Adv. Appl. Prob. **48,** 812–831 (2016)
*doi:*10.1017/apr.2016.29

G/G/∞ QUEUES WITH RENEWAL ALTERNATING INTERRUPTIONS

GUODONG PANG * ** AND
YUHANG ZHOU,* *** *Pennsylvania State University*

Abstract

We study G/G/∞ queues with renewal alternating service interruptions, where the service station experiences 'up' and 'down' periods. The system operates normally in the up periods, and all servers stop functioning while customers continue entering the system during the down periods. The amount of service a customer has received when an interruption occurs will be conserved and the service will resume when the down period ends. We use a two-parameter process to describe the system dynamics: $X^r(t, y)$ tracking the number of customers in the system at time t that have residual service times strictly greater than y. The service times are assumed to satisfy either of the two conditions: they are independent and identically distributed with a distribution of a finite support, or are a stationary and weakly dependent sequence satisfying the ϕ-mixing condition and having a continuous marginal distribution function. We consider the system in a heavy-traffic asymptotic regime where the arrival rate gets large and service time distribution is fixed, and the interruption down times are asymptotically negligible while the up times are of the same order as the service times. We show the functional law of large numbers and functional central limit theorem (FCLT) for the process $X^r(t, y)$ in this regime, where the convergence is in the space $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$ endowed with the Skorokhod M_1 topology. The limit processes in the FCLT possess a stochastic decomposition property.

Keywords: G/G/∞ queue; dependent service time; service interruption; two-parameter stochastic process; FLLN; FCLT; Skorokhod M_1 topology

2010 Mathematics Subject Classification: Primary 60K25; 60F17; 90B22; 60J75
Secondary 60G44; 54A20

1. Introduction

We study G/G/∞ queues with service interruptions. The service station operates in a renewal alternating environment with 'up' and 'down' periods. All servers function normally in an 'up' period, but break down in a 'down' period. Customers enter the system according to some point process, without being affected by the interruptions. They are served immediately upon arrival if arriving during an 'up' period, and wait for service until the next 'up' period starts otherwise. Services received when interruptions occur will be conserved, and when interruptions end, services resume from where they are left. Customers may experience multiple interruption 'down' periods before completing their services. To describe the system dynamics, we use a two-parameter process $X^r(t, y)$, representing the number of customers in the system at time t that have residual service times strictly greater than y. The second parameter will help track

Received 25 November 2014; revision received 28 August 2015.
* Postal address: The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA.
** Email address: gup3@psu.edu
*** Email address: yxz197@psu.edu

the amount of interrupted service times. The total number of customers in the system at time t is $X(t) = X^r(t, 0)$.

We consider the system in a heavy-traffic asymptotic regime. The arrival rate becomes large while the service time distributions are fixed. The alternating renewal 'up' and 'down' times are scaled such that the 'up' time periods are of the same order as service times, while the 'down' time periods are asymptotically negligible compared with service times. Specifically, we let the 'down' times be of order $O(1/n^\gamma)$ for some $\gamma > 0$ while the service times are $O(1)$. We study the system under two different assumptions on the service times:

(i) independent and identically distributed (i.i.d.) service times with a finite support, and

(ii) the sequence of service times is stationary and weakly dependent, satisfying the ϕ-mixing condition and having a continuous marginal distribution function.

We show the functional law of large numbers (FLLNs) and functional central limit theorems (FCLTs) for the processes $X^r(t, y)$ and $X(t)$ in the asymptotic regime under the two assumptions on the service times.

The limit processes in the FLLNs and FCLTs of the process $X^r(t, y)$ have sample paths in the space $\mathbb{D}_{\mathbb{D}} := \mathbb{D}([0, \infty), \mathbb{D}([0, \infty), \mathbb{R}))$. To prove the convergence, due to unmatched jumps of prelimit and limit processes, we employ the Skorokhod M_1 topology in the space $\mathbb{D}_{\mathbb{D}}$, where the interior space $\mathbb{D} := \mathbb{D}([0, \infty), \mathbb{R})$ is endowed with the L_1 norm. The fluid limits take the same forms as those in the corresponding G/G/∞ queues without interruptions [11], [25], [28]. However, the convergence is proved in the Skorokhod M_1 topology. The limits in the FCLTs have extra terms to capture the effect of service interruptions, in addition to the same terms as to those in the corresponding G/G/∞ queues without interruptions [11], [25], [28]. We obtain a stochastic decomposition property for the limiting processes in the FCLTs, that is, the variabilities of arrivals, services, and interruptions are captured in mutually independent processes. When the arrival limit process is a Brownian motion and the limiting counting process for the number of 'up' and 'down' periods is Poisson, we characterize the transient and stationary distributions for the limiting processes (Corollaries 1 and 2).

1.1. Literature and contributions

There has been a large body of literature on queues with service interruptions; see, e.g. [16], [17], [23], and [22] and the references therein. Infinite-server queues with service interruptions (or in a random environment) were studied in [1], [2], [6], [8], [20], [21], and [22]; however, all these studies have focused on either Poisson arrivals and/or exponential service times. Infinite-server queues with general service times and service interruptions (or in a random environment) were studied in [4], [5], [9], and [15]. To the best of the authors' knowledge this work is the first to establish two-parameter heavy-traffic limits for G/G/∞ queues with service interruptions under general assumptions on the arrival, service, and interruption processes.

Infinite-server queues with general arrival and service processes have been studied in [10], [11], [18], and [34] and more recently in [25], [28], and [30]. In [11], when the service time distribution has a finite support, the key idea is to treat that as a mixture of deterministic service time distributions, and split the arrival process into corresponding arrival processes associated with each deterministic service time. The main step relies on the FCLT for the split counting processes. That key idea can be used to show an FLLN and an FCLT for the total count process $X(t)$ in the system as well as the two-parameter process $X^r(t, y)$. We first generalize that approach to G/G/∞ queues with service interruptions when the service times have finite support. In [25], and [28], Pang and Whitt established an FLLN and an FCLT for the

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

two-parameter process $X^r(t, y)$ when the service times are i.i.d. with continuous distributions, and when the service times form a stationary and weakly dependent sequence satisfying certain mixing conditions and have continuous marginal distributions. The approach in [25], and [28] relied on the important observation first made in [18] that the queue length process in G/GI/∞ queues can be represented via sequential empirical processes driven by service times. Here we make a novel and important observation that for G/G/∞ queues with service interruptions, the two-parameter process $X^r(t, y)$ can be represented via sequential empirical processes driven by service times together with the cumulative up time process. This enables us to establish the limits in the fluid and diffusion scales. Here we assume that the service time distributions are either of finite support or weakly dependent with continuous marginal distributions as in [11] and [28]. It is conceivable that the results can be extended to the most general conditions on service times, but that is beyond the scope of this paper. Our approach is extended to study the total count processes for G/G/N(+G) queues in the Halfin–Whitt regime in [19] when the services are i.i.d. with a continuous distribution function. It remains to prove the two-parameter limits for these models. We also conjecture that it can be potentially extended to study other many-server non-Markovian models.

An interesting stochastic decomposition property has been established for M/M/∞ and M/G/∞ queues with service interruptions or vacations in [2], [5], [8], [15], and [21]. Namely, the steady-state distribution of the number of customers in the system can be decomposed into two independent components: the steady-state distribution of the number of customers in the system without interruptions and the distribution taking into account the effects of service interruptions. In the heavy-traffic setting, similar stochastic decomposition properties were proved for G/M/∞ queues in [22] and for G/G/1 queues in [16] and [17]. In this paper we establish a stochastic decomposition property for the two-parameter process $X^r(t, y)$ and the total count process $X(t)$ of G/G/∞ queues with service interruptions under the two assumptions on the service time distributions. As a result, their steady-state distributions also possess a stochastic decomposition property. In fact, as shown in Theorems 2 and 4, the limiting two-parameter processes and total count processes in the diffusion scale are decomposed into independent processes which capture variabilities in the arrival processes, service processes, and service interruptions.

To prove the convergence of the fluid and diffusion-scaled processes for queues with service interruptions, the Skorokhod M_1 topology in the space $\mathbb{D}$ is necessary to take into account the unmatched jumps and discontinuity when the interruption 'down' times are scaled properly. For example, single-server queues with service interruptions in the conventional heavy-traffic regime [16], [17] and G/M/N(+M) queues with service interruptions in the many-server heavy-traffic regimes [22], [24] all require the convergence in $(\mathbb{D}, M_1)$. To prove the convergence of the two-parameter process $X^r(t, y)$ of G/G/∞ queues with service interruptions, it requires to use the Skorokhod M_1 topology in the space $\mathbb{D}_{\mathbb{D}}$. Since the Skorokhod M_1 topology is only well defined for the space $\mathbb{D}([0, \infty), \mathcal{S})$ when $\mathcal{S}$ is a Banach space [29], [32], [35], we endow the interior space $\mathbb{D}$ with the L_1 norm. This topology is discussed in the study of G/GI/∞ queues in [35, Section 10.3.1, p. 350]. We establish some useful continuity properties in the space $\mathbb{D}_{\mathbb{D}}$ with Skorokhod M_1 topology, which may be of separate interest.

1.2. Notation

We use $\mathbb{R}^k$ (and $\mathbb{R}^k_+$), $k \geq 1$, to denote real-valued k-dimensional (nonnegative) vectors, and write $\mathbb{R}$ and $\mathbb{R}_+$ for $k = 1$. Let $\mathbb{D}^k = \mathbb{D}([0, \infty), \mathbb{R}^k)$ denote the $\mathbb{R}^k$-valued function space of all right-continuous functions on $[0, \infty)$ with left limits everywhere in $(0, \infty)$. Denote $\mathbb{D} \equiv \mathbb{D}^1$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Let $(\mathbb{D}_k, J_1) = (\mathbb{D}, J_1) \times \cdots (\mathbb{D}, J_1)$ be the k-fold product of $(\mathbb{D}, J_1)$ with the product topology. Similarly, let $(\mathbb{D}_k, M_1) = (\mathbb{D}, M_1) \times \cdots (\mathbb{D}, M_1)$ be the k-fold product of $(\mathbb{D}, M_1)$ with the product topology. We use $\mathbb{C}_\uparrow$ and $\mathbb{D}_\uparrow$ to denote the space of nondecreasing functions in $\mathbb{C}$ and $\mathbb{D}$, respectively. Denote $\|\cdot\|$ the uniform norm: for any real-valued function x and $T > 0$, $\|x\|_T = \sup_{t\in[0,T]} |x(t)|$. The notation '$\to$' and '$\Rightarrow$' mean convergence of real numbers and convergence in distributions, respectively. Without abusing notation, we write '$\Rightarrow$' for both weak convergence and convergence in distributions. The abbreviations 'w.p.1' and 'a.e.' mean *with probability* 1 and *almost everywhere*, respectively. We use the familiar upper-case O and lower-case o notation for deterministic functions: for two real-valued functions f and nonzero g, we write $f(x) = O(g(x))$ if $\limsup_{x\to\infty} |f(x)/g(x)| < \infty$ and $f(x) = o(g(x))$ if $\limsup_{x\to\infty} |f(x)/g(x)| = 0$.

