J. Appl. Probab. **61,** 490–514 (2024)
*doi:*10.1017/jpr.2023.50

HEAVY-TRAFFIC LIMITS FOR PARALLEL SINGLE-SERVER QUEUES WITH RANDOMLY SPLIT HAWKES ARRIVAL PROCESSES

BO LI,* *Nankai University*
GUODONG PANG,** *Rice University*

Abstract

We consider parallel single-server queues in heavy traffic with randomly split Hawkes arrival processes. The service times are assumed to be independent and identically distributed (i.i.d.) in each queue and are independent in different queues. In the critically loaded regime at each queue, it is shown that the diffusion-scaled queueing and workload processes converge to a multidimensional reflected Brownian motion in the non-negative orthant with orthonormal reflections. For the model with abandonment, we also show that the corresponding limit is a multidimensional reflected Ornstein–Uhlenbeck diffusion in the non-negative orthant.

Keywords: Hawkes process; random splitting/sampling; functional central limit theorem; parallel single-server queues; multidimensional reflected Brownian motion; multidimensional reflected Ornstein–Uhlenbeck diffusion

2020 Mathematics Subject Classification: Primary 60G55; 60K25; 60F17; 90B22
Secondary 60H10; 60J70

1. Introduction

Single-server queues are fundamental models in applied probability and queueing theory, and heavy-traffic limits have been well studied; see the recent surveys in [7], [31], and [32], and the recent work on the model with abandonment in [28] and [30]. In these studies, the arrival processes are usually assumed to be Poisson or renewal processes. Recently, to account for the excessive burstiness, clustering effects, and path-dependence, other point processes such as Hawkes and Pólya processes have been used to model arrivals in queueing models. They can be used to model, for example, internet/social media traffic flows, patient flows during a pandemic, neuron interaction processes, and high-frequency transaction processes in limit order books. Queues with Hawkes input are studied in [8], [11], [16], [22], and [26], and single-server queues with Pólya arrival processes are considered in [14] and [15]. On the other hand, many systems have parallel servers, such as data centers and internet processors, and various randomized routeing schemes have been developed in the literature (see e.g. the recent survey [12]). A simple scheme is to send incoming jobs/customers randomly to any of the servers upon arrival.

Received 15 July 2022; accepted 9 June 2023.
* Postal address: School of Mathematics and LPMC, Nankai University, Tianjin, 300071 China. Email address: libo@nankai.edu.cn
** Postal address: Department of Computational Applied Mathematics and Operations Research, George R. Brown School of Engineering, Rice University, Houston, TX 77005, USA. Email address: gdpang@rice.edu

In this paper we consider parallel single-server queues in heavy traffic, where the arrival process for each queue is randomly split from one Hawkes process. That is, upon each arrival, a customer or job is randomly assigned to one of the queues. The service times in the different queues are mutually independent and may have different distributions, and within a queue the service times are assumed to be independent and identically distributed (i.i.d.). The service discipline in each queue is first-come first-served (FCFS). We assume that each queue is critically loaded in heavy traffic, that is, the traffic intensity gets close to one. We also consider the queueing model with abandonment, where customers joining each queue may abandon the system while waiting in the queue and before receiving service. We focus on the joint queue length and workload processes, and prove the functional central limit theorem (FCLT) for the associated diffusion-scaled processes.

Randomly split Hawkes processes have recently been studied by the present authors [23]. Unlike Poisson processes, the split Hawkes processes become a multivariate Hawkes process with a particular dependence structure (see the covariance function of scaling limits in Proposition 2.1). As a consequence, the vector of queueing (workload) processes of the parallel server queues will be correlated. In contrast, in the case of randomly split Poisson processes, because of the independence property of Poisson thinning/sampling, the queues at the parallel servers are mutually independent and each queue can be analyzed as an M/G/1 queue. However, our model becomes more challenging to analyze directly. This is in a similar flavor to the open problem recently posed by Mandjes [24] about multivariate M/G/1 queues with coupled input and parallel service. Our model presents another example of multivariate single-server queues with correlated input, and our results provide heavy-traffic diffusion approximations for such a system.

For the queueing models without abandonment, the queueing limit process is a multidimensional reflected Brownian motion (RBM) in the non-negative orthant with orthonormal reflections. In many queueing networks, RBMs arise as the heavy-traffic limits of the queueing and workload processes; see e.g. the recent surveys [10] and [34]. In the very original article by Harrison [17], as a diffusion approximation for a pair of single-server queues in series, a two-dimensional RBM in the non-negative quadrant was introduced, which has a normal reflection at one axis and a tangential reflection at the other axis. On the other hand, our model with two queues provides an example of RBMs in the non-negative quadrant with normal reflections at both axes. A multidimensional RBM limit with orthonormal reflections was derived for a network of parallel single-server queues with Markov-modulated service speeds in [13]. This work contributes to the literature of queueing network scaling limits by providing a concrete example of RBMs in a non-negative orthant with orthonormal reflections.

Although the correlation structure of the split Hawkes processes can be explicitly characterized (as can the Brownian motion without reflection), it is challenging to compute the covariance functions for the limiting RBM in the parallel server queueing model. Moreover, one can check that the conditions of Propositions 8 and 9 in [9] (see also [33]) are not satisfied so that the limit process does not possess a product-form stationary distribution. Fortunately, the numerical approach developed in [9] to compute the stationary distribution can be used for the limiting RBM in our model. We implement that algorithm in a two-queue example to illustrate how the parameters in the Hawkes process and the splitting probabilities as well as the heavy-traffic scalings affect the correlation between the steady state of the two queues in the limit (see Section 3.1).

For the queueing models with abandonment, the queueing limit process is a multidimensional reflected Ornstein–Uhlenbeck (OU) diffusion in the non-negative orthant with

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

orthonormal reflections. This extends the one-dimensional results in [30]. This result is of interest in two respects. First, there is a very limited result in the queueing network literature that gives a multidimensional reflected OU limit. Huang and Zhang [20] study open Jackson networks with reneging, and obtain a multidimensional reflected OU diffusion limit with a reflection being characterized via the routeing probability matrix (extending the RBM approximations for open Jackson networks without abandonment in [25]). The reflection is very complex in that model, but the reflections in the limits of our model are orthonormal, which is of interest in its own right. It is worth noting that in the studies of dynamic scheduling in multiclass GI/GI/1+GI queues [1, 21], although the queueing processes are multidimensional, the Brownian control problem is solved via the so-called workload process as a one-dimensional controlled OU process with reflection. Second, there have been studies of the stationary distributions and ergodic properties of reflected OU processes in one dimension; see e.g. [29] and [35]. However, such results beyond one dimension are wide open. Our limit process provides a concrete example for which an explicit characterization of the stationary distribution could potentially be obtained. We leave this as an open problem for future work. (We also refer the readers to a relevant article [2] for the recurrence properties of multidimensional reflected diffusions in convex polyhedral cones.) It would also be interesting to explore whether the numerical scheme in [9] could be further developed for the multidimensional reflected OU processes such as those arising from our model.

The proof for the model without abandonment uses the continuous mapping approach for the multidimensional reflection maps (see [31]). This is standard, but the particular scaling involved in the Hawkes process and its splitting scheme must be taken into account. For the queueing models with abandonment, we adapt the approach in [30] by comparing with the model without abandonment. It is therefore also important to first study the model without abandonment. In the meantime, the comparison approach is non-trivial, since this requires us to use the martingale representations for the Hawkes process and the randomly split Hawkes processes [3, 23], as well as some martingale inequalities and properties.

We also discuss two related models in which the arrival processes for the parallel servers come from a multivariate Hawkes process in Section 5. The limits for the models with and without abandonment are of the same type with orthonormal reflections but the driving Brownian motions have a different covariance function. Although the split Hawkes process is also a multivariate Hawkes process, the splitting scheme introduces a particular structure together with only one self-exciting function. On the other hand, a general multivariate Hawkes process has a matrix of self-exciting functions. This is yet another example of the open problems posed in [24], and also complements the parallel server queueing model with correlated services due to Markov modulation in [13]. The limits for these models will also be of interest in their own right.

1.1. Organization of the paper

In Section 2 we discuss random splitting of Hawkes processes and review the FCLT results. In Section 3 we present the parallel single-server queueing model with split Hawkes arrival processes and the FCLT on the joint queueing and workload processes. A numerical example is provided in Section 3.1 to illustrate the stationary distribution of the limiting queueing process in a model with two queues. In Section 4 we discuss the model with abandonment and state the corresponding FCLT. In Section 5 we consider the parallel server queueing model with a general multivariate Hawkes process and state the corresponding scaling limits. The proofs for these models are given in Section 6 and the Appendix.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

1.2. Notation

All random variables and processes are defined in a common complete probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. Throughout the paper, $\mathbb{N}$ denotes the set of natural numbers. $\mathbb{R}^d(\mathbb{R}_+^d)$ denotes the space of d-dimensional real (non-negative) vectors; we write $\mathbb{R}(\mathbb{R}_+)$ when $d=1$. Let $\mathbb{D}^d = \mathbb{D}^d(\mathbb{R}_+, \mathbb{R}^d)$ denote the space of $\mathbb{R}^d$-valued càdlàg functions on $\mathbb{R}_+$. $(\mathbb{D}, J_1)$ denotes the space $\mathbb{D}$ equipped with Skorokhod J_1 topology (see [4]), which is complete and separable. $\mathbb{D}^d$ denotes the space of $\mathbb{R}^d$-valued càdlàg functions endowed with the weak Skorokhod J_1 topology [32], for which we write $(\mathbb{D}^d, J_1)$. For an integrable function $f:\mathbb{R}\to\mathbb{R}$, its L^1 norm is denoted by $\|f\|_1$. Notations $\to$ and $\Rightarrow$ mean convergence of real numbers and convergence in distribution, respectively. For a vector a, $\mathrm{diag}(a)$ denotes the diagonal matrix with the elements of vector a on the main diagonal. For two vectors $a=(a_k)_k$ and $b=(b_k)_k$, $ab=(a_kb_k)_k=\mathrm{diag}(a)b$ denotes their elementwise product. We use $\mathfrak{e}$ to denote the identity function $\mathfrak{e}(t)=t$ for $t\in\mathbb{R}_+$. Additional notation is introduced in the paper whenever necessary.

2. Random splitting of Hawkes processes

A one-dimensional Hawkes process, $N=\{N(t), t\geq 0\}$, is a simple counting process with conditional intensity

$$\lambda(t)=\lambda_0+\sum_{j=1}^{N(t)}H(t-\tau_j)=\lambda_0+\int_0^t H(t-s)\,\mathrm{d}N(s), \tag{2.1}$$

where $\lambda_0>0$ is a constant, called the baseline intensity, $H:\mathbb{R}_+\to\mathbb{R}_+$ is the self-exciting function, and τ_j denotes the jth event time of N.

We consider the randomly split Hawkes processes $N_k=\{N_k(t), t\geq 0\}$, $k=1,\dots,d$, defined as follows:

$$N_k(t)=\sum_{j\geq 1}\mathbf{1}(\xi_j=k,\tau_j\leq t),$$

where $\{\xi_j, j\geq 1\}$ is a sequence of i.i.d. random variables, independent of N, with the distribution

$$\mathbb{P}(\xi_j=k)=p_k\in(0,1)\quad\text{and}\quad\sum_{k=1}^d p_k=1.$$

It is shown in [23] that the splitting Hawkes process $(N_k)_k$ is a multivariate Hawkes process with conditional intensity

$$\lambda_k(t)=p_k\lambda(t)=p_k\lambda_0+\sum_{k'=1}^d\int_0^t(p_kH(t-s))\,\mathrm{d}N_{k'}(s). \tag{2.2}$$

Notice that the split processes N_k are not independent.

We consider a sequence of Hawkes processes indexed by n, that is, $N^{(n)}$ is a Hawkes process for the nth system with intensity process $\lambda^{(n)}(\cdot)$ whose baseline intensity in (2.1) is $\lambda_0^{(n)}$ while the self-exciting function H stays the same. The splitting variables are denoted by $\{\xi_j^{(n)}, j\geq 1\}$ with distribution $(p_k^{(n)})_k$.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Define the LLN and CLT-scaled processes for the splitting $\big(N_k^{(n)}\big)_k$ by

$$\bar{N}_k^{(n)}(t) = \frac{1}{n} N_k^{(n)}(nt) \quad \text{and} \quad \hat{N}_k^{(n)}(t) = \sqrt{n}\big(\bar{N}_k^{(n)}(t) - \mathbb{E}\big[\bar{N}_k^{(n)}(t)\big]\big), \tag{2.3}$$

respectively. It is easy to see (see e.g. [3]) that

$$\mathbb{E}\big[\bar{N}_k^{(n)}(t)\big] = \lambda_0^{(n)} p_k^{(n)} \int_0^t (1 + \varphi * \mathfrak{e}(ns))\, \mathrm{d}s, \tag{2.4}$$

where $\varphi = \sum_{j\geq 1} H^{*j}$ is the renewal function of H,

$$f * g(x) = \int_0^x f(y) g(x-y)\, \mathrm{d}y$$

denotes the convolution of f and g on $\mathbb{R}_+$, and $\mathfrak{e}$ is the identity function.

