*J. Appl. Probab.* 1–17 (2025)
*doi:*10.1017/jpr.2025.10046

# STOCHASTIC VOLTERRA INTEGRAL EQUATIONS WITH RANKS AS SCALING LIMITS OF PARALLEL INFINITE-SERVER QUEUES UNDER WEIGHTED SHORTEST QUEUE POLICY

TOMOYUKI ICHIBA,* *University of California, Santa Barbara*
GUODONG PANG,** *Rice University*

## Abstract

We study a queueing system with a fixed number of parallel service stations of infinite servers, each having a dedicated arrival process, and one flexible arrival stream that is routed to one of the service stations according to a 'weighted' shortest queue policy. We consider the model with general arrival processes and general service time distributions. Assuming that the dedicated arrival rates are of order $n$ and the flexible arrival rate is of order $\sqrt{n}$, we show that the diffusion-scaled queueing processes converge to a stochastic Volterra integral equation with 'ranks' driven by a continuous Gaussian process. It reduces to the limiting diffusion with a discontinuous drift in the Markovian setting.

*Keywords:* Stochastic Volterra integral equation with ranks; 'weighted' shortest queue routing policy; parallel infinite-server queues

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

## 1. Introduction

We consider a system of parallel service stations, each of which has a dedicated arrival process and an infinite number of servers. There is also a flexible arrival stream that can be served by any of the service stations, according to a 'weighted' shortest queue routing policy. This model was previously studied in [5, 10, 18] when the arrival processes are Poisson and the service times are independent and identically distributed (i.i.d.) exponential. In this paper we study the model with general arrival processes and service times with general distributions. The model has many applications, such as CDMA cellular systems [25] and customer service systems. The model is also related to studies of 'load balancing' in the sense that the 'weighted' shortest queue policy for the flexible arrivals balances the load of each of the service stations. We refer readers to the recent development on load balancing in [7], in which the studies focused on parallel server stations with one or multiple servers while the number of stations also goes to infinity. In this literature, [11] studied a Markovian model with an infinite number of servers in each station while the number of stations also goes to infinity, and a self-learning

Received 11 August 2025; accepted 11 August 2025.

* Postal address: Department of Statistics & Applied Probability, University of California, Santa Barbara, CA 9310. Email: ichiba@pstat.ucsb.edu
** Postal address: Department of Computational Applied Mathematics and Operations Research, George R. Brown School of Engineering, Rice University, Houston, TX 77005. Email: gdpang@rice.edu

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

threshold-based load balancing policy is proposed for the model. Our study has a fixed number of stations as in [5, 10, 18]. We also refer to [29, 30] for studies on joining the shortest queue policy in infinite-server queues, and to [6] for load balancing with a fixed number of single-server stations.

We study the system behavior under heavy traffic, i.e. when the arrival rates are scaled to grow to infinity, while the service times are unscaled. In the Markovian setting (Poisson arrivals and exponential service times), when the dedicated arrivals are of order $n$ and the flexible arrival stream is of order $\sqrt{n}$, [10] conjectured the diffusion limit with a discontinuous drift, which was then formally proved in [5]. A similar model was considered in [18] with a modification in the service process when the queues are over a threshold, and a diffusion limit with both discontinuous drift and variance coefficients was proved. The discontinuity in the drift is a consequence of the 'weighted' shortest queue policy, and in fact, the diffusion limit can also be regarded as a diffusion with 'ranks' in the drift. The discontinuity in the drift prevents us from applying the standard techniques in establishing heavy traffic scaling limits for queueing processes (such as the continuous mapping theorem applied to the integral mapping for standard Markovian many-server queues [21]). The idea to circumvent the discontinuity in the drift and/or variance coefficient in [5, 18] is to show that the time spent by the process in the set of their discontinuity is almost surely Lebesgue measure zero. The methods in [5, 18] rely heavily on semimartingales (constructed from the Poisson arrival and exponential service processes). However, their approaches through the (super-)martingale characterization do not apply to the non-Markovian setting we are considering.