All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any two complete separable metric spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, we denote $\mathcal{S}_1 \times \mathcal{S}_2$ as their product space, endowed with the product topology. Let $\mathbb{D}_\mathbb{D} = \mathbb{D}([0,\infty), \mathbb{D})$ denote the $\mathbb{D}$-valued function space of all right-continuous functions on $[0,\infty)$ with left limits everywhere in $(0,\infty)$. Let $\mathbb{C}_\mathbb{C} = \mathbb{C}([0,\infty), \mathbb{C})$, where $\mathbb{C}$ denotes the real-valued function space of all continuous functions on $[0,\infty)$. For $x(t,s) \in \mathbb{D}_\mathbb{D}$, we often denote $x(t) := x(t,\cdot)$. We use $(\mathbb{D}_\mathbb{D}, J_1)$ to denote the space $\mathbb{D}_\mathbb{D}$ with both the interior and exterior $\mathbb{D}$ spaces endowed with the Skorokhod J_1 topology. We refer the reader to [33] and [25] for an introduction and some properties of $(\mathbb{D}_\mathbb{D}, J_1)$. We use $(\mathbb{D}_\mathbb{D}, M_1)$ to denote the space $\mathbb{D}_\mathbb{D}$ with the exterior $\mathbb{D}$ space endowed with the Skorokhod M_1 topology, while the interior $\mathbb{D}$ space is endowed with the L_1 norm, that is, for any $x \in \mathbb{D}$, the L_1 norm is defined by $\|x\|_{L_1} = \int_0^\infty |x(t)|\,\mathrm{d}t$. See [35, Sections 10.3 and 11.5]. Note that the space $(\mathbb{D}, L_1)$ is a Banach space, and, thus, the space $\mathbb{D}([0,\infty), (\mathbb{D}, L_1))$ with Skorokhod M_1 topology is well defined [29], [32], [35]. Let $((\mathbb{D}_\mathbb{D})_k, M_1) = (\mathbb{D}_\mathbb{D}, M_1) \times \cdots \times (\mathbb{D}_\mathbb{D}, M_1)$ be the k-fold product of $(\mathbb{D}_\mathbb{D}, M_1)$ with the product topology.

1.3. Organization of the paper

The rest of the paper is organized as follows. In Section 2 we describe the model in detail, and the assumptions on the interruptions are given in Section 2.1. We present the results for service times with finite support in Section 3 and their proofs in Section 5. The results for service times that form a stationary and weakly dependent sequence and have continuous marginal distributions are presented in Section 4 and their proofs are in Section 6. Some technical proofs in Sections 5 and 6 are given in the Appendix A.

2. The model

Consider a G/G/∞ queue subject to service interruptions. The service station is in a renewal alternating environment with *up* and *down* periods, where all servers function normally during the up period but breakdown during the down period. During an up period, customers will enter service immediately (no delay or queue). During a down period, customers in service will wait at their associated servers until the end of the down period and resume their services when the next up period starts, and new customers will continue to enter the system and get assigned to some free servers and wait there for service to start until the next up period begins.

Assume that customers arrive to the system according to a general arrival process $A = \{A(t) : t \ge 0\}$ with τ_i representing the arrival time of the ith customer, that is,

$$A(t) = \max\{i \ge 1 : \tau_i \le t\} \quad \text{for each } t > 0 \quad \text{and} \quad A(0) = 0.$$

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Let $\{\eta_i : i \geq 1\}$ be a sequence of random variables, with η_i representing the service times of the ith customer. We will consider two assumptions on the sequence $\{\eta_i : i \geq 1\}$:

(i) i.i.d. with a finite support, and

(ii) stationary and weakly dependent, satisfying the ϕ-mixing condition and having a general continuous marginal distribution function (see Assumption 4 for the precise statement).

This second assumption includes i.i.d. service times with a general continuous distribution function as a special case. Let $\{(u_i, d_i) : i \geq 1\}$ be a sequence of i.i.d. random vectors with u_i and d_i representing the up and down times in the ith *up–down* cycle of the underlying renewal process. We assume that the arrival, service, and service-interruption processes are mutually independent.

Let $X^r(t, y)$ be the number of customers in the system at time t that have a remaining amount of service time strictly greater than y. Let $X(t)$ be the total number of customers in the system at time t. Then $X(t) = X^r(t, 0)$ for each $t \geq 0$. We assume that the system starts empty at the starting epoch of an up period.

To describe the dynamics of $X^r(t, y)$, we first define some auxiliary processes associated with the underlying renewal process $\{(u_i, d_i) : i \geq 1\}$. Let the sequence $\{T_i : i \geq 1\}$ be the renewal times, defined by $T_i = \sum_{k=1}^{i}(u_k + d_k) \, for \, i \geq 1 \, and \, T_0 = 0$. Let $N = \{N(t) : t \geq 0\}$ be the associated renewal counting process, defined by $N(t) = \max\{i \geq 0 : T_i \leq t\}$, $t \geq 0$. Let $\xi = \{\xi(t) : t \geq 0\}$ be the service-availability process, defined by $\xi(t) = 1$ when $T_i \leq t \leq T_i + u_{i+1}$, and $\xi(t) = 0$ when $T_i + u_i < t < T_{i+1}$ for $i \geq 0$. The cumulative up-time process $U = \{U(t) : t \geq 0\}$ is defined by $U(t) = \int_0^t \xi(s) \, ds$, $t \geq 0$. The cumulative down-time process $D = \{D(t) : t \geq 0\}$ is defined by $D(t) = t - U(t)$ for each $t \geq 0$.

We now give a representation of the processes $X^r(t, y)$ and $X(t)$:

$$X^r(t, y) = \sum_{i=1}^{A(t)} \mathbf{1}_{\{\eta_i > U(t) - U(\tau_i) + y\}}, \qquad t, y \geq 0, \quad \text{and} \quad X(t) = X^r(t, 0), \tag{1}$$

where $\mathbf{1}$ is the indicator function. We give an intuitive explanation of (1). For the process $X^r(t, y)$, we count the arrivals up to time t that have residual service times strictly greater than y. The amount of service that the ith customer has received by time t, if he/she is still in the system, is equal to $U(t) - U(\tau_i)$, the cumulative up time from his/her arrival time τ_i to time t. Thus, for the ith customer to be counted at time t, the amount of received service $U(t) - U(\tau_i)$ must be strictly less than the service requirement η_i minus y, that is, $\eta_i > U(t) - U(\tau_i) + y$.

We will consider a sequence of systems indexed by n and let $n \to \infty$, where the arrival processes and the underlying interruption processes are scaled while the service time distribution does not change with n. We let the arrival rate increase to ∞ as $n \to \infty$ and the underlying interruption processes have the up times $O(1)$ and the down times $O(1/n^\gamma)$ for some positive constant $\gamma > 0$ (see Section 2.1).

2.1. Assumptions on service interruptions

We consider a scaling regime for the underlying service interruption process, where the down times are asymptotically negligible.

Assumption 1. *The sequence of up and down times* $\{(u_i^n, d_i^n) : i \geq 1\}$ *satisfies*

$$\{(u_i^n, n^\gamma d_i^n) : i \geq 1\} \Rightarrow \{(u_i, d_i) : i \geq 1\} \text{ in } (\mathbb{R}^2)^\infty \quad \text{as } n \to \infty,$$

for some $\gamma > 0$, *where* $u_i, d_i > 0$ *for each* i *w.p.*1.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Let $\{\hat{T}_i : i \geq 0\}$ be the associated sequence of cycle renewal times in the limit, defined by $\hat{T}_i = \sum_{k=1}^{i} u_k$, $i \geq 1$, and $\hat{T}_0 = 0$. Assumption 1 implies that $\hat{T}_i < \hat{T}_{i+1}$ for all $i \geq 0$. Let $\hat{N} = \{\hat{N}(t) : t \geq 0\}$ be the associated renewal counting process in the limit, defined by

$$\hat{N}(t) = \max\{i \geq 0 : \hat{T}_i \leq t\}, \qquad t \geq 0. \tag{2}$$

Define the diffusion-scaled processes $\hat{U}^n = n^\gamma(U^n - e)$ and $\hat{D}^n = n^\gamma D^n = -\hat{U}^n$, where $e(t) \equiv t$ for $t \geq 0$, and U^n and D^n are the cumulative up and down times.

Lemma 1. *Under Assumption 1,*

$$(U^n, D^n, N^n, \hat{U}^n, \hat{D}^n) \Rightarrow (e, 0, \hat{N}, -\hat{J}, \hat{J}) \text{ in } (\mathbb{D}_3, J_1) \times (\mathbb{D}_2, M_1) \quad \text{as } n \to \infty,$$

where the limit process $\hat{J} = \{\hat{J}(t) : t \geq 0\}$ *is defined by* $\hat{J}(t) = \sum_{i=1}^{\hat{N}(t)} d_i$ *for* $t \geq 0$, *and* $\hat{N}(t)$ *is defined in (2).*

Proof. The proof of the convergence of (U^n, D^n, N^n) follows directly from Assumption 1 and the proof of $(\hat{U}^n, \hat{D}^n)$ can be found in [22, Section 5.4]. □

3. Service times with finite support

In this section we consider service time distributions that are i.i.d. and have a finite support. In particular, we assume that the service times $\{\eta_i : i \geq 1\}$ are i.i.d. and have a distribution F with a finite positive support $\{x_1, \ldots, x_m\}$ with associated probabilities $p_1, \ldots, p_m$ such that $\sum_{i=1}^{m} p_i = 1$. We say that a customer requiring service time x_i is a type i customer. Let $A_i^n(t)$ be the cumulative number of arrivals of type i customers up to time t in the nth system. Denote $\boldsymbol{A}^n = (A_1^n, \ldots, A_m^n)$. Let $\mathrm{disc}(x)$ be the set of discontinuity points of x in $[0, \infty)$ and $\theta_s : \mathbb{D} \to \mathbb{D}$ be the shift operator defined by $\theta_s(x)(t) = x(t+s)$ for $t + s \geq 0$ and $\theta_s(x)(t) = 0$ for $t + s < 0$ and $t \geq 0$. We make the following assumption on the diffusion-scaled arrival processes.