The following FCLT for the splitting processes $\big(N_k^{(n)}\big)_k$ is proved in [23].

Proposition 2.1. *Suppose that*

$$\lambda_0^{(n)} \to \lambda_0 \quad \textit{and} \quad \big(p_k^{(n)}\big) \to (p_k)_k \quad \textit{as } n \to \infty \quad \textit{and} \quad \|H\|_1 = \int_0^\infty H(t)\, \mathrm{d}t \in (0, 1). \tag{2.5}$$

We have

$$\big(\hat{N}_k^{(n)}\big)_k \Rightarrow (\hat{N}_k)_k \quad \textit{in } (\mathbb{D}^d, J_1) \quad \textit{as } n \to \infty,$$

where $(\hat{N}_k)_k$ is a d-dimensional Brownian motion with mean zero and covariance matrix

$$\operatorname{cov}(\hat{N}_k(t), \hat{N}_{k'}(s)) = (t \wedge s) \cdot \left(\frac{\lambda_0 p_k (\delta_{kk'} - p_{k'})}{1 - \|H\|_1} + \frac{\lambda_0 p_k p_{k'}}{(1 - \|H\|_1)^3} \right),$$

where $\delta_{kk'} = 1$ if $k = k'$ and $\delta_{kk'} = 0$ if $k \neq k'$.

The process $\hat{N}_k$ admits the representations

$$\begin{aligned} \hat{N}_k &= \frac{\lambda_0^{1/2}}{(1 - \|H\|_1)^{1/2}} \sqrt{p_k} W_k + p_k \frac{\lambda_0^{1/2} \|H\|_1}{(1 - \|H\|_1)^{3/2}} W \\ &= \frac{\lambda_0^{1/2}}{(1 - \|H\|_1)^{1/2}} \hat{S}_k + p_k \frac{\lambda_0^{1/2}}{(1 - \|H\|_1)^{3/2}} W, \end{aligned} \tag{2.6}$$

where $(W_k)_k$ is a standard d-dimensional Brownian motion, $W = \sum_{k=1}^d \sqrt{p_k} W_k$, and $\hat{S}_k = \sqrt{p_k} W_k - p_k W$.

The condition $\|H\|_1 \in (0, 1)$ is referred to as *the stability condition* in the literature of Hawkes processes, under which a stationary version of Hawkes process exists and $\varphi * \mathfrak{e}(t) \to \|H\|_1/(1 - \|H\|_1)$ as $t \to \infty$. The first representation in (2.6) is a direct consequence of [3, Theorem 2] for the vector-valued Hawkes process $\big(N_k^{(n)}\big)_k$ characterized by (2.2), while the second representation follows from [32, Chapter 9.5], where $\hat{S}$ is also a d-dimensional Brownian motion, independent of W, with covariance function

$$\operatorname{cov}(\hat{S}_k(t), \hat{S}_{k'}(s)) = p_k(\delta_{kk'} - p_{k'})(t \wedge s).$$

It is necessary that $\sum_k \hat{S}_k = 0$.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

In our study of parallel server queues with a split Hawkes process as arrivals, we will also need the following alternative diffusion-scaled process by replacing the centering term $\mathbb{E}\big[\bar{N}_k^{(n)}(t)\big]$ in (2.3) with a linear function:

$$\check{N}_k^{(n)}(t) : eq\sqrt{n}\left(\bar{N}_k^{(n)}(t) - \frac{\lambda_0^{(n)} p_k^{(n)} t}{1 - \|H\|_1}\right), \quad t \geq 0. \tag{2.7}$$

The proof of the proposition is given in the Appendix.

Proposition 2.2. *Under the conditions in Proposition* 2.1, *if, in addition,*

$$t^{1/2} \int_t^\infty H(s)\, \mathrm{d}s \to 0 \quad \text{as } t \to \infty, \tag{2.8}$$

then

$$\big(\check{N}_k^{(n)}\big)_k \Rightarrow (\hat{N}_k)_k \quad \text{in } (\mathbb{D}^d, J_1) \quad \text{as } n \to \infty, \tag{2.9}$$

where $(\hat{N}_k)_k$ *is the same limit as given in Proposition* 2.1.

3. Parallel single-server queues with split Hawkes arrival processes

In this model, there are d parallel servers and each server has its own queue. Arrivals in each queue come from the randomly split Hawkes process and are served in the FCFS discipline. Let $\{N(t) : t \geq 0\}$ be the Hawkes process, arriving in the system, and let $(N_k)_k$ be the splitting Hawkes arrival processes with the splitting mechanism as described in Section 2. Recall the baseline rate λ_0, splitting probability $(p_k)_k$, and self-exciting function H. For every $k = 1, 2, \ldots, d$, let $\{\eta_{j,k}\}_j$ represent the service times of customers in the kth queue. We assume that $\{\eta_{j,k}\}_j$ are i.i.d. with finite mean m_k and variance σ_k^2, and are independent of the Hawkes arrival process and the random splitting process. Denote the service rate of the kth queue by $\mu_k := 1/m_k$, and write $\mu = (\mu_k)_k$.

We now give a definition of the model. For every $k = 1, \ldots, d$, let

$$V_k(m) = \sum_{j=1}^{m} \eta_{j,k}$$

be the partial sum with the i.i.d. service times, and let $U_k(t) := \sup\{m \geq 0 : V_k(m) \leq t\}$ be its right-continuous inverse. Then $U_k(t)$ is a renewal process representing the number of jobs that the server k can potentially complete by time t. Let $Z = (Z_k)_k$ be the *workload processes*, and let $Q = (Q_k)_k$ be the *queue length processes*, with $Q(0) = (Q_k(0))_k$ being the number of initial jobs in each queue. Under our assumptions, the processes N and (V, U) are independent, and we further assume that $Q(0)$ is independent of the new arrivals $N = (N_k)_k$ and the service processes. Then we have the following flow-balance equations for the dynamics at each queue: for $k = 1, \ldots, d$,

$$\begin{aligned} Z_k(t) &= V_k(Q_k(0) + N_k(t)) - B_k(t), \\ Q_k(t) &= Q_k(0) + N_k(t) - U_k(B_k(t)), \end{aligned} \tag{3.1}$$

where

$$B_k(t) = \int_0^t \mathbf{1}(Q_k(s) > 0)\, \mathrm{d}s = \int_0^t \mathbf{1}(Z_k(s) > 0)\, \mathrm{d}s$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

is the *busy-time process*, and its paired *idle-time process* is given by

$$I_k(t) = t - B_k(t) = \int_0^t \mathbf{1}(Q_k(s) = 0)\,\mathrm{d}s = \int_0^t \mathbf{1}(Z_k(s) = 0)\,\mathrm{d}s.$$

The processes $(Q_k, Z_k)_k$ take values in the non-negative orthant $\mathbb{R}_+^{2d}$, and can be characterized by the following version of reflection maps (see [32, Chapter 14.2]).

Definition 3.1. For any $x \in \mathbb{D}^d$ and any reflection matrix, i.e. a matrix with $\mathtt{Q}_{ij} \geq 0$ and $\sum_i \mathtt{Q}_{ij} \leq 1$ such that $(\mathtt{Q})^n \to 0$ as $n \to \infty$, let the feasible regulator set be

$$\Psi_{\mathtt{Q}}(x) \equiv \left\{w \in \mathbb{D}_{\uparrow}^d : x + (\mathtt{I} - \mathtt{Q})w \geq 0\right\},$$

where $\mathtt{I}$ is the identity matrix and $\mathbb{D}_{\uparrow}^d$ is the subset of functions in $\mathbb{D}^d$ that are non-decreasing and non-negative in each coordinate. Then the *reflection mapping* is defined as

$$(z, y) := (\phi_{\mathtt{Q}}, \psi_{\mathtt{Q}})(x) : \mathbb{D}^d \to \mathbb{D}^{2d},$$

where y is called the *regulator component,* given by

$$y := \psi_{\mathtt{Q}}(x) := \inf \Psi_{\mathtt{Q}}(x) = \inf\{w : w \in \Psi_{\mathtt{Q}}(x)\},$$

that is, for all i and t,

$$y_i(t) := \inf\{w_i(t) \in \mathbb{R} : w \in \Psi_{\mathtt{Q}}(x)\},$$

and where z is called the *content component,* given by

$$z = \phi_{\mathtt{Q}}(x) = x + (\mathtt{I} - \mathtt{Q})y.$$

In the definition above, the matrix $\mathtt{I} - \mathtt{Q}$ is the direction of reflection (see [9, 18]), that is, whenever the boundary face $\{z \in \mathbb{R}_+^d : z_j = 0\}$ is hit for some j, the process w_j increases and causes an instantaneous displacement of z in the direction given by $\mathrm{col}_j(\mathtt{I} - \mathtt{Q})$, the jth column of $(\mathtt{I} - \mathtt{Q})$; $y = \psi_{\mathtt{Q}}$ is the minimal element in $\Psi_{\mathtt{Q}}$ such that $z \geq 0$. Moreover, the *complementarity property* holds ([32, Theorem 14.2.3]), that is,

$$\int_0^\infty z_i\,\mathrm{d}y_i = 0 \quad \text{for all } i,$$

which also characterizes the regulator uniquely. It is proved in [32, Theorems 14.2.5 and 14.2.7] that $\psi_{\mathtt{Q}}$ is well-defined and Lipschitz under both the uniform norm and the Skorokhod J_1 topology.

In particular, if the reflection matrix $\mathtt{Q} = 0$, then *the angle of reflection* is 0 (see [19]), and the reflection direction is orthogonal to the boundary, which is the case in our paper. Thus the reflection is also referred to as *orthonormal*. Letting $(\phi, \psi) \equiv (\phi_0, \psi_0)$ denote the operator in this case, and recalling that $\mathfrak{e}(t) \equiv t$ is the identity function on $\mathbb{R}_+$, we can rewrite the processes in (3.1) in terms of Definition 3.1 as follows:

$$(Z, I) = (\phi, \psi)(V \circ (Q(0) + N) - \mathfrak{e}), \tag{3.2}$$

and the queue length process can be rewritten similarly.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

We consider a sequence of such queueing systems in heavy traffic and indexed by the parameter n with $n \to \infty$. The traffic intensity for the kth queue is given by

$$\rho_k^{(n)} = \frac{\lambda_0^{(n)} p_k^{(n)}}{(1 - \|H\|_1)\mu_k^{(n)}}, \quad k = 1, \ldots, d. \tag{3.3}$$

Define the following scaled processes:

$$\begin{gathered} \bar{V}^{(n)}(t) = \frac{1}{n} V^{(n)}([nt]), \quad \bar{U}^{(n)}(t) = \frac{1}{n} U^{(n)}(nt), \quad \bar{B}^{(n)}(t) = \frac{1}{n} B^{(n)}(nt), \\ \hat{V}^{(n)}(t) = \sqrt{n}\big(\bar{V}^{(n)}(t) - tm^{(n)}\big), \quad \hat{U}^{(n)}(t) = \sqrt{n}\big(\bar{U}^{(n)}(t) - t\mu^{(n)}\big), \end{gathered} \tag{3.4}$$

and

$$\hat{Q}^{(n)}(t) = \frac{1}{\sqrt{n}} Q^{(n)}(nt), \quad \hat{Z}^{(n)}(t) = \frac{1}{\sqrt{n}} Z^{(n)}(nt), \quad t \geq 0.$$

We assume the following conditions on the service processes.