For $G/GI/\infty$ queues with a general arrival process and service times, a functional central limit theorem (FCLT) is established in [16]. In that paper, the FCLT for the diffusion-scaled queueing process gives a Gaussian process limit, which has two independent components, capturing the randomness in the arrival and service processes, respectively. We adapt that approach to our setting, and in fact we directly use the results on the convergence of these components. We will show that the limit in the FCLT for our model is a multidimensional stochastic Volterra integral equation with 'ranks' in the integral term and driven by Gaussian processes (essentially the same Gaussian components as those in the limit for standard $G/GI/\infty$ queues with an additional Gaussian component resulting from the initial quantities); see (2.7). However, similarly to [5, 18], we have to tackle the issue of the 'ranks' in the integral term of the limit as a consequence of the 'weighted' shortest queue policy. We refer to [3] for existence, uniqueness, and perturbation results for the multidimensional stochastic Volterra integral equation driven by Brownian motions with Lipschitz continuous coefficients. Here, (2.7) does not satisfy the sufficient conditions for existence and uniqueness discussed in [3], because of the discontinuity of the drift coefficients due to the 'ranks'.

We show that the limiting stochastic Volterra integral equation has a unique weak solution with continuous paths. In order to prove the existence of a weak solution, we employ the Girsanov change-of-measure theorem for Brownian motion and Brownian sheets (noting that the driving Gaussian processes are functionals of either Brownian motion or a Brownian sheet). We include the terms with 'ranks' in the construction of a semimartingale that is a Brownian motion with a random drift, which will become a Brownian motion under a new measure. We also prove an important property of the limit process regarding the 'ranks', as in [5, 18]: the cumulative time that the limit process lies at the boundary (that is, when any two 'weighted' queues are equal) has Lebesgue measure zero with probability 1. For that purpose, we again exploit the Girsanov theorem and representations under the new measure.

In order to prove the convergence of the diffusion-scaled queueing processes to the limit process, we exploit some existing convergence results for the driving Gaussian components

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

for standard G/GI/$\infty$ queues [16]. However, from the representation in (4.1), the integral component with 'ranks' under the 'weighted' shortest queue policy forbids us from applying the continuous mapping theorem directly. To tackle this issue, we exploit the property of the limit process at the boundary mentioned above.

Our limit process (2.7) is related to the works on diffusions with discontinuous drifts in various applications. For example, the weak uniqueness of the diffusions with piecewise constant coefficients are established in [2], weak uniqueness for diffusions with discontinuous coefficients in [17], and the strong uniqueness of the diffusions with rank-based coefficients are discussed in [13]. See, e.g., [1, 8, 14] for the related stationary distributions and applications to the performance of functionally generated portfolios in financial markets. More generally, it is relevant to the theory of (stochastic) differential equations with discontinuous right-hand sides; see, e.g., [9] for differential equations.

The remainder of the paper is organized as follows. In Section 2 we describe the model in detail and state the main results in Theorems 2.1 and 2.2. In Section 3, we prove the existence and uniqueness (Theorem 3.1) of a weak solution to the limiting stochastic Volterra integral equation. We prove the convergence of diffusion-scaled processes in Section 4. We show how the limiting stochastic Volterra integral equation is reduced to the limiting diffusion with discontinuous drift in the Markovian case in Appendix A.

## 2. Model and results

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space where all the random variables and processes are defined. We consider a queueing system with $K$ service stations, each of which has its own dedicated arrival process and an infinite number of parallel servers. In addition, there is a flexible arrival stream, which can be served by any of the stations. Let $A_k = \{A_k(t)\colon t \geq 0\}$ be the dedicated arrival process at station $k = 1, \ldots, K$, with arrival rate $\lambda_k$ and arrival times $\tau_{k,i}$, $i \in \mathbb{N}$, and $A_0 = \{A_0(t)\colon t \geq 0\}$ be the flexible arrival process, with arrival rate $\lambda_0$ and arrival times $\tau_{0,i}$, $i \in \mathbb{N}$. Assume that these arrival processes are mutually independent. Let $X_k = \{X_k(t)\colon t \geq 0\}$ be the process counting the number of jobs in station $k$ for $k = 1, \ldots, K$, and denote the counting process of jobs in the $K$ service stations by $X = (X_1, \ldots, X_K)$.