Assumption 2. *There exists a deterministic nondecreasing function* $\boldsymbol{\Lambda}(t) = (\Lambda_1(t), \ldots, \Lambda_m(t))$ *in* $\mathbb{D}^m$ *and a stochastic process* $\hat{\boldsymbol{A}} = (\hat{A}_1, \ldots, \hat{A}_m)$ *in* $\mathbb{D}^m$ *such that*

$$\mathbb{P}(\mathrm{disc}(\hat{A}_i) \cap \mathrm{disc}(\theta_s(\hat{A}_j))) = 0 \quad \text{for all } i, j = 1, \ldots, m,\ s \leq 0,$$

and $\hat{\boldsymbol{A}}^n(t) \Rightarrow \hat{\boldsymbol{A}}(t)$ *in* $(\mathbb{D}_m, M_1)$ *as* $n \to \infty$, *where* $\hat{\boldsymbol{A}}^n = (\hat{A}_1^n, \ldots, \hat{A}_m^n)$ *is defined by*

$$\hat{A}_i^n(t) := n^\gamma(n^{-1}A_i^n(t) - \Lambda_i(t)), \qquad t \geq 0,\ i = 1, \ldots, m.$$

The functions $\Lambda_i(t)$, $i = 1, \ldots, m$, are centering terms in the FCLT. This assumption implies that the fluid-scaled processes satisfy an FLLN [35]: $\bar{\boldsymbol{A}}^n = n^{-1}\boldsymbol{A}^n \Rightarrow \boldsymbol{\Lambda}$ in $(\mathbb{D}_m, M_1)$ as $n \to \infty$.

Let $X_{n,i}^r(t, y)$ be the associated two-parameter processes for type i customers and $X_{n,i}(t)$ be the associated total count process for type i customers. Then by (1), we have

$$X_{n,i}^r(t, y) = \begin{cases} A_i^n(t) - A_i^n(U^{n,-1}(U^n(t) - (x_i - y))), & t \geq 0,\ 0 \leq y < x_i, \\ 0, & t \geq 0,\ y \geq x_i, \end{cases} \tag{3}$$

where $U^{n,-1}(t) = \sup\{s \geq 0 : U^n(s) \leq t\}$, $t \geq 0$. We also let $U^{n,-1}(t) \equiv 0$ for $t < 0$ so that (3) is well defined. Note that $X_{n,i}(t) = A_i^n(t) - A_i^n(U^{n,-1}(U^n(t) - x_i))$ for $t \geq 0$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

We remark that the inverse process $U^{n,-1}$ has sample paths in $\mathbb{D}$ and can be represented as

$$U^{n,-1}(t) = t + \sum_{k=1}^{\hat{N}_n^u(t)} d_k^n, \qquad t \geq 0,$$

where $\hat{N}_n^u(t) = \max\{i \geq 0\colon \hat{T}_{n,i}^u \leq t\}$, $t \geq 0$, and $\hat{T}_{n,i}^u = \sum_{k=1}^i u_k^n$, $i \geq 1$, and $\hat{T}_{n,0}^u = 0$. We remark that $X_{n,i}^r(t, y)$ in (3) cannot be simply written as

$$A_i^n(t) - A_i^n(U^{n,-1}(U^n(t) - (x_i - y)^+)) \quad \text{for all } y \geq 0,$$

as in the case without interruptions. By the definition of $U^{n,-1}(t)$, the expression is correct only for $0 \leq y < x_i$. But in the FLLN and FCLT limits for $X_{n,i}^r(t, y)$ in (4) and (5), we can use the expression $t - (x_i - y)^+$ to combine the two scenarios with $0 \leq y < x_i$ and $y \geq x_i$.

We write

$$\boldsymbol{X}_n^r(t, y) = (X_{n,1}^r(t, y), \ldots, X_{n,m}^r(t, y)), \qquad X_n^r(t, y) = \sum_{i=1}^m X_{n,i}^r(t, y),$$

$$\boldsymbol{X}_n(t) = (X_{n,1}(t), \ldots, X_{n,m}(t)), \qquad X_n(t) = \sum_{i=1}^m X_{n,i}(t).$$

Now define the fluid-scaled processes $\bar{\boldsymbol{X}}_n^r(t, y) = n^{-1}\boldsymbol{X}_n^r(t, y)$, $\bar{X}_n^r(t, y) = n^{-1}X_n^r(t, y)$, $\bar{\boldsymbol{X}}_n(t) = n^{-1}\boldsymbol{X}_n(t)$, and $\bar{X}_n(t) = n^{-1}X_n(t)$. The fluid limits for these processes $\bar{\boldsymbol{X}}_n^r(t, y)$, $\bar{X}_n^r(t, y)$, $\bar{\boldsymbol{X}}_n(t)$, and $\bar{X}_n(t)$ are stated in the following theorem. We remark that it is understood that for each $i = 1, \ldots, m$, $\Lambda_i(t) \equiv 0$ whenever $t < 0$ so that the limit processes in Theorem 1 are all well defined, and similarly for other relevant processes throughout the paper.

Theorem 1. (FLLN.) *Under Assumptions 1 and 2,*

$$(\bar{\boldsymbol{X}}_n^r(t, y), \bar{X}_n^r(t, y), \bar{\boldsymbol{X}}_n(t), \bar{X}_n(t)) \Rightarrow (\boldsymbol{x}^r(t, y), x^r(t, y), \boldsymbol{x}(t), x(t)) \text{ in } ((\mathbb{D}_{\mathbb{D}})_{m+1}, M_1) \times (\mathbb{D}_{m+1}, M_1) \quad \text{as } n \to \infty,$$

where $\boldsymbol{x}^r(t, y) = (x_1^r(t, y), \ldots, x_m^r(t, y))$, $\boldsymbol{x}(t) = (x_1(t), \ldots, x_m(t))$, *and for each* $i = 1, \ldots, m$,

$$x_i^r(t, y) = \Lambda_i(t) - \Lambda_i(t - (x_i - y)^+), \qquad x_i(t) = x_i^r(t, 0) = \Lambda_i(t) - \Lambda_i(t - x_i), \tag{4}$$

$$x^r(t, y) = \sum_{i=1}^m x_i^r(t, y), \quad \text{and} \quad x(t) = \sum_{i=1}^m x_i(t).$$

Define the diffusion-scaled processes

$$\hat{X}_{n,i}^r(t, y) = n^\gamma(\bar{X}_{n,i}^r(t, y) - x_i^r(t, y)), \qquad \hat{X}_n^r(t, y) = \sum_{i=1}^m \hat{X}_{n,i}^r(t, y),$$

$$\hat{X}_{n,i}(t) = \hat{X}_{n,i}^r(t, 0) = n^\gamma(\bar{X}_{n,i}(t) - x_i(t)), \qquad \hat{X}_n(t) = \hat{X}_n^r(t, 0) = \sum_{i=1}^m \hat{X}_{n,i}(t),$$

and write

$$\hat{\boldsymbol{X}}_n^r(t, y) = (\hat{X}_{n,1}^r(t, y), \ldots, \hat{X}_{n,m}^r(t, y)) \quad \text{and} \quad \hat{\boldsymbol{X}}_n(t) = (\hat{X}_{n,1}(t), \ldots, \hat{X}_{n,m}(t)).$$

We show an FCLT for these diffusion-scaled processes.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Theorem 2. (FCLT.) *Under Assumptions 1 and 2, assuming that* $\boldsymbol{\Lambda}$ *is absolutely continuous with density* $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_m) \in \mathbb{C}^m$ *a.e.,*

$$(\hat{\boldsymbol{X}}_n^r(t,y), \hat{X}_n^r(t,y), \hat{\boldsymbol{X}}_n(t), \hat{X}_n(t)) \Rightarrow (\hat{\boldsymbol{X}}^r(t,y), \hat{X}^r(t,y), \hat{\boldsymbol{X}}(t), \hat{X}(t)) \text{ in } ((\mathbb{D}_{\mathbb{D}})_{m+1}, M_1) \times (\mathbb{D}_{m+1}, M_1) \quad \text{as } n \to \infty,$$

where $\hat{\boldsymbol{X}}^r(t,y) = (\hat{X}_1^r(t,y), \ldots, \hat{X}_m^r(t,y))$, $\hat{\boldsymbol{X}}(t) = (\hat{X}_1(t), \ldots, \hat{X}_m(t))$,

$$\hat{X}_i^r(t,y) = \hat{A}_i(t) - \hat{A}_i(t-(x_i-y)^+) + \hat{J}_i^r(t,y), \qquad \hat{X}_i(t) = \hat{X}_i^r(t,0), \tag{5}$$

$$\hat{J}_i^r(t,y) = \lambda_i(t-(x_i-y)^+) \sum_{k=\hat{N}(t-(x_i-y)^+)}^{\hat{N}(t)} d_k, \qquad \hat{J}_i(t) = \hat{J}_i^r(t,0), \tag{6}$$

$$\hat{X}^r(t,y) = \sum_{i=1}^m \hat{X}_i^r(t,y), \qquad \hat{X}(t) = \sum_{i=1}^m \hat{X}_i(t),$$

and $\hat{N}(t)$ *is defined in (2). The processes* $\hat{J}_i^r(t,y)$ *and* $\hat{J}_i(t)$ *are independent of the arrival limit processes* $\hat{A}_i(t)$ *for each* $i = 1, \ldots, m$.

We remark that when the arrival rates are constant, that is, $\Lambda_i(t) = \lambda_i t$, the fluid limits $x_i^r(t,y)$ and $x_i^r(t)$ in (4) become

$$x_i^r(t,y) = \lambda_i(t-(t-(x_i-y)^+)^+), \qquad x_i(t) = \lambda_i(t-(t-x_i)^+). \tag{7}$$

and the processes $\hat{J}_i^r(t,y)$ and $\hat{J}_i(t)$ in (6) become

$$\hat{J}_i^r(t,y) = \lambda_i \sum_{k=\hat{N}(t-(x_i-y)^+)}^{\hat{N}(t)} d_k, \qquad \hat{J}_i(t) = \lambda_i \sum_{k=\hat{N}(t-x_i)}^{\hat{N}(t)} d_k.$$

As a consequence of Theorem 2, there exists a *stochastic decomposition property* for the limit process $\hat{\boldsymbol{X}}^r(t,y)$, that is, each of the processes $\hat{X}_i^r(t,y)$ is equal to a sum of two independent processes, the arrival limit processes $\hat{A}_i(t) - \hat{A}_i(t-(x_i-y)^+)$ and the jump processes due to service interruptions $\hat{J}_i^r(t,y)$. This stochastic decomposition property also holds for $\hat{X}^r(t,y)$, $\hat{\boldsymbol{X}}(t)$, and $\hat{X}(t)$. This can be intuitively explained as follows. Since the service times are deterministic, the realized service times with the interruptions are in fact the deterministic duration 'inserted' with the random down times from the arrival time to service completion. The randomness from arrivals is captured in $\hat{A}_i(t) - \hat{A}_i(t-(x_i-y)^+)$ as in [11], while the randomness from interruptions is captured in $\hat{J}_i^r(t,y)$, which is exactly the number of down times (in the diffusion scale) during the time period to be counted in $\hat{X}_i^r(t,y)$. This is also demonstrated in the covariance formulas below.