Assumption 3.1. *Assume that the service times $\{\eta_{j,k}\}_j$ for $k = 1, 2, \ldots, d$, satisfy*

$$m_k^{(n)} = \mathbb{E}\big[\eta_{1,k}^{(n)}\big] \to m_k =: \mu_k^{-1} \quad \textit{and} \quad \sigma_k^{(n)} = \sqrt{\mathrm{var}\big(\eta_{1,k}^{(n)}\big)} \to \sigma_k \quad \textit{as } n \to \infty, \tag{3.5}$$

and letting $F_k^{(n)}$ be the cumulative distribution function (CDF) of $\eta_{1,k}^{(n)}$, we have for every $\varepsilon > 0$

$$\int_{y > \sqrt{n}\varepsilon} y^2 \, \mathrm{d}F_k^{(n)}(y) \to 0 \quad \textit{as } n \to \infty. \tag{3.6}$$

The traffic intensity at each queue satisfies the following: for $k = 1, \ldots, d$,

$$\sqrt{n}\big(1 - \rho_k^{(n)}\big) \to \hat{\rho}_k > 0 \quad \textit{as } n \to \infty. \tag{3.7}$$

Observe that by the definition in (3.3), under (3.7), we have

$$\lambda_0 p_k m_k = \frac{\lambda_0 p_k}{\mu_k} = 1 - \|H\|_1 \quad \text{for each } k = 1, 2, \ldots, d. \tag{3.8}$$

Remark 3.1. Condition (3.6) is known as *Lindeberg's condition* for the triangular array $\{\eta_{j,k}^{(n)}\}_{j,k}$ (see [5, Theorem 27.2]), under which one can show that as $n \to \infty$,

$$\sup_{j \leq n} \frac{1}{\sqrt{n}} \eta_{j,k}^{(n)} \Rightarrow 0,$$

and the processes in (3.4) satisfies

$$\begin{gathered} \big(\bar{V}^{(n)}, \bar{U}^{(n)}\big) \to (\mu^{-1}, \mu)\mathfrak{e} \quad \text{u.o.c. in probability}, \\ \big(\hat{V}^{(n)}, \hat{U}^{(n)}\big) \Rightarrow (\hat{V}, \hat{U}) \quad \text{in } (\mathbb{D}^{2d}, J_1), \end{gathered} \tag{3.9}$$

where u.o.c. is short for 'uniformly on every compact set on $\mathbb{R}_+$', $\hat{V}$ is a d-dimensional Brownian motion with mean zero and covariance matrix

$$\mathrm{cov}(\hat{V}_k(t), \hat{V}_{k'}(s)) = (t \wedge s)\sigma_k^2 \delta_{kk'},$$

and $\hat{U}_k(t) = -\mu_k \hat{V}_k(\mu_k t)$ for $t \geq 0$.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Remark 3.2. In addition to (3.5), if we assume

$$\sqrt{n}\big(\lambda_0 - \lambda_0^{(n)}\big) \to \hat{\lambda}_0, \quad \sqrt{n}\big(\mu_k - \mu_k^{(n)}\big) \to \hat{\mu}_k, \quad \sqrt{n}\big(p_k - p_k^{(n)}\big) \to \hat{p}_k$$

for some $\hat{\lambda}_0, \hat{\mu}_k, \hat{p}_k \in \mathbb{R}$, as $n \to \infty$, where p_k is given in (2.5), then (3.7) holds with

$$\hat{\rho}_k = \frac{\hat{\lambda}_0}{\lambda_0} + \frac{\hat{p}_k}{p_k} - \frac{\hat{\mu}_k}{\mu_k}.$$

The process limits remain the same.

We have the following FCLT for the diffusion-scaled processes $\big(\hat{Q}^{(n)}, \hat{Z}^{(n)}\big)$.

Theorem 3.1. *Suppose that Assumption 3.1 and the conditions in (2.5) and (2.8) hold, and that* $\hat{Q}^{(n)}(0) \Rightarrow \hat{Q}(0) \geq 0$. *Then*

$$\big(\hat{Q}^{(n)}, \hat{Z}^{(n)}\big) \Rightarrow (\hat{Q}, \hat{Z}) \quad \text{in } (\mathbb{D}^{2d}, J_1) \quad \text{as } n \to \infty, \tag{3.10}$$

with the limit

$$\hat{Q} = \phi(\hat{Q}(0) + \hat{W} - \hat{\theta}\mathfrak{e}) \quad \text{and} \quad \hat{Z} = m\hat{Q} = (m_k\hat{Q}_k)_k,$$

where

$$\hat{\theta} = \mu\hat{\rho} = (\mu_k\hat{\rho}_k)_k$$

with $\hat{\rho}_k$ *in* (3.7), *and* $\hat{W}$ *is a* d*-dimensional Brownian motion with covariance function*

$$\begin{aligned} \operatorname{cov}(\hat{W}_k(t), \hat{W}_{k'}(s)) &= (t \wedge s) \cdot \hat{R}_{kk'}, \\ \hat{R}_{kk'} &:= \frac{\lambda_0 p_k p_{k'}}{(1 - \|H\|_1)^3} - \frac{\lambda_0 p_k p_{k'}}{1 - \|H\|_1} + \mu_k\big(1 + \mu_k^2\sigma_k^2\big)\delta_{kk'} \\ &= \mu_k\left(\frac{p_{k'}}{(1 - \|H\|_1)^2} - p_{k'} + \big(1 + \mu_k^2\sigma_k^2\big)\delta_{kk'}\right). \end{aligned} \tag{3.11}$$

3.1. The stationary distribution of $\hat{Q}$

Observe that $\hat{Q}$ in Theorem 3.1 is a reflected Brownian motion on the orthant $\mathbb{R}_+^d$ with drift vector $\hat{\theta}$, covariance matrix $\hat{R}$, and reflection matrix $\mathtt{I}$. Since $\hat{\theta}_k > 0$ for every $k = 1, 2, \ldots, d$, there exists a unique stationary distribution of $\hat{Q}$.

Recall that a probability measure π on $\mathbb{R}_+^d$ is called a stationary distribution for the reflected Brownian motion $\hat{Q}$ if, for every bounded Borel function f on $\mathbb{R}_+^d$ and every $t > 0$,

$$\int_{\mathbb{R}_+^d} \mathbb{E}_x[f(\hat{Q}(t))]\pi(\mathrm{d}x) = \int_{\mathbb{R}_+^d} f(x)\pi(\mathrm{d}x).$$

It is shown in [18] and [33] that the stationary distribution for a reflected Brownian motion in the orthant is equivalent to Lebesgue measure on $\mathbb{R}_+^d$, and the occupation measures on faces are absolutely continuous with respect to the $(d-1)$-dimensional Lebesgue measure $\nu_k(\cdot)$ on the face $F_k := \{z \in \mathbb{R}_+^d : z_k = 0\}$, that is,

$$\pi(\mathrm{d}x) = q_0(x)\,\mathrm{d}x \quad \text{and} \quad \mathbb{E}_\pi\left[\int_0^t \mathbf{1}_A(\hat{Q}(s))\,\mathrm{d}\hat{Y}_k(s)\right] = \frac{1}{2}t\int_A q_k(y)\nu_k(\mathrm{d}y) \quad \text{for all } A \in \mathscr{B}(F_k).$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

TABLE 1. A numerical example to illustrate the stationary distribution of the limiting queueing process in a model with two queues.

p_1	$\mathbb{E}[\hat{Q}_1(\infty)]$	$\mathbb{E}[\hat{Q}_2(\infty)]$	Corr-of-$\hat{Q}$	$\mathbb{E}[\hat{Q}_1(\infty)]$	$\mathbb{E}[\hat{Q}_2(\infty)]$	Corr-of-$\hat{Q}$	Corr-of-$\hat{W}$
0.1	1.87415	2.99150	0.05862	0.56224	1.19660	0.06673	0.15541
0.3	2.49320	2.27381	0.10852	0.74796	0.90952	0.12338	0.23608
0.5	3.38435	1.69218	0.12530	1.01531	0.67687	0.14123	0.25628
0.7	4.54762	1.24660	0.12126	1.36429	0.49864	0.13438	0.23608
0.9	5.98299	0.93707	0.08832	1.79490	0.37483	0.08906	0.15541
	$(\hat{\rho}_1, \hat{\rho}_2) = (0.3, 0, 6)$			$(\hat{\rho}_1, \hat{\rho}_2) = (1, 1.5)$			

Furthermore, $(q_0; q_1, \ldots, q_d)$ in this paper jointly satisfy the basic adjoint relationship (BAR):

$$\int_{\mathbb{R}_+^d} \left(\frac{1}{2} \sum_{i,j} \hat{R}_{ij} \frac{\partial^2 f(x)}{\partial x_i \partial x_j} - \sum_k \hat{\theta}_k \frac{\partial f(x)}{\partial x_k} \right) q_0(x)\, \mathrm{d}x + \frac{1}{2} \sum_k \int_{F_k} \frac{\partial f}{\partial x_k}(x) q_k(x) \nu_k(\mathrm{d}x) = 0$$

for all $f \in C_b^2(\mathbb{R}_+^d)$. It can be checked that the coefficients for $\hat{Q}$ do not satisfy the conditions of Propositions 8 and 9 in [9] (see also [33]) which means the stationary density cannot be a product function. The numerical approach developed in [9] to compute the stationary distribution is applicable to our model.

Noticing that the marginal distribution $\hat{Q}_k(\infty)$ for each k is exponentially distributed from the one-dimensional reflected Brownian motion results, we readily obtain

$$\mathbb{E}[\hat{Q}_k(\infty)] = \frac{\hat{R}_{kk}}{2\hat{\theta}_k} \quad \text{and} \quad \mathrm{var}(\hat{Q}_k(\infty)) = \left(\frac{\hat{R}_{kk}}{2\hat{\theta}_k} \right)^2.$$

We apply the numerical algorithm in [9] to a model with two queues and compute some characteristics of the stationary distribution. Similar to [9], we compute the mean of the queueing limit and the correlation coefficients of $(\hat{W}_1, \hat{W}_2)$ and $(\hat{Q}_1(\infty), \hat{Q}_2(\infty))$, and examine the effect of the splitting probability parameter p_k. We fix $\lambda_0 = 1$, $\sigma_1 = \sigma_2 = 1$, $\|H\|_1 = 0.3$, and change the values of p_1. Observe that by equation (3.8), the value of μ_k changes accordingly as p_k changes. We also consider two scenarios of $(\hat{\rho}_1, \hat{\rho}_2)$, taking values (0.3, 0.6) and (1, 1.5). Notice that the correlation of $(\hat{W}_1, \hat{W}_2)$ from (3.11) is independent of the choice of $(\hat{\rho}_1, \hat{\rho}_2)$ by definition. Table 1 shows how the splitting probability p_1 and the parameters $(\hat{\rho}_1, \hat{\rho}_2)$ affect the correlation of $(\hat{Q}_1(\infty), \hat{Q}_2(\infty))$.

4. Parallel single-server queues with split Hawkes arrival processes and abandonment

In this section we consider the parallel single-server queues as described in the previous section but with an additional feature of abandonment, that is, each customer joining the queue has a patience time and will leave the queue if the patience time runs out before entering service. The arrival and service processes are modeled in the same way as the previous section. Patience times are assumed to be independent of the arrival and service processes as well as the splitting process. Let $\{\vartheta_{j,k,p}\}_{j \in \mathbb{Z}}$ represent the patience times for every customers with CDF G_k, $k = 1, \ldots, d$. The dependence on k may be interpreted as the impact of each queue upon patience, since their services may have different distributions. Of course, one may also consider the homogeneous scenario with patience times having the same distribution in all the

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

queues. In this model, variables and processes will be indexed with an additional subscript p to distinguish from the model in the previous section whenever necessary.

To give a rigorous definition of the model, we adapt the definition from [30] and introduce the *offered workload process*

$$\tilde{Z}_{k,p}(t) = \Theta_{k,p}(Q_k(0)) + \sum_{j\geq 1} \eta_{j,k}\mathbf{1}(\tau_{j,k} \leq t, \tilde{Z}_{k,p}(\tau_{j,k}-) < \vartheta_{j,k,p}) - B_{k,p}(t) \tag{4.1}$$

for $t > 0$, where $\{\tau_{j,k} : j \geq 1\}$ are the arrival times of the split Hawkes process $N_k(t)$,

$$\Theta_{k,p}(j) = \Theta_{k,p}(j-1) + \eta_{-j,k}\mathbf{1}(\Theta_{k,p}(j-1) < \vartheta_{-j,k,p}) \quad \text{for } j \geq 1,$$

with $\Theta_{k,p}(0) = 0$ representing the waiting time for the $(j+1)$th customers initially in the kth queue. Note that $\tilde{Z}_{k,p}(0) = \Theta_{k,p}(Q_k(0))$. The cumulative busy-time process is defined by

$$B_{k,p}(t) = \int_0^t \mathbf{1}(\tilde{Z}_{k,p}(s) > 0)\,\mathrm{d}s. \tag{4.2}$$

Different from the offered workload process, the observed workload process is defined by

$$\begin{aligned} Z_{k,p}(t) = \tilde{Z}_{k,p}(t) &+ \sum_{j\geq 1} \eta_{-j,k}\mathbf{1}(j \leq Q_k(0),\, \Theta_{k,p}(j-1) \geq \vartheta_{-j,k,p} > t) \\ &+ \sum_{j\geq 1} \eta_{j,k}\mathbf{1}(\tilde{Z}_{k,p}(\tau_{j,k}-) \geq \vartheta_{j,k,p} > t - \tau_{j,k} \geq 0), \end{aligned} \tag{4.3}$$

which retrospectively counts both the workload from customers that eventually receive service, $\tilde{Z}_{k,p}$, and from those that are currently in a queue but eventually renege, $\mathbf{1}(\cdot \geq \vartheta_{j,k,p} > \cdot)$, in contrast to the prospective definition for Z_k in (3.1).

