ queue. *Math. Operat. Res.* **30,** 1–27.

[41] Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the M/M/$n$+G queue. *Queueing Syst. Theory Appl.* **51,** 361–402.

https://doi.org/10.1239/aap/1214950216 Published online by Cambridge University Press