Corollary 1. *In addition to the assumptions of Theorem 2, suppose that*

(i) *the arrival rates are constant and the arrival limit process* $\hat{\boldsymbol{A}}$ *is a Brownian motion with mean* $\mathbf{0}$ *and covariance coefficient matrix* $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,m}$, *where* $\sigma_{ii} = \lambda_i c_{a,i}^2$ *for some constant* $c_{a,i}^2 > 0$, *and* $\sigma_{ij} \geq 0$ *are some positive constants for all* $i \neq j$, *and*

(ii) *the limit counting process* $\hat{N}$ *is a Poisson process with rate* $\lambda^u = 1/\mathbb{E}[u_i] \in (0, \infty)$ *and the limit up times* u_i *have a finite second moment.*

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Then for each $t > 0$ and $y \geq 0$, the distribution of $\hat{\boldsymbol{X}}^r(t, y)$ is a sum of two independent random vectors, a multivariate Gaussian random vector $N(0, \Sigma^r(t, y))$, where $\Sigma^r_{ii}(t, y) = c^2_{a,i} x^r_i(t, y)$ and $\Sigma^r_{ij}(t, y) = \sigma_{ij}(x^r_i(t, y) \wedge x^r_j(t, y))$ with $x^r_i(t, y)$ being defined in (7), and a random vector of compound Poisson variables, $\hat{\boldsymbol{J}}^r = (\hat{J}^r_1, \ldots, \hat{J}^r_m)$, which has mean $\mathbb{E}[\hat{J}^r_i(t, y)] = \lambda^u \mathbb{E}[d_1] x^r_i(t, y)$, $i = 1, \ldots, m$, and covariances: for $i, j = 1, \ldots, m$,

$$\begin{aligned}&\operatorname{cov}(\hat{J}^r_i(t, y),\ \hat{J}^r_j(t, y))\\ &\quad = \lambda_i \lambda_j (\lambda^u((x_i - y)^+ \wedge (x_j - y)^+)\mathbb{E}[d_1^2] - (\lambda^u)^2((x_i - y)^+ \wedge (x_j - y)^+)^2(\mathbb{E}[d_1])^2).\end{aligned}$$

4. Dependent service times with continuous distributions

In this section we consider G/G/∞ systems with stationary and weakly dependent service times and a general continuous distribution function. Throughout this section, we assume that $\gamma = \frac{1}{2}$ (see also Remark 2). We assume that the arrival processes A^n satisfy an FCLT.

Assumption 3. *There exists a deterministic function $\Lambda \in \mathbb{C}$ with density $\lambda \in \mathbb{D}$ and a stochastic process $\hat{A}$ such that $\hat{A}^n(t) := \sqrt{n}(n^{-1}A^n(t) - \Lambda(t)) \Rightarrow \hat{A}(t)$ in $(\mathbb{D}, M_1)$ as $n \to \infty$.*

The service times are fixed with respect to n and so is their distribution function.

Assumption 4. *The sequence of service times $\{\eta_i : i \geq 1\}$ is weakly dependent and constitutes a one-sided stationary sequence. All the η_i have the same continuous cumulative distribution function F with $F(0) = 0$, and $\mathbb{E}[\eta_1^2] < \infty$, satisfying*

$$\sum_{i=1}^{\infty} (\mathbb{E}[(\mathbb{E}[\eta_{i+k} \mid \mathcal{F}^s_k])^2])^{1/2} < \infty \quad \text{for } k = 1, 2, \ldots,$$

where $\mathcal{F}^s_k = \sigma\{\eta_i : 1 \leq i \leq k\}$. Moreover, the sequence $\{\eta_i : i \geq 1\}$ satisfies the ϕ-mixing condition, that is, $\sum_{k=1}^{\infty} \phi_k < \infty$, where

$$\phi_k = \sup\{|\mathbb{P}(B \mid A) - \mathbb{P}(B)| : A \text{ in } \mathcal{F}^s_m,\ \mathbb{P}(A) > 0,\ B \in \mathcal{G}^s_{m+k},\ m \geq 1\}$$

with $\mathcal{G}^s_k = \sigma\{\eta_i : i \geq k\}$. Let $F^c = 1 - F$.

Note that Assumption 4 includes many interesting examples of correlated service times as studied in Pang and Whitt [26], [27], [28]. For example, customers arrive in batches (deterministic or random batch sizes) and service times within each batch are (symmetrically) correlated while service times across different batches are independent. Other examples include EARMA (exponential autoregressive moving average) sequences and first-order autoregressive sequences with exponential or general marginal distributions [12]–[14], [31], which have been used to study correlated service times in queueing; see Remark 3.

We first state the following FLLN for the fluid-scaled processes $\bar{X}^r_n(t, y) = n^{-1}X^r_n(t, y)$ and $\bar{X}_n(t) = n^{-1}X_n(t)$. The proof follows directly from Theorem 4.

Theorem 3. *Under Assumptions 1, 3, and 4,*

$$(\bar{X}^r_n(t, y), \bar{X}_n(t)) \Rightarrow (\tilde{X}^r(t, y), \tilde{X}(t)) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \times (\mathbb{D}, M_1) \quad \text{as } n \to \infty,$$

where $\tilde{X}^r(t, y) = \int_0^t F^c(t - s + y)\, \mathrm{d}\Lambda(s)$ and $\tilde{X}(t) = \tilde{X}^r(t, 0)$ for $t, y \geq 0$.

Define the diffusion-scaled processes $\hat{X}^r_n(t, y) = \sqrt{n}(\bar{X}^r_n(t, y) - \tilde{X}^r(t, y))$ and $\hat{X}_n(t) = \sqrt{n}(\bar{X}_n(t) - \tilde{X}(t))$. We next state the FCLT for the processes $\hat{X}^r_n(t, y)$ and $\hat{X}_n(t)$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Theorem 4. *Under Assumptions 1, 3, and 4,*

$$(\hat{X}_n^r(t,y), \hat{X}_n(t)) \Rightarrow (\hat{X}^r(t,y), \hat{X}(t)) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \times (\mathbb{D}, M_1) \quad \text{as } n \to \infty,$$

where

$$\hat{X}^r(t,y) = \hat{X}_1^r(t,y) + \hat{X}_2^r(t,y) + \hat{X}_3^r(t,y), \qquad \hat{X}(t) = \hat{X}_1(t) + \hat{X}_2(t) + \hat{X}_3(t),$$

$$\hat{X}_1^r(t,y) = \int_0^t F^c(t-s+y)\,\mathrm{d}\hat{A}(s), \qquad \hat{X}_2^r(t,y) = \int_0^t \int_0^\infty \mathbf{1}_{\{x>t-s+y\}}\,\mathrm{d}\hat{K}(\Lambda(s),x),$$

$$\hat{X}_3^r(t,y) = \int_0^t (\hat{J}(t) - \hat{J}(s))\lambda(s)\,\mathrm{d}F^c(t-s+y),$$

$$\hat{X}_1(t) = \hat{X}_1^r(t,0), \qquad \hat{X}_2(t) = \hat{X}_2^r(t,0), \qquad \hat{X}_3(t) = \hat{X}_3^r(t,0),$$

the process $\hat{J}(t)$ is defined in Lemma 1, the process $\hat{K}(t,x)$ is a two-parameter continuous Gaussian process with mean 0 and covariance function $\mathrm{cov}(\hat{K}(t,x), \hat{K}(s,y)) = (t \wedge s)\Gamma_K(x,y)$ *with* $\Gamma_K(x,y) = (F(x \wedge y) - F(x)F(y)) + \Gamma_K^c(x,y) < \infty$ *and*

$$\Gamma_K^c(x,y) = \sum_{k=2}^{\infty} (\mathbb{E}[\bar{\gamma}_1(x)\bar{\gamma}_k(y)] + \mathbb{E}[\bar{\gamma}_1(y)\bar{\gamma}_k(x)]) \quad \text{for } x, y \geq 0$$

and with $\bar{\gamma}_k(x) := \mathbf{1}_{\{\eta_k \leq x\}} - F(x)$ for $x \geq 0$ and $k \geq 1$, and the integral in $\hat{X}_2^r(t,y)$ with respect to the process $\hat{K}$ is defined in the sense of mean-square integrals. The three processes $\hat{X}_i^r(t,y), i = 1,2,3$, are mutually independent, and so are the three processes $\hat{X}_i(t), i = 1,2,3$.

When the service times are i.i.d., the covariance function of $\hat{K}(t,x)$ in Theorem 4 becomes $\mathrm{cov}(\hat{K}(t,x), \hat{K}(s,y)) = (t \wedge s)(F(x \wedge y) - F(x)F(y))$, $t, s, x, y \geq 0$, and the process $\hat{K}(t,x)$ is called a Kiefer process [18], [25]. When $\hat{N}(t)$ is Poisson, the jump process $\hat{J}(t)$ in Lemma 1 is a compound Poisson process. If, in addition, the arrival rate is constant, by the stationarity of $\hat{J}(t)$ and integration by parts, we can represent the process $\hat{X}_3^r(t,y)$ in Theorem 4 as

$$\hat{X}_3^r(t,y) = -\lambda F^c(t+y)\hat{\mathcal{J}}([0,t] \times \mathbb{R}_+) + \lambda \int_0^t \int_0^\infty x F^c(s+y)\hat{\mathcal{J}}(\mathrm{d}s, \mathrm{d}x),$$

where $\hat{\mathcal{J}}(s,x)$ is a Poisson random measure defined on $[0,\infty) \times \mathbb{R}_+$ with intensity $\lambda^u\,\mathrm{d}s \times \mathrm{d}G(x)$, where $\lambda^u = 1/\mathbb{E}[u_i] \in (0,\infty)$ and $G(\cdot)$ is the distribution function of the limiting down times $\{d_k : k \geq 1\}$.

Remark 1. There is a *stochastic decomposition property* for limit process $\hat{X}^r(t,y)$: the variability from the arrival process is captured in $\hat{X}_1^r(t,y)$, the variability and correlation of service times are captured in $\hat{X}_2^r(t,y)$, and the impact of interruptions is captured in $\hat{X}_3^r(t,y)$.

Remark 2. In Theorems 3 and 4, we have assumed that $\gamma = \frac{1}{2}$. This is the only scaling when a proper limit can be established to capture the variabilities from service times, the processes $\hat{X}_2^r(t,y)$ and $\hat{X}_2(t)$ in Theorem 4. This is because the FCLT for the sequential empirical processes driven by the service times requires $\sqrt{n}$ scaling, see Lemma 7. When $\gamma \in (0, \frac{1}{2})$, no limit exists for $(\hat{X}_n^r(t,y), \hat{X}_n(t))$, while when $\gamma > \frac{1}{2}$, an FCLT can be established for $(\hat{X}_n^r(t,y), \hat{X}_n(t))$, which will have the same limits in Theorem 4 with $\hat{X}_2^r(t,y) = \hat{X}_2(t) \equiv 0$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

When the arrival limit process is a Brownian motion and the limiting counting process $\hat{N}(t)$ is Poisson, we can characterize the transient and stationary distribution of the limit processes as in the following corollary. Its proof follows from direct calculations and is thus omitted.