Similarly, we define the observed queue length process as follows:

$$\begin{aligned} Q_{k,p}(t) = &\sum_{j=1}^{Q_k(0)} \big(\mathbf{1}(\Theta_{k,p}(j-1) < \vartheta_{-j,k,p},\, t < \Theta_{k,p}(j)) + \mathbf{1}(\Theta_{k,p}(j-1) \geq \vartheta_{-j,k,p} > t)\big) \\ &+ \sum_{j=1}^{N_k(t)} \big(\mathbf{1}(\tilde{Z}_{k,p}(\tau_{j,k}-) < \vartheta_{j,k,p},\, t < \tilde{Z}_{k,p}(\tau_{j,k})) + \mathbf{1}(\tilde{Z}_{k,p}(\tau_{j,k}-) \geq \vartheta_{j,k,p} > t - \tau_{j,k})\big), \end{aligned}$$

which counts the number of customers currently in the queue. Observe that the busy-time process in (4.2) also satisfies

$$B_{k,p}(t) = \int_0^t \mathbf{1}(Z_{k,p}(s) > 0)\,\mathrm{d}s = \int_0^t \mathbf{1}(\tilde{Z}_{k,p}(s) > 0)\,\mathrm{d}s = \int_0^t \mathbf{1}(Q_{k,p}(s) > 0)\,\mathrm{d}s$$

and the idle-time process is given by

$$I_{k,p}(t) := t - B_{k,p}(t) = \int_0^t \mathbf{1}(Q_{k,p}(s) = 0)\,\mathrm{d}s = \int_0^t \mathbf{1}(\tilde{Z}_{k,p}(s) = 0)\,\mathrm{d}s.$$

We also need the following definition generalizing the multidimensional reflection mapping in Definition 3.1, which is adapted from Appendix A.1 in [30].

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Definition 4.1. Let Γ be a $d \times d$ matrix of real entries and let $\mathtt{Q}$ be a reflection matrix in Definition 3.1. Then, for each $x \in \mathbb{D}^d$ with $x(0) \geq 0$, let $(z, y) \in \mathbb{D}^{2d}$ be a unique pair satisfying

(i) $z(t) = x(t) - \int_0^t \Gamma z(s)\,\mathrm{d}s + (\mathtt{I} - \mathtt{Q})y(t) \geq 0$ for all $t \geq 0$,

(ii) $y(0) = 0$, $y \in \mathbb{D}^d_\uparrow$ and $\int_0^\infty z_k(t)\,\mathrm{d}y_k(t) = 0$ for every k.

Furthermore, let u be the unique solution to the integral equation

$$u(t) + \int_0^t \Gamma u(s)\,\mathrm{d}s = x(t). \tag{4.4}$$

Define the mapping $\mathcal{M}_\Gamma : \mathbb{D}^d \to \mathbb{D}^d$ by $\mathcal{M}_\Gamma(x) = u$. Then

$$(z, y) = (\phi_\mathtt{Q}, \psi_\mathtt{Q})(\mathcal{M}_\Gamma(x)),$$

where $\phi_\mathtt{Q}$ and $\psi_\mathtt{Q}$ are the content and reflection operators, respectively, in Definition 3.1.

By [30, Lemma 3], $\mathcal{M}_\Gamma$ is Lipschitz-continuous with respect to uniform topology on every compact set. One can find from the integral equation (4.4) that

$$\begin{aligned} u(t) &= \mathcal{M}_\Gamma(x)(t) \\ &= x(0) + \int_0^t \mathrm{e}^{\Gamma(s-t)} x(\mathrm{d}s) \\ &= x(t) - \Gamma \int_0^t \mathrm{e}^{\Gamma(s-t)} x(s)\,\mathrm{d}s \\ &= x(t) - \sum_{j\geq 0} \Gamma^{j+1} \frac{(-1)^j}{j!} \int_0^t (t-s)^j x(s)\,\mathrm{d}s. \end{aligned}$$

In our model, $\Gamma = \mathrm{diag}(\gamma)$ is a diagonal matrix with $\gamma = (\gamma_k)_k \in \mathbb{R}^d$ on the diagonal, and we have

$$u_k(t) = x_k(t) - \gamma_k \int_0^t \mathrm{e}^{\gamma_k(s-t)} x_k(s)\,\mathrm{d}s$$

for each coordinate process. In this case, we simply write $u = \mathcal{M}_\gamma(x)$. If $x(t) = x_0 + at + \sigma B(t)$ in Definition 4.1, where B is a d-dimensional standard Brownian motion, $a \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^{d\times d}$, then $u = \mathcal{M}_\gamma(x)$ is the unique strong solution to the stochastic differential equation

$$\mathrm{d}u(t) = (a - \gamma u(t))\,\mathrm{d}t + \sigma\,\mathrm{d}B_t = -\gamma u(t)\,\mathrm{d}t + (a\,\mathrm{d}t + \sigma\,\mathrm{d}B_t), \quad u(0) = x_0.$$

This is well-defined and called an Ornstein–Uhlenbeck (OU) process. Moreover, $z = \phi(\mathcal{M}_\gamma(x))$ is the regulated (reflected) OU process. Recall that $\phi = \phi_0$ with the matrix $\mathtt{Q} \equiv 0$.

As in the previous section, let $\big(Q_p^{(n)}, Z_p^{(n)}\big) = \big(Q_{k,p}^{(n)}, Z_{k,p}^{(n)}\big)_k$ be the associated observed queue length and workload processes for the nth queueing system. We are interested in the diffusion-scaled observed processes in the heavy-traffic regime, and define

$$\hat{Q}_{k,p}^{(n)}(t) := \frac{1}{\sqrt{n}} Q_{k,p}^{(n)}(nt) \quad \text{and} \quad \hat{\tilde{Z}}_{k,p}^{(n)}(t) := \frac{1}{\sqrt{n}} \tilde{Z}_{k,p}^{(n)}(nt).$$

We make the following assumption on $G_k^{(n)}$ for the variables $\vartheta_{j,k,p}^{(n)}$.

Assumption 4.1. *Assume that $G_k^{(n)}$ is continuous, and for some $\gamma_k > 0$,*

$$\hat{G}_k^{(n)}(t) := \sqrt{n} G_k^{(n)}(\sqrt{n}t) \to \gamma_k t \quad \text{as } n \to \infty,$$

where the convergence holds uniformly on compacts (u.o.c.) over $\mathbb{R}_+$.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Note that in the case $\vartheta_{j,k,p}^{(n)} := n \times \vartheta_{j,k,p}$, that is, $G_k^{(n)}(t) = G_k(t/n)$ for some common CDF G_k, the assumption above is equivalent to G_k being differentiable at 0 with $\gamma_k = G'_k(0)$. This is the so-called critical case studied in [30].

Theorem 4.1. *Suppose that Assumptions 3.1–4.1 and the conditions in (2.5) and (2.8) hold, and that* $\hat{Q}^{(n)}(0) \Rightarrow \hat{Q}(0) \geq 0$. *Then*

$$\left(\hat{Q}_p^{(n)}, \hat{Z}_p^{(n)}\right) \Rightarrow (\hat{Q}_p, \hat{Z}_p) \quad \text{in } (\mathbb{D}^{2d}, J_1) \quad \text{as } n \to \infty,$$

with the limit

$$\hat{Q}_p = \phi(\mathcal{M}_\gamma(\hat{Q}(0) + \hat{W} - \hat{\theta}\mathfrak{e})) \quad \text{and} \quad \hat{Z}_p = \mu^{-1}\hat{Q}_p,$$

where $\hat{\theta}$ *and* $\hat{W}$ *are as given in Theorem 3.1.*

Here $\hat{Q}_p$ is a d-dimensional reflected OU process with initial value $\hat{Q}(0)$. Although the limit in Theorem 4.1 formally reduces to that in Theorem 3.1 if $\gamma = 0$ in Assumption 4.1, it is worth mentioning that the proof of Theorem 4.1 relies on Theorem 3.1, and in the case $\gamma > 0$,

$$(Z_p, I_p) \neq (\phi, \psi)(\mathcal{M}_\gamma(V \circ (Q(0) + N) - \mathfrak{e})),$$

in contrast to (Z, I) in (3.2).

5. A parallel server model with multivariate Hawkes arrivals

As noted in Section 2, a splitting Hawkes process is a multivariate Hawkes process with a particular conditional intensity process given by (2.2). In this section we present the limits for a parallel server model with a multivariate Hawkes arrival process, defined as a simple point process with conditional intensity process given by

$$\lambda_i(t) = \lambda_{i,0} + \sum_{j=1}^{d} \int_0^t H_{ij}(t-s)\, \mathrm{d}N_j(s), \quad i = 1, \ldots, d, \quad t \geq 0,$$

where $\lambda_0 = (\lambda_{i,0})_i$ is the baseline intensity vector, and $H = (H_{ij})_{ij}$ is the kernel matrix function with $H_{ij} : \mathbb{R}_+ \to \mathbb{R}_+$. The *non-explosion criterion* in [3] is given by $\int_0^t H_{ij}(s)\, \mathrm{d}s < \infty$, for all $t > 0$, under which the point process is well-defined. Given the multivariate Hawkes process as arrivals and the service process as defined in Section 3 in the parallel single-server queues, the queue length process and workload process can be defined in the same way.

We consider a sequence of such queueing systems in heavy traffic with index n, where the baseline intensity is $\lambda^{(n)}$ and the kernel function is H.

Assumption 5.1. *Suppose the following conditions hold:*

(i) $\lambda_0^{(n)} \to \lambda_0$ *as* $n \to \infty$,

(ii) $\int_0^\infty H_{ij}(t)\, \mathrm{d}t < \infty$ *and the spectral radius of the matrix* $\|H\|_1 = \left(\int_0^\infty H_{ij}(t)\, \mathrm{d}t\right)_{ij}$ *is less than 1,*

(iii) $\lim_{t\to\infty} \sqrt{t} \int_t^\infty H_{ij}(s)\, \mathrm{d}s = 0$ *for every i and j.*

We note in particular that the notation differs from Section 3: λ_0 is a vector and H is a matrix.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Under Assumption 5.1, it is shown (see [3, Theorem 2] and also Proposition 2.2) that

$$\check{N}^{(n)}(t) = \sqrt{n}\left(\frac{1}{n}N^{(n)}(nt) - (\mathtt{I} - \|H\|_1)^{-1}\lambda_0^{(n)}t\right) \Rightarrow \hat{N}(t) \tag{5.1}$$

in $(\mathbb{D}^d, J_1)$, where $\hat{N}$ is a d-dimensional Brownian motion with covariance function

$$\operatorname{cov}(\hat{N}_i(t), \hat{N}_j(s)) = (t \wedge s) \cdot \operatorname{ent}_{ij}\big((\mathtt{I} - \|H\|_1)^{-1} \cdot \operatorname{diag}((\mathtt{I} - \|H\|_1)^{-1}\lambda_0) \cdot (\mathtt{I} - \|H\|_1^{\top})^{-1}\big)$$

for $t, s \geq 0$, where the superscript $\top$ denotes the transpose of a matrix, comparing it with Proposition 2.1 Note that in this model setup, the kernel matrix H is assumed to be independent of n. The traffic intensity for the kth queue in the nth system is given by (compare with (3.3))

$$\rho_k^{(n)} = \operatorname{ent}_k\big((\mathtt{I} - \|H\|_1)^{-1}\lambda_0^{(n)}\big)\big/\mu_k^{(n)}.$$