For jobs initially in service in station $k$, let $\eta^0_{k,j}$, $j = 1, \ldots, X_k(0)$, be their remaining service times, and for the new arrivals from the stream $A_k$, let $\eta_{k,i}$, $i \in \mathbb{N}$, be their service times. For the jobs from the arrival stream $A_0$, let $\eta_{0,i}$, $i \in \mathbb{N}$, be their service times. We assume that the remaining service times $\{\eta^0_{k,j}\}_{k,j}$ are i.i.d. continuous random variables with cumulative distribution function (c.d.f.) $F_0$, and the service times $\{\eta_{k,i}\}_{k,i}$ are all i.i.d. with c.d.f. $F$. Let us denote the upper tail probabilities by $F_0^{\mathrm{c}} = 1 - F_0$ and $F^{\mathrm{c}} = 1 - F$ for $F_0$ and $F$, respectively. Without loss of generality, assume that the mean of $F$ is one.

We associate a 'weight' $\alpha_k > 0$ with each station $k$ in order to evaluate the status of station $k$ by a score $\alpha_k X_k(\cdot)$, and our routing policy depends on the scores. If the score $\alpha_k X_k$ of station $k$ is lower than those of the other stations, then we route the newly arriving jobs to station $k$. More specifically, for the $i$th job from the arrival stream $A_0$, at the arrival time $\tau_{0,i}$, it is routed to station $k$ if the score of station $k$ is the lowest among the other scores, i.e.

$$\alpha_k X_k(\tau_{0,i}) < \min_{\ell \neq k} \alpha_\ell X_\ell(\tau_{0,i}). \tag{2.1}$$

If multiple stations have the lowest scores, then job $i$ is routed to the station with the smallest number. For example, if stations 1 and 2 have the same score, i.e. $\alpha_1 X_1(\tau_{0,i}) = \alpha_2 X_2(\tau_{0,i})$ at the

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

time $\tau_{0,i}$ of the arrival of job $i$, then job $i$ is routed to station 1. Ties in the scores are resolved in this lexicographic way. For each $x=(x_1, \ldots, x_K)\in\mathbb{R}_+^K$, define an indicator function

$$\delta_k(x)=\begin{cases}1 & \text{if } k=\min\{j\colon \alpha_j x_j=\min_{1\le \ell\le K}\alpha_\ell x_\ell\},\\ 0 & \text{otherwise,}\end{cases} \tag{2.2}$$

and the set $\mathcal{R}_k := \{x\in\mathbb{R}_+^K\colon \delta_k(x)=1\}$ for $k=1, \ldots, K$. Then we have $\delta_k(x)=\mathbf{1}_{\mathcal{R}_k}(x)$ for $x\in\mathbb{R}_+^K$, $k=1, \ldots, K$. Moreover, $\mathcal{R}_j\cap\mathcal{R}_k=\emptyset$ for $j\neq k$, and $\cup_{k=1}^K\mathcal{R}_k=\mathbb{R}_+^K$. Since $\delta_k(cx)=\delta_k(x)$ for every $x\in\mathbb{R}_+^K$, $c>0$, we see that each set $\mathcal{R}_k$ forms a cone for $k=1, \ldots, K$.

Then, with these indicator functions, the process $X_k(t)$ can be described as

$$X_k(t)=\sum_{j=1}^{X_k(0)}\mathbf{1}_{\{\eta^0_{k,j}>t\}}+\sum_{i=1}^{A_k(t)}\mathbf{1}_{\{\tau_{k,i}+\eta_{k,i}>t\}}+\sum_{i=1}^{A_0(t)}\delta_k(X(\tau_{0,i}-))\mathbf{1}_{\{\tau_{0,i}+\eta_{0,i}>t\}},\qquad t\ge 0.$$

We consider a sequence of such queueing systems, indexed by $n$, and the scaling limit. We first make the following assumption on the arrival rates.