Corollary 2. *Under the assumptions of Theorem 4, if, in addition, $\hat{A}(t) = c_a B(\Lambda(t))$ for some $c_a > 0$, $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ and a standard Brownian motion $B(t)$, and the limit counting process $\hat{N}(t)$ is a Poisson process with rate $\lambda^u = 1/\mathbb{E}[u_i] \in (0, \infty)$ and the limit up times u_i have a finite second moment, the processes $\hat{X}^r(t, y)$ and $\hat{X}(t)$ have mean*

$$\mathbb{E}[\hat{X}^r(t, y)] = \lambda^u \mathbb{E}[d_1] \int_0^t [(t-s)\lambda(s)]\,\mathrm{d}F^c(t-s+y), \qquad \mathbb{E}[\hat{X}^r(t)] = \mathbb{E}[\hat{X}^r(t, 0)],$$

and variance functions

$$\begin{aligned}\mathrm{var}(\hat{X}^r(t, y)) = {} & \int_0^t \lambda(s)(F^c(t+y-s) \\ & + (c_a^2 - 1)((F^c(t+y-s))^2 + \Gamma_K^c(t+y-s, t+y-s)))\,\mathrm{d}s \\ & + \lambda^u \mathbb{E}[d_1^2] \int_0^t \int_0^t (t - s \vee u)\lambda(s)\lambda(u)\,\mathrm{d}F^c(t-s+y)\,\mathrm{d}F^c(t-u+y),\end{aligned}$$

and, $\mathrm{var}(\hat{X}(t)) = \mathrm{var}(\hat{X}^r(t, 0))$. *When the arrival rate is constant, we obtain*

$$\begin{aligned}\mathbb{E}[\hat{X}^r(\infty, y)] &= \lambda\lambda^u \mathbb{E}[d_1] \int_y^\infty (s-y)\,\mathrm{d}F^c(s), \\ \mathbb{E}[\hat{X}(\infty)] &= \mathbb{E}[\hat{X}^r(\infty, 0)] = \lambda \mathbb{E}[\eta_1]\lambda^u \mathbb{E}[d_1], \\ \mathrm{var}(\hat{X}^r(\infty, y)) &= \lambda \int_y^\infty (F^c(s) + (c_a^2 - 1)((F^c(s))^2 + \Gamma_K^c(s, s)))\,\mathrm{d}s \\ &\quad + \lambda^2\lambda^u \mathbb{E}[d_1^2] \int_y^\infty \int_y^\infty (s \wedge u - y)\,\mathrm{d}F^c(s)\,\mathrm{d}F^c(u),\end{aligned}$$

and $\mathrm{var}(\hat{X}(\infty)) = \mathrm{var}(\hat{X}^r(\infty, 0))$.

Remark 3. The impact of correlation among service times was characterized in [26] and [27]. In particular, the effects of the terms Γ_k^c in $\mathrm{var}(\hat{X}^r(t, y))$ upon the variance of the queue in the infinite-server models and upon the delay in the finite-server models were carefully studied. For example, consider the case of the sequence of service times forming a first-order discrete autoregressive process, DAR(1), that is, $\eta_i = \zeta_{i-1}\eta_{i-1} + (1 - \zeta_{i-1})\tilde{\eta}_i$, $i \geq 2$, where $\{\zeta_i : i \geq 1\}$ is a sequence of i.i.d. Bernoulli random variables with $\mathbb{P}(\zeta_i = 1) = p \in (0, 1)$, and $\{\tilde{\eta}_i : i \geq 2\}$ is a sequence of i.i.d. random variables with distribution F. Then, as in [27], the term $\Gamma_K^c(s, s)$ in $\mathrm{var}(\hat{X}^r(t, y))$ is equal to $(2p/(1-p))F(s)F^c(s)$. Our contribution lies in the additional variabilities caused by service interruptions, characterized by the last terms in $\mathrm{var}(\hat{X}^r(t, y))$ and $\mathrm{var}(\hat{X}^r(\infty, y))$, which depend on the second moment of the limiting downtimes.

5. Proofs for service times with finite support

In this section we prove Theorems 1 and 2, when the service times are i.i.d. and have a finite support. We will need Lemmas 3–6 below, whose proofs are given in Appendix A.

Lemma 2. *Let $(\mathcal{S}, r)$ be a Banach space, and $x_n, x \in \mathbb{D}([0, T], \mathcal{S})$ and $x \in \mathbb{C}([0, T], \mathcal{S})$ for $n \geq 1$ and $T > 0$. The following is a necessary and sufficient condition for $x_n \to x$ in the*

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

space $\mathbb{D}([0,T],\mathcal{S})$ *endowed with the Skorokhod* M_1 *topology: whenever* $t_n \to t$ *as* $n \to \infty$ *for* $t_n, t \in [0,T]$, $r(x_n(t_n), x(t)) \to 0$ *as* $n \to \infty$.

Proof. See, e.g. [7, Section 3.6] and [33, Lemma 2.1]. □

Lemma 3. *Suppose that* $(x_n(t), y_n(t,s)) \to (x(t), y(t,s))$ *in* $(\mathbb{D}, M_1) \times (\mathbb{D}_{\mathbb{D}}, J_1)$ *as* $n \to \infty$ *and* $y(t,s) \in \mathbb{C}_{\mathbb{C}}$. *Define* $z_n(t,s) := x_n(y_n(t,s))$ *and* $z(t,s) := x(y(t,s))$, *and denote* $z_n(t) := z_n(t,\cdot)$ *and* $z(t) := z(t,\cdot)$. *Suppose that* $y_n(t,s)$ *and* $y(t,s)$ *are nondecreasing in* t *and strictly increasing in* s. *Then* $z_n(t,s)$ *and* $z(t,s)$ *are both in* $\mathbb{D}_{\mathbb{D}}$, $z_n(t)$ *and* $z(t)$ *are both continuous in* t *in* $\mathbb{D}([0,\infty),(\mathbb{D},L_1))$, *and* $z_n \to z$ *in* $(\mathbb{D}_{\mathbb{D}}, M_1)$ *as* $n \to \infty$.

Lemma 4. *If* $x_n(t) \to x(t)$ *in* $(\mathbb{D}, M_1)$ *as* $n \to \infty$ *then* $x_n(t) \to x(t)$ *in* $(\mathbb{D}_{\mathbb{D}}, M_1)$ *as* $n \to \infty$.

Lemma 5. *Suppose that* $(x^n(t), y^n(t,s)) \to (x(t), y(t,s))$ *in* $(\mathbb{D}, M_1) \times (\mathbb{D}_{\mathbb{D}}, M_1)$ *as* $n \to \infty$, *and* $x(t)$ *and* $y(t) := y(t,\cdot)$ *do not have common discontinuity points. Then* $x^n(t) + y^n(t,s) \to x(t) + y(t,s)$ *in* $(\mathbb{D}_{\mathbb{D}}, M_1)$ *as* $n \to \infty$.

Proof of Theorem 1. First, by (3), for $i = 1, \dots, m$, we can write $\bar{X}^r_{n,i}(t,y)$ as, for $t \geq 0$ and $0 \leq y < x_i$,

$$\bar{X}^r_{n,i}(t,y) = \bar{A}^n_i(t) - \bar{A}^n_i\left(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))} d^n_k\right), \tag{8}$$

and for $t \geq 0$ and $y \geq x_i$, $\bar{X}^r_{n,i}(t,y) = 0$. Thus, the main focus is on proving the convergence of $\bar{X}^r_{n,i}(t,y)$ for $t \geq 0$ and $0 \leq y < x_i$, $i = 1, \dots, m$. In the remainder of this proof, the space $\mathbb{D}_{\mathbb{D}}$ is restricted to functions $z(t,y)$ with $t \in [0,\infty)$ and $y \in [0, x_i)$ for each $i = 1, \dots, m$. By Assumption 1, we have $\hat{N}^u_n \Rightarrow \hat{N}$ in $(\mathbb{D}, J_1)$ as $n \to \infty$, where $\hat{N}$ is defined in (2). By Lemma 1, this implies that

$$\begin{aligned}&(U^n(t) - (x_i - y),\ \hat{N}^u_n(U^n(t) - (x_i - y)),\ i = 1, \dots, m)\\ &\quad \Rightarrow (t - (x_i - y),\ \hat{N}(t - (x_i - y)),\ i = 1, \dots, m) \text{ in } ((\mathbb{D}_{\mathbb{D}})_{2m}, J_1) \quad \text{as } n \to \infty.\end{aligned}$$

Now by Assumption 1,

$$\left(\sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))} d^n_k,\ i = 1, \dots, m\right) \Rightarrow (0,\ i = 1, \dots, m) \text{ in } ((\mathbb{D}_{\mathbb{D}})_m, J_1) \quad \text{as } n \to \infty.$$

Thus, by the continuous mapping theorem applied to the addition mapping in $(\mathbb{D}_{\mathbb{D}}, J_1)$, we have

$$\begin{aligned}&\left(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))} d^n_k,\ i = 1, \dots, m\right)\\ &\quad \Rightarrow (t - (x_i - y),\ i = 1, \dots, m) \text{ in } ((\mathbb{D}_{\mathbb{D}})_m, J_1) \quad \text{as } n \to \infty.\end{aligned}$$

By Assumption 2 on the arrival processes and Lemma 3, we obtain

$$\begin{aligned}&\left(\bar{A}^n_i\left(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))} d^n_k\right),\ i = 1, \dots, m\right)\\ &\quad \Rightarrow (\Lambda_i(t - (x_i - y)),\ i = 1, \dots, m) \text{ in } ((\mathbb{D}_{\mathbb{D}})_m, M_1) \quad \text{as } n \to \infty.\end{aligned}$$

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

To apply Lemma 3, it is easy to see that the sample paths of the processes

$$U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U^n(t)-(x_i-y))} d_k^n$$

and the functions $t - (x_i - y)^+$ are nondecreasing in t and strictly increasing in y for $y \in [0, x_i)$ for each $i = 1, \ldots, m$.

Now, by applying the continuous mapping theorem, Lemmas 4 and 5, and by combining the convergence of $\bar{X}_{n,i}^r(t, y)$ for $t \geq 0$ and $y \geq x_i$, we obtain the convergence of $\bar{\boldsymbol{X}}_n(t, y)$. The convergence of $\bar{\boldsymbol{X}}_n(t)$, and, thus, that of $\bar{X}_n(t)$, follow directly from the relationship between the total count process and the two-parameter process. It can also be proved directly by applying the continuous mapping theorem to the composition and addition mappings in the space $(\mathbb{D}, M_1)$. This completes the proof. □

We then prove Theorem 2. For the proof, we need the following lemma.