The heavy-traffic condition will then imply that

$$(\mathtt{I} - \|H\|_1)^{-1}\lambda_0 = \mu.$$

Under Assumptions 3.1 and 5.1 together with the new traffic intensity above, and $\hat{Q}^{(n)}(0) \Rightarrow \hat{Q}(0) \geq 0$, we obtain

$$\big(\hat{Q}^{(n)}, \hat{Z}^{(n)}\big) \Rightarrow (\hat{Q}, \hat{Z}) \quad \text{in } (\mathbb{D}^{2d}, J_1) \quad \text{as } n \to \infty,$$

with the limit

$$\hat{Q} = \phi(\hat{Q}(0) + \hat{W} - \hat{\theta}\mathfrak{e}) \quad \text{and} \quad \hat{Z} = \mu^{-1}\hat{Q} = \big(\mu_k^{-1}\hat{Q}_k\big)_k,$$

where

$$\hat{\theta} = \mu\hat{\rho} = (\mu_k\hat{\rho}_k)_k,$$

and $\hat{W}$ is a d-dimensional Brownian motion with the following covariance function: for $t, s \geq 0$,

$$\begin{aligned} \operatorname{cov}(\hat{W}(t), \hat{W}(s)) &= (t \wedge s) \cdot \hat{R}_{kk'}, \\ (\hat{R}_{kk'})_{kk'} &= (\mathtt{I} - \|H\|_1)^{-1} \cdot \operatorname{diag}(\mu) \cdot \big(\mathtt{I} - \|H\|_1^{\top}\big)^{-1} + \operatorname{diag}\big(\mu^3\sigma^2\big). \end{aligned} \tag{5.2}$$

For the parallel server queueing model with abandonment as described in Section 4, suppose the arrival process is also a multivariate Hawkes process as described above. Let $Q_p^{(n)}$ and $Z_p^{(n)}$ be the queue length process and the workload process for the model, respectively, and suppose that Assumptions 3.1, 4.1, and 5.1 hold, and that $\hat{Q}^{(n)}(0) \Rightarrow \hat{Q}(0) \geq 0$; then

$$\big(\hat{Q}_p^{(n)}, \hat{Z}_p^{(n)}\big) \Rightarrow \big(\hat{Q}_p, \hat{Z}_p\big) \quad \text{in } (\mathbb{D}^{2d}, J_1) \quad \text{as } n \to \infty, \tag{5.3}$$

with the limit

$$\hat{Q}_p = \phi(\mathcal{M}_\gamma(\hat{Q}(0) + \hat{W} - \hat{\theta}\mathfrak{e})) \quad \text{and} \quad \hat{Z}_p = \mu^{-1}\hat{Q}_p,$$

where $\hat{W}$ is as given in (5.2).

We remark that the proofs for both models follow similar arguments by adapting those with a split Hawkes process, which we omit for brevity. However, we note that the split Hawkes processes as arrivals to the parallel server queues have a particular structure due to the splitting mechanism, while the multivariate Hawkes process has a matrix kernel $H = (H_{ij})$ (independent of n). It is important to highlight that the limit processes in both parallel server queueing models with a split Hawkes arrival process and a general multivariate Hawkes process are essentially of the same nature (with orthonormal reflections), except that the driving Brownian motions have different covariance functions (comparing (3.11) and (5.2)).

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

6. Proofs

6.1. Proof of Theorem 3.1

The proof follows from some modification of the arguments for the heavy-traffic convergence in single-server queues; see e.g. [7, Chapter 6]. We highlight the differences here for the parallel single-server queueing model.

Recall the diffusion-scaled processes $\hat{Q}_k^{(n)}$ and $\hat{Z}_k^{(n)}$ for the nth system from (3.1):

$$\begin{aligned}\hat{Z}_k^{(n)}(t) &= \frac{1}{\sqrt{n}} Z_k^{(n)}(nt) = \sqrt{n}\big(\bar{V}_k^{(n)}\big(\bar{Q}_k^{(n)}(0) + \bar{N}_k^{(n)}(t)\big) - \bar{B}_k^{(n)}(t)\big),\\ \hat{Q}_k^{(n)}(t) &= \frac{1}{\sqrt{n}} Q_k^{(n)}(nt) = \sqrt{n}\big(\bar{Q}_k^{(n)}(0) + \bar{N}_k^{(n)}(t) - \bar{U}_k^{(n)}\big(\bar{B}_k^{(n)}(t)\big)\big),\end{aligned} \tag{6.1}$$

where

$$\bar{Q}_k^{(n)}(0) = \frac{1}{n} Q_k^{(n)}(0)$$

and $\bar{N}_k^{(n)}$ is the process in (2.3). We will apply the FCLT established for the split Hawkes processes and renewal process, Proposition 2.2 and Remark 3.1 respectively, and the continuous mapping approach to the multidimensional reflection mapping; see e.g. [7, Theorems 6.1 and 7.2] and [32, Chapter 14.2]. The following lemma is a modification of the so-called *random change of time* result from [4, page 151], which is also called the *continuity of composition* in [32, Theorem 13.2.2]. In this version, each sub-counting process has a different time scaling, and all the limit processes are assumed to have continuous sample paths. The conditions are slightly different from those in [32, Chapter 14.2]. Recall that $\mathbb{C}_\uparrow^d = \mathbb{C}^d \cap \mathbb{D}_\uparrow^d$, where $\mathbb{C}^d$ denotes the space of $\mathbb{R}^d$-valued continuous functions.

Lemma 6.1. *Let* $\big(x^{(n)}, \chi^{(n)}\big) = \big(x_k^{(n)}, \chi_k^{(n)}\big)_k \in \mathbb{D}^d \times \mathbb{D}_\uparrow^d$ *and* $(x, \chi) = (x_k, \chi_k)_k \in \mathbb{C}^d \times \mathbb{C}_\uparrow^d$. *If*

$$\big(x^{(n)}, \chi^{(n)}\big) \to (x, \chi) \quad \text{in } (\mathbb{D}^{2d}, J_1) \quad \text{as } n \to \infty,$$

then

$$x^{(n)} \circ \chi^{(n)} = \big(x_k^{(n)} \circ \chi_k^{(n)}\big)_k \to (x_k \circ \chi_k)_k = x \circ \chi \quad \text{u.o.c.} \quad \text{as } n \to \infty.$$

Proof of Theorem 3.1. We have from (6.1), for every $k = 1, 2, \ldots, d$,

$$\begin{aligned}\hat{Q}_k^{(n)}(t) &= \sqrt{n}\bar{Q}_k^{(n)}(0) + \mu_k^{(n)}\sqrt{n}\left(\frac{\lambda_0^{(n)} p_k^{(n)}}{(1 - \|H\|_1)\mu_k^{(n)}} - 1\right)t\\ &\quad + \sqrt{n}\left(\bar{N}_k^{(n)}(t) - \frac{\lambda_0^{(n)} p_k^{(n)} t}{1 - \|H\|_1}\right) - \sqrt{n}\big(\bar{U}_k^{(n)}(s) - \mu_k^{(n)} s\big)\big|_{s=\bar{B}_k^{(n)}(t)}\\ &\quad + \mu_k^{(n)}\sqrt{n}\big(t - \bar{B}_k^{(n)}(t)\big)\\ &= \hat{Q}_k^{(n)}(0) - \hat{\theta}_k^{(n)} t + \check{N}_k^{(n)}(t) - \hat{U}_k^{(n)} \circ \bar{B}_k^{(n)}(t) + \mu_k^{(n)}\hat{I}_k^{(n)}(t),\end{aligned}$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

where

$$\hat{\theta}_k^{(n)} = \mu_k^{(n)} \cdot \sqrt{n}\big(1 - \rho_k^{(n)}\big) \quad \text{and} \quad \hat{I}_k^{(n)}(t) = \sqrt{n}\big(t - \bar{B}_k^{(n)}(t)\big),$$

$\rho_k^{(n)}$ is defined in (3.3), and $\check{N}_k^{(n)}(t)$ is defined in (2.7). Observe that what differs from a single-server queue is that the $\check{N}_k^{(n)}(t)$ are correlated Hawkes processes, which introduces dependence among the different queueing processes.

By the definition of $\hat{Z}^{(n)}$ in (6.1) and $\big(\hat{V}^{(n)}, \hat{U}^{(n)}\big)$ in (3.4), we also obtain

$$\hat{Z}^{(n)} - m^{(n)}\hat{Q}^{(n)} = \hat{V}^{(n)} \circ \big(\bar{Q}^{(n)}(0) + \bar{N}^{(n)}\big) + m^{(n)}\hat{U}^{(n)} \circ \bar{B}^{(n)}. \tag{6.2}$$

We now consider the convergence of various components in these representations.

Under Assumption 3.1 and the conditions in Proposition 2.2 for Hawkes processes, from (2.9) for $\check{N}^{(n)}$, (3.9) for $(\hat{V}^{(n)}, \hat{U}^{(n)})$ and their independences, we have the joint weak convergence

$$\big(\check{N}^{(n)}, \hat{V}^{(n)}, \hat{U}^{(n)}\big) \Rightarrow (\hat{N}, \hat{V}, \hat{U}) \quad \text{in } (\mathbb{D}^{3d}, J_1) \quad \text{as } n \to \infty,$$

where $\hat{N}$ is the Brownian motion in Proposition 2.1, $(\hat{V}, \hat{U})$ is the Brownian motion in (3.9) and independent of $\hat{N}$.

Moreover, it is easy to see that

$$\big(\bar{Q}^{(n)}(0), \bar{N}^{(n)}, \bar{B}^{(n)}\big) \to \left(0, \frac{\lambda_0 (p_k)_k}{1 - \|H\|_1}\mathfrak{e}, \mathfrak{e}\right) = (0, \mu\mathfrak{e}, \mathfrak{e}) \quad \text{u.o.c.}, \tag{6.3}$$

in probability as $n \to \infty$, where the identity (3.8) is used in the equality, and abusing notation, $\mathfrak{e}$ is a vector of identity functions.

Thus, applying the continuous mapping theorem and the composition mapping (Lemma 6.1), we obtain the joint convergence

$$\big(\check{N}_k^{(n)}, \hat{V}_k^{(n)}\big(\bar{Q}_k^{(n)}(0) + \bar{N}_k^{(n)}\big), \hat{U}_k^{(n)}\big(\bar{B}_k^{(n)}\big)\big)_k \Rightarrow \big(\hat{N}_k, \hat{V}_k(\mu_k \cdot), \hat{U}_k\big)_k \tag{6.4}$$

in $(\mathbb{D}^{3d}, J_1)$ as $n \to \infty$.

Now, applying the continuous mapping theorem to the multi-dimensional reflection mapping in Definition 3.1 and using the convergence results in (6.3) and (6.4), together with the fact $\hat{U}_k(t) = -\mu_k \hat{V}_k(\mu_k t)$ from (3.9), we obtain the joint convergence of $\big(\hat{Q}^{(n)}, \hat{\tilde{Z}}^{(n)}\big)$ in (3.10).

Finally, since $\hat{N}$ and $\hat{U}$ are independent, it is easy to check that $\hat{W} = \hat{N} - \hat{U}$ is a Brownian motion with the covariance function given in (3.11). This completes the proof. □

6.2. Proof of Theorem 4.1

We adapt the proof idea in [30] and highlight the main differences in the proof. For the system indexed by n, recall that the diffusion-scaled processes are defined by

$$\hat{Q}_{k,p}^{(n)}(t) = \frac{1}{\sqrt{n}} Q_{k,p}^{(n)}(nt), \quad \hat{Z}_{k,p}^{(n)}(t) = \frac{1}{\sqrt{n}} Z_{k,p}^{(n)}(nt), \quad \hat{\tilde{Z}}_{k,p}^{(n)}(t) = \frac{1}{\sqrt{n}} \tilde{Z}_{k,p}^{(n)}(nt),$$

and

$$\hat{I}_{k,p}^{(n)}(t) = \sqrt{n}\big(t - \bar{B}_{k,p}^{(n)}(t)\big).$$

Theorem 4.1 is proved following the procedure from [30], where we start from the analysis of the offered workload process in (4.1). Specifically, the proof proceeds in the following steps.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Step 1. The weak convergence of the diffusion-scaled process $\big(\hat{\tilde{Z}}_p^{(n)}, \hat{I}_p^{(n)}\big)$ in Theorem 6.1, by relating to $\big(\hat{Z}^{(n)}, \hat{I}^{(n)}\big)$ in the corresponding model without abandonment.

Step 2. The following asymptotic equivalence properties: for every $T > 0$ and k, as $n \to \infty$, we have

$$\sup_{t\le T}\big|\hat{Z}_{k,p}^{(n)}(t) - \hat{\tilde{Z}}_{k,p}^{(n)}(t)\big| \Rightarrow 0 \quad \text{and} \quad \sup_{t\le T}\big|\hat{\tilde{Z}}_{k,p}^{(n)}(t) - m_k^{(n)}\hat{Q}_{k,p}^{(n)}(t)\big| \Rightarrow 0.$$

They follow the same procedure as [30], so their proofs are given in the Appendix for completeness.

Step 3. Completing the proof: given the convergence results in the two steps, the joint convergence in Theorem 4.1 follows immediately.

Therefore we focus only on the proof of the following theorem.