Lemma 6. *Suppose that $x(t, s) \in \mathbb{D}_{\mathbb{D}}$ as a function of t is continuous in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$ and $y_n(t, s) \to y(t, s)$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$, and $y(t, s)$ as a function of t is continuous in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$. Then $x(t, s)y_n(t, s) \to x(t, s)y(t, s)$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$ as $n \to \infty$.*

Proof of Theorem 2. First, by (8), for each $i = 1, \ldots, m$, we can write the processes $\hat{X}_{n,i}^r(t, y)$ as, for $t \geq 0$ and $0 \leq y < x_i$,

$$\hat{X}_{n,i}^r(t, y) = \hat{A}_i^n(t) - \hat{A}_i^n\left(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U^n(t)-(x_i-y))} d_k^n\right) + \hat{J}_{n,i}^r(t, y),$$

where $\hat{J}_{n,i}^r(t, y) = n^\gamma(\Lambda_i(t - (x_i - y)) - \Lambda_i(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U^n(t)-(x_i-y))} d_k^n))$, and for $t \geq 0$ and $y \geq x_i$, $\hat{X}_{n,i}^r(t, y) = 0$. As in the proof of Theorem 1, we need focus only on proving the convergence of $\hat{X}_{n,i}^r(t, y)$ for $t \geq 0$ and $0 \leq y < x_i$, $i = 1, \ldots, m$, and in the remainder of the proof, the space $\mathbb{D}_{\mathbb{D}}$ is restricted to functions $z(t, y)$ with $t \in [0, \infty)$ and $y \in [0, x_i)$ for each $i = 1, \ldots, m$.

By Assumption 2, the convergence of $U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U_n(t)-(x_i-y))} d_k^n$ in the proof of Theorem 1, and Lemma 3, we obtain

$$\left(\hat{A}_i^n\left(U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U_n(t)-(x_i-y))} d_k^n\right),\ i = 1, \ldots, m\right)$$
$$\Rightarrow (\hat{A}_i(t - (x_i - y)),\ i = 1, \ldots, m) \quad \text{in } ((\mathbb{D}_{\mathbb{D}})_m, M_1) \quad \text{as } n \to \infty.$$

Again, note that the sample paths of the processes $U^n(t) - (x_i - y) + \sum_{k=1}^{\hat{N}_n^u(U^n(t)-(x_i-y))} d_k^n$ and the functions $t - (x_i - y)$ are continuous as functions of t in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$, and are nondecreasing in t and strictly increasing in y for $y \in [0, x_i)$, for $i = 1, \ldots, m$.

Next, we focus on the processes $\hat{J}_{n,i}^r(t, y)$. We note that the continuity assumption of Λ_i is necessary because if $t - (x_i - y)^+$ is a discontinuity point of Λ_i, the processes $\hat{J}_{n,i}^r(t, y)$ will converge to ∞ at those discontinuity points. In the expression of $\hat{J}_{n,i}^r(t, y)$, we apply a Taylor expansion and obtain

$$\hat{J}_{n,i}^r(t, y) = \lambda_i(t - (x_i - y))\hat{Z}_{n,i}^r(t, y) + o\left(\frac{1}{n^\gamma}\right),$$

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

where $\hat{Z}^r_{n,i}(t,y) = \hat{D}^n(t) - \sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))}(n^\gamma d^n_k)$. By Assumption 1, the convergence of $\hat{N}^u_n$ and Lemma 1, it follows that

$$\left(\sum_{k=1}^{\hat{N}^u_n(U^n(t)-(x_i-y))}(n^\gamma d^n_k),\ i=1,\dots,m\right)$$
$$\Rightarrow \left(\sum_{k=1}^{\hat{N}(t-(x_i-y))} d_k,\ i=1,\dots,m\right) \quad \text{in } ((\mathbb{D}_{\mathbb{D}})_m, J_1) \quad \text{as } n\to\infty.$$

Thus, by Assumption 1, the convergence of $\hat{D}^n(t)$ in Lemma 1, and by Lemma 5, we obtain

$$(\hat{Z}^r_{n,i}(t,y),\ i=1,\dots,m)$$
$$\Rightarrow \left(\sum_{k=\hat{N}(t-(x_i-y))}^{\hat{N}(t)} d_k,\ i=1,\dots,m\right) \quad \text{in } ((\mathbb{D}_{\mathbb{D}})_m, M_1) \quad \text{as } n\to\infty.$$

We denote $\hat{\boldsymbol{J}}^r_n = (\hat{J}^r_{n,1},\dots,\hat{J}^r_{n,m})$. Recall the associated processes $\hat{\boldsymbol{J}}^r = (\hat{J}^r_1,\dots,\hat{J}^r_m)$, where $\hat{J}^r_i$ is defined in (6). By Lemma 6, we obtain $\hat{\boldsymbol{J}}^r_n \Rightarrow \hat{\boldsymbol{J}}^r$ in $((\mathbb{D}_{\mathbb{D}})_m, M_1)$ as $n\to\infty$.

Finally, by the fact that the limiting processes $\hat{A}_i$ and $\hat{J}^r_i$ do not have simultaneous jumps, by the continuity of summation in Lemma 5, and by combining the convergence of $\hat{X}^r_{n,i}(t,y)$ for $t\ge 0$ and $y\ge x_i$, $i=1,\dots,m$, we obtain the convergence of $\hat{X}^r_n(t,y)$. The convergence of $\hat{\boldsymbol{X}}_n(t)$, and, thus, that of $\hat{X}_n(t)$, follow directly from the relationship between the total count process and the two-parameter processes. In fact, a direct proof for the convergence of $\hat{\boldsymbol{X}}_n(t)$ and $\hat{X}_n(t)$ can be performed by simply applying the continuous mapping theorem of composition, multiplication, and summation in the $(\mathbb{D}, M_1)$ topology. The details are omitted for brevity. This completes the proof. □

6. Proofs for dependent service times with continuous distributions

In this section we prove Theorem 4. As in [18], [25], and [28], we can represent the processes $\hat{X}^r_n(t,y)$ as Lebesgue–Stieltjes integrals with respect to the sequential empirical process driven by the sequence of service times. Let $\hat{K}_n(t,x)$ be a sequential empirical process driven by the service times $\{\eta_i : i\ge 1\}$, defined by

$$\hat{K}_n(t,x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor}(\mathbf{1}_{\{\eta_i\le x\}} - F(x)), \qquad t\ge 0,\ x\ge 0.$$

We now state an FCLT for the processes $\hat{K}_n(t,x)$, whose proof can be found in [28, Section 4.1 for Theorem 2.1].

Lemma 7. *Under Assumption 4, $\hat{K}_n(t,x) \Rightarrow \hat{K}(t,x)$ in $(\mathbb{D}_{\mathbb{D}}, J_1)$ as $n\to\infty$, where $\hat{K}$ is the two-parameter continuous Gaussian process defined in Theorem 4.*

Lemma 8. *The processes $\hat{X}^r_n(t,y)$ can be represented as*

$$\hat{X}^r_n(t,y) = \hat{X}^r_{n,1}(t,y) + \hat{X}^r_{n,2}(t,y) + \hat{X}^r_{n,3}(t,y),$$

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

where

$$\hat{X}^r_{n,1}(t,y) = \int_0^t F^c(U^n(t) - U^n(s) + y)\,\mathrm{d}\hat{A}^n(s),$$
$$\hat{X}^r_{n,2}(t,y) = \int_0^t \int_0^\infty \mathbf{1}_{\{x > U^n(t) - U^n(s) + y\}}\,\mathrm{d}\hat{K}_n(\bar{A}^n(s), x),$$
$$\hat{X}^r_{n,3}(t,y) = \int_0^t \sqrt{n}(F^c(U^n(t) - U^n(s) + y) - F^c(t - s + y))\,\mathrm{d}\Lambda(s).$$

The integrals in $\hat{X}^r_{n,1}(t,y)$ and $\hat{X}^r_{n,2}(t,y)$ with respect to $\hat{A}^n$ and $\hat{K}_n$, respectively, are all defined as Stieltjes integrals for functions of bounded variation as integrators.

Proof. The proof of this lemma follows from a direct calculation and the observation that the processes $X^r_n(t,y)$ defined in (1) can be written as

$$X^r_n(t,y) = \int_0^t \int_0^\infty \mathbf{1}_{\{x > U^n(t) - U^n(s) + y\}}\,\mathrm{d}\left(\sum_{i=1}^{A^n(t)} \mathbf{1}_{\{\eta_i \le x\}}\right).$$

We omit the details of this proof for brevity. □

We first show the convergence of $\hat{X}^r_{n,1}(t,y)$. We need the following lemma, whose proof can be found in Appendix A.

Lemma 9. *For $x \in \mathbb{C}$ and $z \in \mathbb{D}$, define a mapping $\phi\colon \mathbb{C} \times \mathbb{D} \to \mathbb{D}_{\mathbb{D}}$ by*

$$\phi(x,z)(t,y) = z(t)F^c(y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s) + y), \qquad t, y \ge 0.$$

Suppose that $(x_n, z_n) \to (x,z)$ in $(\mathbb{C}, \|\cdot\|) \times (\mathbb{D}, M_1)$ as $n \to \infty$, where $x_n, x \in \mathbb{C}$. Then $\phi(x_n, z_n) \to \phi(x,z)$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$ as $n \to \infty$.

We are ready to prove the convergence of $\hat{X}^r_{n,1}(t,y)$ and $\hat{X}_{n,1}(t)$.

Lemma 10. *Under Assumptions 1, 3, and 4,*

$$(\hat{X}^r_{n,1}, \hat{X}_{n,1}) \Rightarrow (\hat{X}^r_1, \hat{X}_1) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \times (\mathbb{D}, M_1) \quad \text{as } n \to \infty.$$

Proof. First, using integration by parts, we have

$$\hat{X}^r_{n,1}(t,y) = \hat{A}^n(t)F^c(y) - \int_0^t \hat{A}^n(s)\,\mathrm{d}F^c(U^n(t) - U^n(s) + y),$$
$$\hat{X}^r_1(t,y) = \hat{A}(t)F^c(y) - \int_0^t \hat{A}(s)\,\mathrm{d}F^c(t - s + y).$$

It is clear that $\hat{X}^r_{n,1}(t,y) = \phi(U^n, \hat{A}^n)(t,y)$ and $\hat{X}^r_1(t,y) = \phi(e, \hat{A})(t,y)$, where $e(t) \equiv t$ is the identity mapping. Recall that U^n has continuous nondecreasing sample paths almost surely. Now by the continuity of summation in $(\mathbb{D}_{\mathbb{D}}, M_1)$, Lemmas 4 and 9, and Assumption 4, we obtain the convergence of $\hat{X}^r_{n,1}(t,y)$. The convergence of $\hat{X}_{n,1}(t)$ follows from that of $\hat{X}^r_{n,1}(t,y)$. It can also be proved directly by showing the continuity in the M_1 topology of the mapping $\psi\colon \mathbb{C}_\uparrow \times \mathbb{D} \to \mathbb{D}$, defined by $\psi(x,z)(t) = z(t)F^c(0) - \int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s))$, $t \ge 0$. It is easy to check that the claim holds. Thus, the proof is complete. □

We next prove the convergence of $\hat{X}^r_{n,3}(t,y)$ and $\hat{X}_{n,3}(t)$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Lemma 11. *Under Assumptions 1, 3, and 4,*

$$(\hat{X}^r_{n,3}, \hat{X}_{n,3}) \Rightarrow (\hat{X}^r_3, \hat{X}_3) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \times (\mathbb{D}, M_1) \quad \text{as } n \to \infty.$$

Proof. First, by a Taylor expansion, we have

$$X^r_{n,3}(t, y) = \hat{D}^n(t) \int_0^t \lambda(s)\, \mathrm{d}F^c(t - s + y) - \int_0^t \hat{D}^n(s)\lambda(s)\, \mathrm{d}F^c(t - s + y) + o\left(\frac{1}{\sqrt{n}}\right). \quad (9)$$

For the first term on the right-hand side of (9), by Lemma 1 we have $\hat{D}^n \Rightarrow \hat{J}$ in $(\mathbb{D}, M_1)$, and by a similar argument as in the proof of Lemma 9, the function $\int_0^t \lambda(s)\, \mathrm{d}F^c(t - s + y)$ as a function of t is continuous in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$. Thus, by Lemmas 4 and 6, it follows that

$$\hat{D}^n(t) \int_0^t \lambda(s)\, \mathrm{d}F^c(t - s + y) \Rightarrow \hat{J}(t) \int_0^t \lambda(s)\, \mathrm{d}F^c(t - s + y) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \quad \text{as } n \to \infty.$$

For the second term on the right-hand side of (9), we can apply Lemmas 1 and 9. To see this, we let $z_n(s)$ and $z(s)$ in Lemma 9 be the sample paths of $\hat{D}^n(s)\lambda(s)$ and $\hat{J}(s)\lambda(s)$, respectively, and $x_n(s)$ and $x(s)$ be $e(s) \equiv s$. Thus, by the continuity of summation in $(\mathbb{D}_{\mathbb{D}}, M_1)$, we obtain $\hat{X}^r_{n,3} \Rightarrow \hat{X}^r_3$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$.