Theorem 6.1. *Under the assumptions of Theorem* 4.1,

$$\big(\hat{\tilde{Z}}_p^{(n)}, \hat{I}_p^{(n)}\big) \Rightarrow \big(\hat{Z}_p, \hat{I}_p\big) \quad \textit{in } (\mathbb{D}^{2d}, J_1) \quad \textit{as } n \to \infty, \tag{6.5}$$

with the limit

$$\big(\hat{Z}_p, \hat{I}_p\big) := (\phi_0, \psi_0)\mathcal{M}_\gamma(\hat{Z} + \hat{I}) = m\,(\phi_0, \psi_0)\mathcal{M}_\gamma(\hat{Q}(0) + \hat{W} - \hat{\theta}\mathfrak{e})$$

where $\hat{W}$ and $\hat{\theta}$ are as given in Theorem 3.1.

The proof of Theorem 6.1 uses the following two lemmas. In the proof we need Theorem 3.1, and it helps to rewrite $Z_k^{(n)}(t)$ in (3.1) in the corresponding queueing model without abandonment as

$$Z_k^{(n)}(t) = Z_k^{(n)}(0) + \sum_{j\ge 1}\eta_{j,k}^{(n)}\mathbf{1}\big(\tau_{j,k}^{(n)} \le t\big) - B_k^{(n)}(t),$$

with

$$Z_k^{(n)}(0) = \sum_{j\ge 1}\eta_{-j,k}^{(n)}\mathbf{1}\big(j \le Q_k^{(n)}(0)\big).$$

The diffusion-scaled process $\hat{Z}^{(n)}$ is defined in the same way. The corresponding convergence results for $\big(\hat{Q}^{(n)}, \hat{Z}^{(n)}\big)$ in Theorem 3.1 will be used.

Define

$$\hat{M}_{k,p,1}^{(n)}(t) = \frac{1}{\sqrt{n}}M_{k,p,1}^{(n)}([nt]) \quad \text{and} \quad \hat{M}_{k,p,2}^{(n)}(t) = \frac{1}{\sqrt{n}}M_{k,p,2}^{(n)}([nt])$$

with

$$M_{k,p,1}^{(n)}(m) = \sum_{j=1}^{m}\big(\eta_{j,k}^{(n)} - m_k^{(n)}\big)\mathbf{1}\big(\vartheta_{j,k,p}^{(n)} \le u\big)\big|_{u=\tilde{Z}_{k,p}^{(n)}(\tau_{j,k}^{(n)}-)},$$

$$M_{k,p,2}^{(n)}(m) = \sum_{j=1}^{m}\big(\mathbf{1}\big(\vartheta_{j,k,p}^{(n)} \le u\big) - G_k^{(n)}(u)\big)\big|_{u=\tilde{Z}_{k,p}^{(n)}(\tau_{j,k}^{(n)}-)}.$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Lemma 6.2. *For every $T > 0$ and every k,*

$$\sup_{t\le T}\big|\hat{M}^{(n)}_{k,p,1}(t)\big| \Rightarrow 0 \quad \textit{and} \quad \sup_{t\le T}\big|\hat{M}^{(n)}_{k,p,2}(t)\big| \Rightarrow 0 \quad \textit{as } n\to\infty. \tag{6.6}$$

Proof. It is easy to see that $\big\{M^{(n)}_{k,p,1}(m)\big\}_{m\ge1}$ is an $\big\{\mathscr{F}^{(n)}_{k,p,m}\big\}_{m\ge1}$-adapted square-integrable martingale, where for $m \ge 1$

$$\mathscr{F}^{(n)}_{k,p,m} = \tilde{Z}^{(n)}_{k,p}(0) \vee \sigma\big\{\eta^{(n)}_{j,k},\, \vartheta^{(n)}_{j,k,p},\, \tau^{(n)}_{j,k}\big\}_{1\le j\le m} \vee \sigma\big\{\tau^{(n)}_{m+1,k}\big\}.$$

Applying Doob's maximal inequality [27, Theorem VII.3.3], we have

$$\begin{aligned}
\mathbb{E}\bigg[\sup_{t\le T}\big(\hat{M}^{(n)}_{k,p,1}(t)\big)^2\bigg] &\le 4\mathbb{E}\big[\big(\hat{M}^{(n)}_{k,p,1}(T)\big)^2\big] \\
&= \frac{4}{n}\mathrm{var}\big(\eta^{(n)}_k\big)\sum_{j=1}^{[nT]}\mathbb{E}\big[G^{(n)}_k\big(\tilde{Z}^{(n)}_{k,p}\big(\tau^{(n)}_{j,,k}-\big)\big)\big] \\
&\le 4\mathrm{var}\big(\eta^{(n)}_k\big)\frac{[nT]}{n}\bigg(\frac{1}{\sqrt{n}}\hat{G}^{(n)}_k(K_0) + \mathbb{P}\bigg(\sup_{\bar{N}^{(n)}_k(t)\le T}\hat{\tilde{Z}}^{(n)}_{k,p}(t)\ge K_0\bigg)\bigg)
\end{aligned}$$

for every T and K_0. Together with Proposition 2.2 for $\bar{N}^{(n)}_k$, the fact that

$$0 \le \hat{\tilde{Z}}^{(n)}_{k,p}(t) \le \hat{Z}^{(n)}_{k,p}(t) \le \hat{Z}^{(n)}_k(t) \quad \text{for all } t\ge 0 \tag{6.7}$$

and Theorem 3.1 is established for $\hat{Z}^{(n)}_k$, we can establish the weak convergence result: $\sup_{t\le T}\big|\hat{M}^{(n)}_{k,p,1}(t)\big| \Rightarrow 0$ as $n\to\infty$. A similar procedure can be used to prove the convergence for $\hat{M}^{(n)}_{k,p,2}$. □

Recall the following martingale representations associated with Hawkes processes [3, 23]:

$$X^{(n)}(t) := N^{(n)}(t) - \int_0^t \lambda^{(n)}(s)\,\mathrm{d}s \quad \text{and} \quad X^{(n)}_k(t) := N^{(n)}_k(t) - \int_0^t \lambda^{(n)}_k(s)\,\mathrm{d}s \tag{6.8}$$

are martingales adapted to the filtrations generated by $N^{(n)}$ and $N^{(n)}_k$, respectively. Denote

$$\bar{X}^{(n)}_k(t) := \frac{1}{n}X^n_k(nt) = \bar{N}^{(n)}_k(t) - \int_0^t \bar{\lambda}^{(n)}_k(s)\,\mathrm{d}s \quad \text{and} \quad \bar{X}^{(n)}(t) = \frac{1}{n}X^{(n)}(nt).$$

Lemma 6.3. *For every $T > 0$ and every k,*

$$\sup_{t\le T}\bigg|\int_0^t \hat{G}^{(n)}_k\big(\hat{\tilde{Z}}^{(n)}_{k,p}(s-)\big)\big(\mathrm{d}\bar{N}^{(n)}_k(s) - \mathbb{E}\big[\bar{\lambda}^{(n)}_k(s)\big]\,\mathrm{d}s\big)\bigg| \Rightarrow 0 \quad \textit{as } n\to\infty.$$

Proof. Using the martingales in (6.8), the integral in the lemma can be split into two martingale integrals and a third component with bounded variation, and we show that each term

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

converges to 0 uniformly on $[0, T]$. Specifically, we have

$$\begin{aligned}
&\int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\big(\mathrm{d}\bar{N}_k^{(n)}(s) - \mathbb{E}\big[\bar{\lambda}_k^{(n)}(s)\big]\,\mathrm{d}s\big)\\
&\quad= \int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\big(\mathrm{d}\bar{X}_k^{(n)}(s) + p_k^{(n)}\big(\bar{\lambda}^{(n)}(s) - \mathbb{E}\big[\bar{\lambda}^{(n)}(s)\big]\big)\,\mathrm{d}s\big)\\
&\quad= \int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\,\mathrm{d}\bar{X}_k^{(n)}(s) + p_k^{(n)}\int_0^t \mathrm{d}\bar{X}^{(n)}(r)\bigg(n\int_r^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\varphi(n(s-r))\,\mathrm{d}s\bigg)\\
&\quad= \int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\big(\mathrm{d}\bar{X}_k^{(n)}(s) + p_k^{(n)}\|\varphi\|_1\,\mathrm{d}\bar{X}^{(n)}(s)\big)\\
&\qquad+ p_k^{(n)}\int_0^t \mathrm{d}\bar{X}^{(n)}(s)\bigg(n\int_s^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(r-)\big)\varphi(n(r-s))\,\mathrm{d}r - \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\|\varphi\|_1\bigg),
\end{aligned} \tag{6.9}$$

where we make use of (2.2) and the following identity from [3]:

$$\lambda^{(n)}(t) - \mathbb{E}\big[\lambda^{(n)}(t)\big] = \int_0^t \varphi(t-s)\,\mathrm{d}X^{(n)}(s),$$

and recall that $\varphi = \sum_{j\ge1} H^{*j}$ is the renewal function of H.

Since $X_k^{(n)}$, $X^{(n)}$ are martingales with quadratic variation $N_k^{(n)}$ and $N^{(n)}$, respectively, we have from Doob's maximal inequality that

$$\begin{aligned}
&\mathbb{E}\bigg[\sup_{t\in[0,T]}\bigg(\int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\,\mathrm{d}\bar{X}_k^{(n)}(s)\bigg)^2\bigg]\\
&\quad\le 4\mathbb{E}\bigg[\bigg(\int_0^T \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\,\mathrm{d}\bar{X}_k^{(n)}(s)\bigg)^2\bigg] = \frac{4}{n}\mathbb{E}\bigg[\int_0^T \big(\hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\big)^2\bar{\lambda}_k^{(n)}(s)\,\mathrm{d}s\bigg]\\
&\quad\le \frac{4}{n}\big(\hat{G}_k^{(n)}(K_0)\big)^2\mathbb{E}\bigg[\int_0^T \bar{\lambda}_k^{(n)}(s)\,\mathrm{d}s\bigg] + 4\mathbb{E}\bigg[\int_0^T \bar{\lambda}_k^{(n)}(s)\,\mathrm{d}s;\ \sup_{s\le T}\hat{\bar{Z}}_{k,p}^{(n)}(s) > K_0\bigg]
\end{aligned}$$

for every $K_0 > 0$, where $\hat{G}_k^{(n)}(z) \le \sqrt{n}$ by definition. Given (6.7), we have

$$\sup_{t\le T}\bigg|\int_0^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\,\mathrm{d}\bar{X}_k^{(n)}(s)\bigg| \Rightarrow 0 \quad \text{as } n\to\infty.$$

One can prove the uniform convergence of the second term in (6.9) similarly.

For the last integral in (6.9), notice that $n\varphi(n(s-r))\,\mathrm{d}s$ degenerates to a Dirac measure at r as $n\to\infty$ and $\bar{X}^{(n)}$ has bounded variation on $[0, T]$. For arbitrary $\delta_n > 0$, if $t - s > \delta_n$, we have

$$\begin{aligned}
&n\int_s^t \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(r-)\big)\varphi(n(r-s))\,\mathrm{d}r - \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\|\varphi\|_1\\
&\quad\le n\int_s^{s+\delta_n}\big|\hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(r-)\big) - \hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(s-)\big)\big|\varphi(n(r-s))\,\mathrm{d}r\\
&\qquad+ 2\sup_{u\le T}\hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(u)\big)\int_{s+\delta_n}^{\infty} n\varphi(n(r-s))\,\mathrm{d}r\\
&\quad\le \sup_{\substack{0<v<u\le T\\ u-v\le\delta_n}}\big|\hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(u)\big) - \hat{G}_{k,p}^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(v)\big)\big|\cdot\|\varphi\|_1 + 2\sup_{u\le T}\hat{G}_k^{(n)}\big(\hat{\bar{Z}}_{k,p}^{(n)}(u)\big)\int_{n\delta_n}^{\infty}\varphi(u)\,\mathrm{d}u.
\end{aligned}$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

On the other hand, if $t-s<\delta_n$, we have

$$n\int_s^t \hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(r-)\big)\varphi(n(r-s))\,\mathrm{d}r - \hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(s-)\big)\|\varphi\|_1 \le \sup_{u\le T}\hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(u)\big)\cdot\|\varphi\|_1.$$

Plugging these into the last integral in (6.9), we obtain the following bound:

$$\begin{aligned}
&\sup_{\substack{0<v<u\le T\\ u-v\le\delta_n}}\big|\hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(u)\big)-\hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(v)\big)\big|\cdot\|\varphi\|_1\cdot\int_0^T\big(\bar{N}^{(n)}(\mathrm{d}s)+\bar{\lambda}^{(n)}(\mathrm{d}s)\big)\\
&\quad+2\int_{n\delta_n}^{\infty}\varphi(u)\cdot\sup_{u\le T}\hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(u)\big)\cdot\int_0^T\big(\bar{N}^{(n)}(\mathrm{d}s)+\bar{\lambda}^{(n)}(\mathrm{d}s)\big)\\
&\quad+\sup_{u\le T}\hat{G}_k^{(n)}\big(\hat{\tilde{Z}}_{k,p}^{(n)}(u)\big)\cdot\|\varphi\|_1\cdot\sup_{\substack{0<v<u\le T\\ u-v\le\delta_n}}\left(\big|\bar{N}^{(n)}(u)-\bar{N}^{(n)}(v)\big|+\int_v^u\bar{\lambda}^{(n)}(s)\,\mathrm{d}s\right).
\end{aligned}$$

We next need the following result. For every $T>0$ and $\delta_n\to 0$ with $\sqrt{n}\delta_n\to 0$, we have for every k

$$\sup_{\substack{0<s<t<T\\ t-s\le\delta_n}}\big|\hat{\tilde{Z}}_{k,p}^{(n)}(t)-\hat{\tilde{Z}}_{k,p}^{(n)}(s)\big|\Rightarrow 0\quad\text{as } n\to\infty. \tag{6.10}$$

It follows from their definitions that

$$\big|\hat{\tilde{Z}}_{k,p}^{(n)}(t)-\hat{\tilde{Z}}_{k,p}^{(n)}(s)\big|\le\big|\hat{Z}_k^{(n)}(t)-\hat{Z}_k^{(n)}(s)\big|+2\sqrt{n}\delta_n$$

for every $t>s>0$ with $t-s\le\delta_n$. Therefore the convergence of $\hat{\tilde{Z}}_k^{(n)}$ in Theorem 3.1 and the fact that $\hat{Z}_k^{(n)}\in\mathbb{C}$ can be applied to show the convergence property in (6.10).