The convergence of $\hat{X}_{n,3}(t)$ follows directly from that of $\hat{X}^r_{n,3}(t, y)$. It can also be shown directly by proving the continuity in the M_1 topology of the mapping $\tilde{\psi}: \mathbb{D} \to \mathbb{D}$, where $\tilde{\psi}(x)(t) = x(t)\int_0^t \lambda(s)\, \mathrm{d}F^c(t - s) - \int_0^t x(s)\lambda(s)\, \mathrm{d}F^c(t - s)$, $t \geq 0$. It is easy to check that the claim holds. Thus, the proof is complete. □

We now focus on the proof of the convergence of $\hat{X}^r_{n,2}(t, y)$ and $\hat{X}_{n,2}(t)$. It turns out that under the assumptions of Theorem 4, we can prove their convergence in the stronger Skorokhod J_1 topology, which implies the convergence in the M_1 topology.

Lemma 12. *Under Assumptions 1, 3, and 4,*

$$(\hat{X}^r_{n,2}, \hat{X}_{n,2}) \Rightarrow (\hat{X}^r_2, \hat{X}_2) \text{ in } (\mathbb{D}_{\mathbb{D}}, J_1) \times (\mathbb{D}, J_1) \quad \text{as } n \to \infty.$$

Proof. Define the process

$$\hat{Y}^r_{n,2}(t, y) = \int_0^t \int_0^\infty \mathbf{1}_{\{x > t - s + y\}}\, \mathrm{d}\hat{K}_n(\bar{A}^n(s), x), \qquad t, y \geq 0.$$

By [28, Theorem 3.2] (note that the convergence therein is directly proved for the associated two-parameter process $X^e(t, y)$ tracking the number of customers in the system at time t that have received a amount of service less than or equal to t, but the convergence of the two-parameter processes $X^r(t, y)$ is implied and can be easily obtained as in [25] by exploiting the relationship between the two-parameter processes tracking elapsed and residual times), we have $\hat{Y}^r_{n,2}(t, y) \Rightarrow \hat{X}^r_2(t, y)$ in $(\mathbb{D}_{\mathbb{D}}, J_1)$ as $n \to \infty$, where $\hat{X}^r_2(t, y)$ is defined in Theorem 4. Thus, by [3, Theorem 3.1], it suffices to show that for each $\varepsilon > 0$ and for each $T > 0$ and $T' > 0$,

$$\lim_{n \to \infty} \mathbb{P}\Big(\sup_{t \in [0,T],\, y \in [0,T']} |\hat{Y}^r_{n,2}(t, y) - \hat{X}^r_{n,2}(t, y)| > \varepsilon \Big) = 0. \quad (10)$$

We provide a sketch proof of (10) here. The tightness of $\hat{X}^r_{n,2}$ follows first a straightforward extension of that for $\hat{Y}^r_{n,2}$. We next discretize the domain into rectangles as in [28, Definition 3.1], and denote the corresponding processes $\hat{Y}^r_{n,2,k}(t, y)$ and $\hat{X}^r_{n,2,k}(t, y)$ for a discretization size k.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

It is then easy to show that the finite-dimensional distributions of $\hat{Y}^r_{n,2,k} - \hat{X}^r_{n,2,k}$ converge to 0 for each k as $n \to \infty$. Finally, by [3, Theorem 3.2], it suffices to show that

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbb{P}(|(\hat{Y}^r_{n,2,k}(t, y) - \hat{X}^r_{n,2,k}(t, y)) - (\hat{Y}^r_{n,2}(t, y) - \hat{X}^r_{n,2}(t, y))| > \varepsilon) = 0$$

for each $t \geq 0$ and $0 \leq y \leq t$. This follows from a similar argument as in the proof of [28, Lemma 4.2] by noting the convergence $U^n \Rightarrow e$. This completes the proof. □

Proof of Theorem 4. The convergence of $\hat{X}^r_n(t, y)$ follows from Lemmas 8, 10, 11, 12, and the continuous mapping theorem. □

Appendix A

In this appendix we collect the proofs for the lemmas used in Sections 5 and 6.

Proof of Lemma 3. We first show that $z_n(t)$ is continuous in t in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$ for each n. To see this, let $t_k, t \in [0, T]$ for $T > 0$, and $t_k \to t$ as $k \to \infty$. We need to show that $\|z_n(t_k) - z_n(t)\|_{L_1} \to 0$ as $k \to \infty$, for each $n \geq 1$, that is, for each $Y > 0$,

$$\int_0^Y |x_n(y_n(t_k, s)) - x_n(y_n(t, s))| \, \mathrm{d}s \to 0 \quad \text{as } k \to \infty. \tag{A.1}$$

Let $\mathrm{disc}^s(x_n, y_n)$ be the set of s values such that s is a discontinuity point of $y_n(t, s)$ or $y_n(t, s)$ is a discontinuity point of x_n for the above $t \geq 0$. Since $y_n \in \mathbb{D}_{\mathbb{D}}$ is strictly increasing in s, it follows that there can only be countably many points in $\mathrm{disc}^s(x_n, y_n)$ for the above $t \geq 0$, which has Lebesgue measure 0.

Now for each $s \notin \mathrm{disc}^s(x_n, y_n)$, we obtain, for each $n \geq 1$,

$$|x_n(y_n(t_k, s)) - x_n(y_n(t, s))| \to 0 \quad \text{as } k \to \infty.$$

Since $x_n(y_n(t, s)) \in \mathbb{D}_{\mathbb{D}}$, $x_n(y_n(t, s))$ is bounded in finite-time intervals, and, thus, by the bounded convergence theorem together with the fact that $\mathrm{disc}^s(x_n, y_n)$ has Lebesgue measure 0, we obtain the convergence in (A.1). Similarly, $z(t)$ is also continuous in t in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$.

Finally, we prove $z_n \to z$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$ as $n \to \infty$. Since $z_n(t)$ and $z(t)$ are continuous in t in $\mathbb{D}([0, \infty), (\mathbb{D}, L_1))$, by Lemma 2 it suffices to show that for any $T > 0$ and any sequence $t_n, t \in [0, T]$ satisfying $t_n \to t$ as $n \to \infty$, $\|z_n(t_n) - z(t)\|_{L_1} \to 0$ as $n \to \infty$, that is, for each $Y > 0$,

$$\int_0^Y |x_n(y_n(t_n, s)) - x(y(t, s))| \, \mathrm{d}s \to 0 \quad \text{as } n \to \infty. \tag{A.2}$$

Let $\mathrm{disc}^s(x)$ be the set of s values such that $y(t, s)$ is a discontinuity point of x for the above $t \geq 0$. Since y is nondecreasing in t and strictly increasing in s, there can only be countably many points in $\mathrm{disc}^s(x)$ which have Lebesgue measure 0.

Now for each $s \notin \mathrm{disc}^s(x)$, since $y_n \to y$ in the space $\mathbb{D}_{\mathbb{D}}$ endowed with the Skorokhod J_1 topology (and, equivalently, with the uniform topology due to the fact that $y \in \mathbb{C}_{\mathbb{C}}$), by [7, Proposition 3.6.5], we have $|y_n(t_n, s) - y(t, s)| \to 0$ as $n \to \infty$. Thus, by the convergence $x_n \to x$ in $(\mathbb{D}, M_1)$, it follows that $x_n(t) \to x(t)$ uniformly at each continuity point of x, and since $s \notin \mathrm{disc}^s(x)$, we obtain $|x_n(y_n(t_n, s)) - x(y(t, s))| \to 0$ as $n \to \infty$. This, together with the fact that $\mathrm{disc}^s(x)$ has Lebesgue measure 0, implies that (A.2) holds and, thus, the lemma is proved. □

Proof of Lemma 4. Since $x_n(t)$ and $x(t)$ are constant functions of the second time parameter when they are regarded as functions in $\mathbb{D}_{\mathbb{D}}$, the same parametric representations for x_n and x

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

used for the convergence of $x_n(t) \to x(t)$ in $(\mathbb{D}, M_1)$ can be also used for the convergence of $x_n(t) \to x(t)$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$. We omit the details for brevity. □

Proof of Lemma 5. The continuity of summation in the M_1 topology follows from Lemma 4 and [29, Theorem III.3.1]. □

Proof of Lemma 6. Due to the continuity of x and y as functions of t in $\mathbb{D}([0,\infty), (\mathbb{D}, L_1))$, by Lemma 2 it suffices to prove that for any $T > 0$ and any sequence $t_n, t \in [0, T]$ satisfying $t_n \to t$ as $n \to \infty$, we have

$$\|x(t_n)y_n(t_n) - x(t)y(t)\|_{L_1} \to 0 \quad \text{as } n \to \infty.$$

Recall that $x(t_n) := x(t_n, \cdot)$ and, similarly, for $y_n(t_n)$, $x(t)$, and $y(t)$. Now,

$$\begin{aligned}\|x(t_n)y_n(t_n) - x(t)y(t)\|_{L_1} &\le \|x(t_n)y_n(t_n) - x(t_n)y(t)\|_{L_1} + \|x(t_n)y(t) - x(t)y(t)\|_{L_1}\\ &\le \|x(t_n)\|_{L_1}\|y_n(t_n) - y(t)\|_{L_1} + \|y(t)\|_{L_1}\|x(t_n) - x(t)\|_{L_1}\\ &\to 0.\end{aligned}$$

The convergence follows from the facts that $\sup_n \|x(t_n)\|_{L_1} < \infty$, $\|y(t)\|_{L_1} < \infty$, $y_n(t,s) \to y(t,s)$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$, and the continuity property of $y(t,s)$ and $x(t,s)$ as functions of t in $\mathbb{D}([0,\infty), (\mathbb{D}, L_1))$. □

Proof of Lemma 9. The proof proceeds in four steps.

Step 1. Since $z_n \to z$ in $(\mathbb{D}, M_1)$ as $n \to \infty$, by Lemma 4, we have $z_n \to z$ in $(\mathbb{D}_{\mathbb{D}}, M_1)$ as $n \to \infty$. Since F is continuous, by Lemma 6, we have

$$z_n(t)F^c(y) \to z(t)F^c(y) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \quad \text{as } n \to \infty.$$

Step 2. We claim that $\int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s) + y)$ is continuous in t in $\mathbb{D}([0,\infty), (\mathbb{D}, L_1))$. Suppose that $t_n, t \in [0, T]$ for $T > 0$ and $t_n \to t$ as $n \to \infty$, for each $y \ge 0$, we have

$$\begin{aligned}&\left|\int_0^{t_n} z(s)\,\mathrm{d}F^c(x(t_n) - x(s) + y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s) + y)\right|\\ &\quad\le \left|\int_0^t z(s)\,\mathrm{d}(F^c(x(t_n) - x(s) + y) - F^c(x(t) - x(s) + y))\right|\\ &\qquad + \left|\int_t^{t_n} z(s)\,\mathrm{d}F^c(x(t_n) - x(s) + y)\right|. \qquad \text{(A.3)}\end{aligned}$$

Since x and F are both continuous functions, $F^c(x(t_n) - x(s) + y) - F^c(x(t) - x(s) + y) \to 0$ as $n \to \infty$, and, thus, the same holds for the first integral on the right-hand side of (A.3). The second integral goes to 0 because $t_n \to t$ and the continuity of F. (So far, we have proved the pointwise convergence.) Given $Y > 0$, and $T > 0$ such that $t_n, t \in [0, T]$, $z(t)$ in $\mathbb{D}$ implies that $z(t)$ is bounded on $[0, T]$. So we have that $\int_0^Y |\int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s) + y)|\,\mathrm{d}y$ is bounded for each $Y > 0$. By (A.3) and the bounded convergence theorem, we obtain

$$\int_0^Y \left|\int_0^{t_n} z(s)\,\mathrm{d}F^c(x(t_n) - x(s) + y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t) - x(s) + y)\right|\mathrm{d}y \to 0 \quad \text{as } n \to \infty.$$

Thus, we have proved the claim. A similar argument also shows that

$$\int_0^t z_n(s)\,\mathrm{d}F^c(x_n(t) - x(s) + y)$$

is continuous in t in $\mathbb{D}([0,\infty), (\mathbb{D}, L_1))$ since we assume that $x_n \in \mathbb{C}$.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

Step 3. We prove that

$$\int_0^t z_n(s)\,\mathrm{d}F^c(x_n(t)-x_n(s)+y) \to \int_0^t z(s)\,\mathrm{d}F^c(x(t)-x(s)+y) \text{ in } (\mathbb{D}_{\mathbb{D}}, M_1) \quad \text{as } n\to\infty.$$

By the claim proved in step 2, by Lemma 2, it suffices to prove that for each $T>0$ and $Y>0$, for $t_n, t\in[0,T]$ and any sequence $t_n\to t$ as $n\to\infty$, we have

$$\int_0^Y \left| \int_0^{t_n} z_n(s)\,\mathrm{d}F^c(x_n(t_n)-x_n(s)+y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t)-x(s)+y)\right| \mathrm{d}y \to 0 \quad \text{(A.4)}$$

as $n\to\infty$. We first prove pointwise convergence and then L_1 convergence. For each $y\geq 0$,

$$\begin{aligned}
&\left| \int_0^{t_n} z_n(s)\,\mathrm{d}F^c(x_n(t_n)-x_n(s)+y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t)-x(s)+y)\right| \\
&\quad\leq \left| \int_0^{t_n} z_n(s)\,\mathrm{d}(F^c(x_n(t_n)-x_n(s)+y) - F^c(x(t)-x(s)+y))\right| \\
&\qquad + \left| \int_0^{t_n} (z_n(s)-z(s))\,\mathrm{d}F^c(x(t)-x(s)+y)\right| + \left| \int_t^{t_n} z(s)\,\mathrm{d}F(x(t)-x(s)+y)\right|.
\end{aligned} \quad \text{(A.5)}$$

The first integral on the right-hand side of (A.5) converges to 0 as $n\to\infty$ because of the continuity of F and $x_n(t_n)\to x(t)$ uniformly as $n\to\infty$ (recall that x_n and x are continuous). Since $z_n(t)\to z(t)$ in $(\mathbb{D}, M_1)$, the set $\{s\in[0,\infty)\colon |z_n(s)-z(s)|\not\to 0\}$ has Lebesgue measure 0, and, thus, the second integral on the right-hand side of (A.5) goes to 0 as $n\to\infty$. It is evident that the third integral on the right-hand side of (A.5) also goes to 0 as $n\to\infty$. Thus, we have proved that, for each $y>0$,

$$\left| \int_0^t z_n(s)\,\mathrm{d}F^c(x_n(t)-x_n(s)+y) - \int_0^t z(s)\,\mathrm{d}F^c(x(t)-x(s)+y)\right| \to 0 \quad \text{as } n\to\infty.$$

Given any $Y>0$, we obtain the convergence in (A.4) by applying the bounded convergence theorem.

Step 4. By the continuity of summation in $(\mathbb{D}_{\mathbb{D}}, M_1)$, the conclusion of the lemma holds and the proof is complete. □

Acknowledgements

This work is supported in part by the National Science Foundation (grant number CMMI-1538149). The authors thank the anonymous referee very much for the valuable and constructive comments that have helped improve the presentation of the paper. The authors also thank the helpful discussions with the doctoral student Hongyuan Lu at Penn State University.

References

[1] Anderson, D. *et al*. (2016). A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodol. Comput. Appl. Prob*. **18,** 153–168.

[2] Baykal-Gursoy, M. and Xiao, W. (2004). Stochastic decomposition in $M/M/\infty$ queues with Markov modulated service rates. *Queueing Systems* **48,** 75–88.

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

[4] Blom, J., Kella, O., Mandjes, M. and Thorsdottir, H. (2014). Markov-modulated infinite-server queues with general service times. *Queueing Systems* **76,** 403–424.

[5] D'Auria, B. (2007). Stochastic decomposition of the $M/G/\infty$ queue in a random environment. *Operat. Res. Lett*. **35,** 805–812.

[6] D'Auria, B. (2008). $M/M/\infty$ queues in semi-Markovian random environment. *Queueing Systems* **58,** 221–237.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press

[7] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence.* John Wiley, New York.
[8] Falin, G. (2008). The $M/M/\infty$ queue in a random environment. *Queueing Systems* **58,** 65–76.
[9] Fralix, B. H. and Adan, I. J. B. F. (2009). An infinite-server queue influenced by a semi-Markovian environment. *Queueing Systems* **61,** 65–84.
[10] Glynn, P. W. (1982). On the Markov property of the $GI/G/\infty$ Gaussian limit. *Adv. Appl. Prob.* **14,** 191–194.
[11] Glynn, P. W. and Whitt, W. (1991). A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Prob.* **23,** 188–209.
[12] Jacobs, P. A. (1980). Heavy traffic results for single-server queues with dependent (EARMA) service and interarrival times. *Adv. Appl. Prob.* **12,** 517–529.
[13] Jacobs, P. A. and Lewis, P. A. W. (1977). A mixed autoregressive-moving average exponential sequence and point process (EARMA 1, 1). *Adv. Appl. Prob.* **9,** 87–104.
[14] Jacobs, P. A. and Lewis, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.* **4,** 19–36.
[15] Jayawardene, A. K. and Kella, O. (1996). $M/G/\infty$ with alternating renewal breakdowns. *Queueing Systems Theory Appl.* **22,** 79–95.
[16] Kella, O. and Whitt, W. (1990). Diffusion approximations for queues with server vacations. *Adv. Appl. Prob.* **22,** 706–729.
[17] Kella, O. and Whitt, W. (1991). Queues with server vacations and Lévy processes with secondary jump input. *Ann. Appl. Prob*. **1,** 104–117.
[18] Krichagina, E. V. and Puhalskii, A. A. (1997). A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems Theory Appl.* **25,** 235–280.
[19] Lu, H., Pang, G. and Zhou, Y. (2016). $G/GI/N(+GI)$ queues with service interruptions in the Halfin–Whitt regime. *Math. Meth. Operat. Res.* **83,** 127–160.
[20] Mitrany, I. L. and Avi-Itzhak, B. (1968). A many-server queue with service interruptions. *Operat. Res*. **16,** 628–638.
[21] O'Cinneide, C. A. and Purdue, P. (1986). The $M/M/\infty$ queue in a random environment. *J. Appl. Prob.* **23,** 175–184.
[22] Pang, G. and Whitt, W. (2009). Heavy-traffic limits for many-server queues with service interruptions. *Queueing Systems* **61,** 167–202.
[23] Pang, G. and Whitt, W. (2009). Service interruptions in large-scale service systems. *Manag. Sci*. **55,** 1499–1512.
[24] Pang, G. and Whitt, W. (2010). Continuity of a queueing integral representation in the M_1 topology. *Ann. Appl. Prob.* **20,** 214–237.
[25] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65,** 325–364.
[26] Pang, G. and Whitt, W. (2012). Infinite-server queues with batch arrivals and dependent service times. *Prob. Eng. Inf. Sci.* **26,** 197–220.
[27] Pang, G. and Whitt, W. (2012). The impact of dependent service times on large-scale service systems. *Manufacturing Service Operat. Manag.* **14,** 262–278.
[28] Pang, G. and Whitt, W. (2013). Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems* **73,** 119–146.
[29] Pomarede, J.-M. L. (1976). A unified approach via graphs to Skorokhod's topologies on the function space D. Ph.D. Doctoral Thesis, Yale University.
[30] Reed, J. and Talreja, R. (2015). Distribution-valued heavy-traffic limits for the $G/GI/\infty$ queue. *Ann. Appl. Prob.* **25,** 1420–1474.
[31] Sim, C. H. (1990). First-order autoregressive models for gamma and exponential processes. *J. Appl. Prob.* **27,** 325–332.
[32] Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory Prob. Appl.* **1,** 261–290.
[33] Talreja, R. and Whitt, W. (2009). Heavy-traffic limits for waiting times in many-server queues with abandonment. *Ann. Appl. Prob*. **19,** 2137–2175.
[34] Whitt, W. (1982). On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv. Appl. Prob.* **14,** 171–190.
[35] Whitt, W. (2002). *Stochastic Process Limits.* Springer, New York.

https://doi.org/10.1017/apr.2016.29 Published online by Cambridge University Press