Now we continue with the last integral in (6.9), taking $\delta_n=n^{-2/3}$ and then $\sqrt{n}\delta_n\to 0$, $n\delta_n\to\infty$, by (6.10); we prove the uniform convergence of the last piece and finish the proof.

Now we are ready to prove Theorem 6.1 by making use of Lemmas 6.2 and 6.3.

Proof of Theorem 6.1. We claim that as $n\to\infty$,

$$\big(\hat{Z}_k^{(n)}(t)-\hat{I}_k^{(n)}(t)\big)-\left(\hat{\tilde{Z}}_{k,p}^{(n)}(t)+\gamma_k\int_0^t\hat{\tilde{Z}}_{k,p}^{(n)}(s)\,\mathrm{d}s-\hat{I}_{k,p}^{(n)}(t)\right)\Rightarrow 0\quad\text{u.o.c.} \tag{6.11}$$

for every k. The claim can also be rephrased as

$$\hat{\tilde{Z}}_p^{(n)}(t)+\mathrm{diag}(\gamma)\int_0^t\hat{\tilde{Z}}_p^{(n)}(s)\,\mathrm{d}s=\big(\hat{Z}^{(n)}(t)-\hat{I}^{(n)}(t)+\hat{\varepsilon}_p^{(n)}(t)\big)+\hat{I}_p^{(n)}(t), \tag{6.12}$$

where $\hat{I}_p^{(n)}$ is the regulator of $\hat{\tilde{Z}}_p^{(n)}$ satisfying the conditions of Definition 4.1 with the reflection matrix $\mathcal{Q}\equiv 0$, and $\hat{\varepsilon}_p^{(n)}$ is the error term which converges weakly to 0 u.o.c. In other words,

$$\big(\hat{\tilde{Z}}_p^{(n)},\hat{I}_p^{(n)}\big)=(\phi_0,\psi_0)\big(\mathcal{M}_\gamma\big(\hat{Z}^{(n)}-\hat{I}^{(n)}+\hat{\varepsilon}_p^{(n)}\big)\big).$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

We have shown in Section 6.1 that $\hat{Z}^{(n)} - \hat{I}^{(n)} \Rightarrow m(\hat{Q}(0) - \hat{\theta}\mathfrak{e} + W)$. By Theorem 3.1 we obtain joint convergence in (6.5).

To obtain (6.11), denote

$$\hat{M}^{(n)}_{k,p,0} := \frac{1}{\sqrt{n}} \sum_{j\geq 1} \eta^{(n)}_{-j,k} \mathbf{1}\big(j \leq Q^{(n)}_k(0),\, \vartheta^{(n)}_{-j,k,p} \leq \Theta^{(n)}_{k,p}(j-1)\big).$$

A simple calculation gives

$$\begin{aligned}
\text{LHS of (6.11)} &= \hat{M}^{(n)}_{k,p,0} + \frac{1}{\sqrt{n}} \sum_{j=1}^{N^{(n)}_k(nt)} \eta^{(n)}_{j,k} \mathbf{1}\big(\vartheta^{(n)}_{j,k,p} \leq \tilde{Z}^{(n)}_{k,p}(\tau^{(n)}_{j,k}-)\big) - \gamma_k \int_0^t \hat{\tilde{Z}}^{(n)}_{k,p}(s)\,\mathrm{d}s \\
&= \hat{M}^{(n)}_{k,p,0} + \hat{M}^{(n)}_{k,p,1}\big(\bar{N}^{(n)}_k(t)\big) + \hat{M}^{(n)}_{k,p,2}\big(\bar{N}^{(n)}_k(t)\big) \\
&\quad + m^{(n)}_k \int_0^t \hat{G}^{(n)}_k\big(\hat{\tilde{Z}}^{(n)}_{k,p}(s-)\big)\big(\mathrm{d}\bar{N}^{(n)}_k(s) - \mathbb{E}\big[\bar{\lambda}^{(n)}_k(s)\big]\,\mathrm{d}s\big) \\
&\quad - \lambda^{(n)}_0 m^{(n)}_k p^{(n)}_k \int_0^t \hat{G}^{(n)}_k\big(\hat{\tilde{Z}}^{(n)}_{k,p}(s)\big)\Big(\int_{ns}^{\infty} \varphi(u)\,\mathrm{d}u\Big)\,\mathrm{d}s \\
&\quad + \int_0^t \big(\rho^{(n)}_k \hat{G}^{(n)}_k(u) - \gamma_k u\big)\big|_{u=\hat{\tilde{Z}}^{(n)}_{k,p}(s)}\,\mathrm{d}s,
\end{aligned} \tag{6.13}$$

where $\bar{\lambda}^{(n)}_k(t) = \lambda^{(n)}_k(nt)$ and we use the expression of $\mathbb{E}\big[\bar{\lambda}^{(n)}_k(t)\big]$ in (2.4). It is thus sufficient to show that all terms on the right-hand side of (6.13) converge weakly to 0 u.o.c.

For the term $\hat{M}^{(n)}_{k,p,0}$, since $\Theta^{(n)}_{k,p}(j) \leq Z^{(n)}_k(0)$ for all $j \leq Q^{(n)}_k(0)$, we have

$$\mathbb{E}\big[\hat{M}^{(n)}_{k,p,0}; \hat{Z}^{(n)}_k(0) \leq K_0 \,\big|\, Q^{(n)}_k(0)\big] \leq \bar{Q}^{(n)}_k(0) m^{(n)}_k \hat{G}^{(n)}_k(K_0) \tag{6.14}$$

for every $K_0 > 0$. The Markov inequality can be applied to show that $\hat{M}^{(n)}_{k,p,0} \Rightarrow 0$ as $n \to \infty$.

Now, given Theorem 3.1 for $\hat{Z}^{(n)}_k$ and (6.7), the desired convergences of the last two terms in (6.13) can be checked directly. Further, applying Lemma 6.2, Lemma 6.3, and Proposition 2.1, we can prove the remaining terms and finish the proof. □

Appendix A. Additional proofs

A.1 Proofs for the asymptotic equivalence properties

This section is dedicated to the proofs of the asymptotic equivalence properties in Step 2 of the proof of Theorem 4.1. The arguments adapt those in [30] with slight modifications to illustrate the role of the split Hawkes processes. We provide the details for completeness.

Lemma A.1. *For every $T > 0$, we have for every k*

$$\sup_{t\leq T} \big|\hat{Z}^{(n)}_{k,p}(t) - \hat{\tilde{Z}}^{(n)}_{k,p}(t)\big| \Rightarrow 0 \quad \text{as } n \to \infty.$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

Proof. Considering the difference between $\hat{Z}_{k,p}^{(n)}$ and $\hat{\tilde{Z}}_{k,p}^{(n)}$, we show that the associated additional terms in (4.3) converge weakly to 0 u.o.c. For every $t > 0$, observe that

$$\frac{1}{\sqrt{n}} \sum_{j \geq 1} \eta_{-j,k}^{(n)} \mathbf{1}\big(j \leq Q_k^{(n)}(0),\, \Theta_{k,p}^{(n)}(j-1) \geq \vartheta_{-j,k,p}^{(n)} > nt\big) \leq \hat{M}_{k,p,0}^{(n)}.$$

Thus, using (6.14), we can show that the first term on the initial quantities converges to zero. On the other hand, for every $T, K_0 > 0$, on the set $\big\{\sup_{t \leq T} \hat{Z}_k^{(n)}(t) \leq K_0\big\}$ we have for every $t < T$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{j \geq 1} \eta_{j,k}^{(n)} \mathbf{1}\big(\tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}-\big) \geq \vartheta_{j,k,p}^{(n)} > nt - \tau_{j,k}^{(n)} \geq 0\big) \\
&\quad \leq \frac{1}{\sqrt{n}} \sum_{j \geq 1} \eta_{j,k}^{(n)} \mathbf{1}\big(\vartheta_{j,k,p}^{(n)} \leq \tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}-\big)\big) \mathbf{1}\big(nt \geq \tau_{j,k}^{(n)} > nt - \sqrt{n} K_0\big) \\
&\quad \leq \big(\big(\hat{M}_{k,p,1}^{(n)}(u) - \hat{M}_{k,p,1}^{(n)}(v)\big) + m_k^{(n)}\big(\hat{M}_{k,p,2}^{(n)}(u) - \hat{M}_{k,p,2}^{(n)}(v)\big) \\
&\qquad + m_k^{(n)} \hat{G}_k^{(n)}(K_0)(u-v)\big)\big|_{u=\bar{N}_k^{(n)}(t),\, v=\bar{N}_k^{(n)}(t-K_0/\sqrt{n})}.
\end{aligned}$$

Lemma 6.2 together with Proposition 2.1 proves the convergence. □

Lemma A.2. *For every $T > 0$ and every k,*

$$\sup_{t \leq T} \big|\hat{\tilde{Z}}_{k,p}^{(n)}(t) - m_k^{(n)} \hat{Q}_{k,p}^{(n)}(t)\big| \Rightarrow 0 \quad \text{as } n \to \infty.$$

Proof. Let $\varsigma_{k,p}^{(n)}(t)$ be the right-continuous increasing version of the arrival time of the customer in service at time t if the server is busy and $\varsigma_{k,p}^{(n)}(t) = t$ if the server is idle. By the same argument as in [30], it can be shown that $\big(t - \bar{\varsigma}_k^{(n)}(t)\big) \Rightarrow 0$ for the nth system, where $\bar{\varsigma}_{k,p}^{(n)}(t) = n^{-1} \varsigma_{k,p}^{(n)}(nt)$.

On the set $\big\{\varsigma_{k,p}^{(n)}(t) > 0\big\}$,

$$\begin{aligned}
&\tilde{Z}_{k,p}^{(n)}(t) - m_k^{(n)} Q_{k,p}^{(n)}(t) \\
&\quad = \sum_{j \geq 1} \big(\eta_{j,k}^{(n)} - m_k^{(n)}\big) \mathbf{1}\big(\varsigma_{k,p}^{(n)}(t) < \tau_{j,k}^{(n)} \leq t\big) \\
&\qquad + \sum_{j \geq 1} \big(\tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}\big) + \tau_{j,k}^{(n)} - t - m_k^{(n)}\big) \mathbf{1}\big(\varsigma_{k,p}^{(n)}(t) = \tau_{j,k}^{(n)} \leq t\big) \\
&\qquad - \sum_{j \geq 1} \big(\eta_{j,k}^{(n)} - m_k^{(n)}\big) \mathbf{1}\big(\varsigma_{k,p}^{(n)}(t) < \tau_{j,k}^{(n)} \leq t,\, \vartheta_{j,k,p}^{(n)} \leq \tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}-\big)\big) \\
&\qquad - m_k^{(n)} \sum_{j \geq 1} \mathbf{1}\big(\varsigma_{k,p}^{(n)}(t) < \tau_{j,k}^{(n)} \leq t,\, \vartheta_{j,k,p}^{(n)} \leq \tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}-\big),\, t < \tau_{j,k}^{(n)} + \vartheta_{j,k,p}^{(n)}\big),
\end{aligned}$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

where the last term is the number of customers currently in a queue and eventually abandoning the system without receiving service, which is less than

$$\sum_{j\ge 1}\mathbf{1}\big(\vartheta_{j,k,p}^{(n)}\le \tilde{Z}_{k,p}^{(n)}\big(\tau_{j,k}^{(n)}-\big),\ \varsigma_{k,p}^{(n)}(t)<\tau_{j,k}^{(n)}\le t\big)$$

$$=\int_{\varsigma_{k,p}^{(n)}(t)}^{t}\big(M_{k,p,1}^{(n)}(\mathrm{d}s)+G_k^{(n)}\big(\tilde{Z}_{k,p}^{(n)}(s-)\big)\,\mathrm{d}N_k^{(n)}(s)\big).$$

We thus have for the nth system, for $t\le T$,

$$\begin{aligned}&\big|\hat{\tilde{Z}}_{k,p}^{(n)}(t)-m_k^{(n)}\hat{Q}_{k,p}^{(n)}(t)\big|\\&\quad\le\big|\hat{V}_k^{(n)}(t)-\hat{V}_k^{(n)}\big(\bar{\varsigma}_{k,p}^{(n)}(t)\big)\big|+m_k^{(n)}\big(\bar{N}_k^{(n)}(t)-\bar{N}_k^{(n)}\big(\bar{\varsigma}_{k,p}^{(n)}(t)\big)\big)\sup_{z\le\sup_{s\le T}\hat{\tilde{Z}}_{k,p}^{(n)}(s)}\hat{G}_k^{(n)}(z)\\&\qquad+\frac{3}{\sqrt{n}}m_k^{(n)}+\sup_{j\le n\bar{N}_k^{(n)}(T)}\frac{1}{\sqrt{n}}\eta_{j,k}^{(n)}+2\sup_{s\le\bar{N}_k^{(n)}(T)}\big|\hat{M}_{k,p,1}^{(n)}(s)\big|\\&\qquad+2m_k^{(n)}\sup_{s\le\bar{N}_k^{(n)}(T)}\big|\hat{M}_{k,p,2}^{(n)}(t)\big|,\end{aligned}$$

where $\hat{V}_k^{(n)}$ is the process in (3.4). Given Lemma 6.2 and the fact that $\hat{V}_k^{(n)}$ and $\bar{N}_k^{(n)}$ both have continuous limits, the lemma is proved. □

A.2 Proof of Proposition 2.2

Proof. For every $k\ge 1$, given (2.4) and the fact $\|\varphi\|_1+1=1/(1-\|H\|_1)$, it is sufficient to show that

$$\sqrt{n}\bigg(\frac{t}{1-\|H\|_1}-\int_0^t\bigg(1+\int_0^{nu}\varphi(v)\,\mathrm{d}v\bigg)\,\mathrm{d}u\bigg)=\sqrt{n}\int_0^t\int_{nu}^{\infty}\varphi(v)\,\mathrm{d}v\,\mathrm{d}u\to 0.\tag{A.1}$$

For every $\varepsilon\in(0,(1-\|H\|_1)/2)$, denote

$$H_\varepsilon(t):=H(t)+\varepsilon(1+t)^{-3/2}\quad\text{for all } t>0$$

and let φ_ε be the associated renewal function, so $\|H_\varepsilon\|_1=\|H\|_1+2\varepsilon\in(0,1)$. Moreover, given (2.8),

$$H_\varepsilon(t)\ge H(t),\quad \varphi_\varepsilon(t)\ge\varphi(t)\quad\text{and}\quad t^{1/2}\int_t^{\infty}H_\varepsilon(s)\,\mathrm{d}s\to 2\varepsilon\quad\text{as } t\to\infty.$$

Thus we have from Karamata's Tauberian theorem (see [6, Theorem 1.7.1]) that

$$\frac{1}{z}\big(\|H_\varepsilon\|_1-\hat{H}_\varepsilon(z)\big)=\int_0^{\infty}\mathrm{e}^{-zt}\bigg(\int_t^{\infty}H_\varepsilon(s)\,\mathrm{d}s\bigg)\,\mathrm{d}t\sim 2\Gamma\bigg(\frac{1}{2}\bigg)\varepsilon z^{-1/2}\quad\text{as } z\to 0+,$$

where

$$\hat{H}_\varepsilon(z):=\int_0^{\infty}\mathrm{e}^{-zt}H_\varepsilon(t)\,\mathrm{d}t$$

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

denotes the Laplace transform of H_ε. It also follows that

$$\begin{aligned}&\int_0^\infty \mathrm{e}^{-zt}\left(\int_t^\infty \varphi_\varepsilon(s)\,\mathrm{d}s\right)\mathrm{d}t\\&=\frac{1}{z}(\|\varphi_\varepsilon\|_1-\hat{\varphi}_\varepsilon(z))\\&=\frac{1}{z}\left(\frac{1}{1-\|H_\varepsilon\|_1}-\frac{1}{1-\hat{H}_\varepsilon(z)}\right)=\frac{1}{z}\frac{\|H_\varepsilon\|_1-\hat{H}_\varepsilon(z)}{(1-\|H_\varepsilon\|_1)(1-\hat{H}_\varepsilon(z))}\sim\frac{2\Gamma(1/2)\varepsilon}{(1-\|H_\varepsilon\|_1)^2}z^{-1/2}\end{aligned}$$

as $z\to 0+$. Then monotone density theory (see [6, Theorem 1.7.2]) can be applied to show that

$$t^{1/2}\int_t^\infty \varphi_\varepsilon(s)\,\mathrm{d}s\to\frac{2\varepsilon}{(1-\|H_\varepsilon\|_1)^2}.$$

Therefore, for every $\delta>0$, applying the uniform convergence theorem (see [6, Theorem 1.5.1]),

$$\lim_{t\to\infty}\sup_{u\in[\delta,1]}\left|t^{1/2}\int_{ut}^\infty \varphi_\varepsilon(s)\,\mathrm{d}s-\frac{2\varepsilon u^{-1/2}}{(1-\|H_\varepsilon\|_1)^2}\right|=0\quad\text{and}\quad\sup_{t>0}t^{1/2}\int_t^\infty\varphi_\varepsilon(s)\,\mathrm{d}s<\infty,$$

which gives

$$\begin{aligned}&\lim_{n\to\infty}\int_\delta^1\left(n^{1/2}\int_{nu}^\infty\varphi_\varepsilon(v)\,\mathrm{d}v\right)\mathrm{d}u=\int_\delta^1\frac{2\varepsilon u^{-1/2}\,\mathrm{d}u}{(1-\|H_\varepsilon\|_1)^2}=\frac{2\varepsilon(1-\sqrt{\delta})}{(1-\|H\|_1-2\varepsilon)^2},\\&\int_0^\delta u^{-1/2}\left((nu)^{1/2}\int_{nu}^\infty\varphi_\varepsilon(v)\,\mathrm{d}v\right)\mathrm{d}u\le\sqrt{\delta}\times\sup_{t>0}t^{1/2}\int_t^\infty\varphi_\varepsilon(s)\,\mathrm{d}s.\end{aligned}$$

Passing $\delta\to 0$ and making use of the fact $\varphi\le\varphi_\varepsilon$ proves the limit in (A.1). □

Acknowledgements

We wish to thank the associate editor and reviewers for helpful comments that have improved the exposition of the paper.

Funding information

G. Pang is partly supported by US National Science Foundation grant DMS-2216765.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Ata, B. and Tongarlak, M. H. (2013). On scheduling a multiclass queue with abandonments under general delay costs. *Queueing Systems* **74**, 65–104.

[2] Atar, R., Budhiraja, A. and Dupuis, P. (2001). On positive recurrence of constrained diffusion processes. *Ann. Prob.* **29**, 979–1000.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press

[3] BACRY, E., DELATTRE, S., HOFFMANN, M. AND MUZY, J. F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stoch. Process. Appl.* **123**, 2475–2499.
[4] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd edn. John Wiley, New York.
[5] BILLINGSLEY, P. (2012). *Probability and Measure*. John Wiley, Hoboken, NJ.
[6] BINGHAM, N. H., GOLDIE, C. M. AND TEUGELS, J. L. (1987). *Regular Variation*. Cambridge University Press.
[7] CHEN, H. AND YAO, D. D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.
[8] CHEN, X. (2021). Perfect sampling of Hawkes processes and queues with Hawkes arrivals. *Stoch. Systems* **11**, 264–283.
[9] DAI, J. G. AND HARRISON, J. M. (1992). Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Ann. Appl. Prob.* **2**, 65–86.
[10] DAI, J. G. AND HARRISON, J. M. (2020). *Processing Networks: Fluid Models and Stability*. Cambridge University Press.
[11] DAW, A. AND PENDER, J. (2018). Queues driven by Hawkes processes. *Stoch. Systems* **8**, 192–229.
[12] DER BOOR, M. V., BORST, S. C., VAN LEEUWAARDEN, J. S. AND MUKHERJEE, D. (2022). Scalable load balancing in networked systems: a survey of recent advances. *SIAM Rev.* **64**, 554–622.
[13] DORSMAN, J.-P. L., VLASIOU, M. AND ZWART, B. (2015). Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds. *Queueing Systems* **79**, 293–319.
[14] FENDICK, K. AND WHITT, W. (2021). Queues with path-dependent arrival processes. *J. Appl. Prob.* **58**, 484–504.
[15] FENDICK, K. AND WHITT, W. (2022). Heavy traffic limits for queues with non-stationary path-dependent arrival processes. *Queueing Systems* **101**, 113–135.
[16] GAO, X. AND ZHU, L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems* **90**, 161–206.
[17] HARRISON, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Adv. Appl. Prob.* **10**, 886–905.
[18] HARRISON, J. M. AND REIMAN, M. I. (1981). Reflected Brownian motion on an orthant. *Ann. Prob.* **9**, 302–308.
[19] HARRISON, J. M., LANDAU, H. J. AND SHEPP, L. A. (1985). The stationary distribution of reflected Brownian motion in a planar region. *Ann. Prob.* **13**, 744–757.
[20] HUANG, J. AND ZHANG, H. (2013). Diffusion approximations for open Jackson networks with reneging. *Queueing Systems* **74**, 445–476.
[21] KIM, J. AND WARD, A. R. (2013). Dynamic scheduling of a GI/GI/1+ GI queue with multiple customer classes. *Queueing Systems* **75**, 339–384.
[22] KOOPS, D. T., SAXENA, M., BOXMA, O. J. AND MANDJES, M. (2018). Infinite-server queues with Hawkes input. *J. Appl. Prob.* **55**, 920–943.
[23] LI, B. AND PANG, G. (2023). On the splitting and aggregating of Hawkes processes. *J. Appl. Prob* **60**, 676–692.
[24] MANDJES, M. (2022). Multivariate M/G/1 systems with coupled input and parallel service. *Queueing Systems* **100**, 1–3.
[25] REIMAN, M. I. (1984). Open queueing networks in heavy traffic. *Math. Operat. Res.* **9**, 441–458.
[26] SELVAMUTHU, D. AND TARDELLI, P. (2022). Infinite-server systems with Hawkes arrivals and Hawkes services. *Queueing Systems* **101**, 329–351.
[27] SHIRYAEV, A. N. (1996). *Probability*. Springer, New York.
[28] WARD, A. R. AND GLYNN, P. W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43**, 103–128.
[29] WARD, A. R. AND GLYNN, P. W. (2003). Properties of the reflected Ornstein–Uhlenbeck process. *Queueing Systems* **44**, 109–123.
[30] WARD, A. R. AND GLYNN, P. W. (2005). A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems* **50**, 371–400.
[31] WHITT, W. (2000). An overview of Brownian and non-Brownian FCLTs for the single-server queue. *Queueing Systems* **36**, 39–70.
[32] WHITT, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York.
[33] WILLIAMS, R. J. (1987). Reflected Brownian motion with skew symmetric data in a polyhedral domain. *Prob. Theory Related Fields* **75**, 459–485.
[34] WILLIAMS, R. J. (2016). Stochastic processing networks. *Annu. Rev. Statist. Appl.* **3**, 323–345.
[35] XING, X., ZHANG, W. AND WANG, Y. (2009). The stationary distributions of two classes of reflected Ornstein–Uhlenbeck processes. *J. Appl. Prob.* **46**, 709–720.

https://doi.org/10.1017/jpr.2023.50 Published online by Cambridge University Press