**Assumption 2.1.** *Assume that for $k=1, \ldots, K$, $\lambda_k^n/n\to\lambda_k$ as $n\to\infty$, and $\lambda_0^n/\sqrt{n}\to\lambda_0$ as $n\to\infty$.*

Let $\bar{X}_k^n := n^{-1}X_k^n$ for $k=1, \ldots, K$. The following functional law of large numbers (FLLN) holds. We use $D(\mathbb{R}_+, \mathbb{R}^d)$ to denote $\mathbb{R}^d$-valued càdlàg functions, endowed with the Skorokhod $J_1$ topology written as $(D(\mathbb{R}_+, \mathbb{R}^d), J_1)$ (see [4] or [26] for its definition). When $d=1$, we write $D=D(\mathbb{R}_+, \mathbb{R})$ and use $(D^d, J_1)$ to denote the $d$-fold product topology. When the space is restricted on the fixed time interval $[0, T]$ for some $T>0$, we write $D_{[0,T]} := D([0, T], \mathbb{R})$ and use $\big(D^d_{[0,T]}, J_1\big)$ to denote the $d$-fold product topology.

**Theorem 2.1.** *Under Assumption 2.1, if there are deterministic constants $\bar{X}_k(0)$, $k=1, \ldots, K$ such that $\big(\bar{X}_1^n(0), \ldots, \bar{X}_K^n(0)\big)\Rightarrow(\bar{X}_1(0), \ldots, \bar{X}_K(0))$ in $\mathbb{R}_+^K$ as $n\to\infty$, then*

$$\big(\bar{X}_1^n, \ldots, \bar{X}_K^n\big)\Rightarrow(\bar{X}_1, \ldots, \bar{X}_K)\quad \text{in } (D^K, J_1) \text{ as } n\to\infty,$$

*where*

$$\bar{X}_k(t)=\bar{X}_k(0)F_0^{\mathrm{c}}(t)+\lambda_k\int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}s,\qquad t\ge 0. \tag{2.3}$$

*In addition, $\bar{X}(t)\to\bar{X}^*=\big(\bar{X}_1^*, \ldots, \bar{X}_K^*\big)$ as $t\to\infty$, with $\bar{X}_k^*=\lambda_k$ for each $k=1, \ldots, K$.*

*Proof.* Under the scaling in Assumption 2.1, the arrival rate $\lambda_0^n/\sqrt{n}\to\lambda_0$ implies that $\bar{A}_0^n(t)=A_0^n(t)/n\to 0$ in $D$ in probability as $n\to\infty$. Hence, the third term in $\bar{X}_k^n(t)$,

$$\frac{1}{n}\sum_{i=1}^{n\bar{A}_0^n(t)}\delta_k(\bar{X}^n(\tau_{0,i}-))\mathbf{1}_{\{\tau^n_{0,i}+\eta_{0,i}>t\}}$$

tends to 0 in $D$ in probability as $n\to\infty$, since the summands are all bounded by one. Then, the convergence of $\bar{X}_k^n(t)$ reduces to the convergence of the first two terms, which is exactly the convergence of the LLN-scaled number of jobs in an infinite-server queueing model with an arrival process $A_k^n(t)$ and i.i.d. service times $\{\eta_{k,i}, i\ge 1\}$. This follows from the existing result in the literature; see, e.g., [16, 21]. □

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

Observe that in the fluid limit, the ‘weighted’ shortest queue policy in (2.1) is irrelevant as indicated in (2.3), and the associated steady-state values $\bar{X}_k^* = \lambda_k$. This is evident since the extra load $A_0^n(t)$ is of order $\sqrt{n}$ while the fluid scale is of order $n$. It is then necessary to consider the queueing dynamics in the diffusion scale.

**Remark 2.1.** We have focused particularly on the case $\lambda_0^n/\sqrt{n} \to \lambda_0$ because the extra load does not affect the behavior of each individual queue in the fluid scale, but only affects it in the diffusion scale, resulting in the interesting limit with ranks as seen in Theorem 2.2. If we assume $\lambda_0^n/n \to \lambda_0$ as $n \to \infty$, then $\bar{A}_0^n(t) = A_0^n(t)/n \to \lambda_0 t$ in $D$ in probability as $n \to \infty$, and hence the third term in $\bar{X}_k^n(t)$ will not vanish, so different behaviors in both fluid and diffusion scales are expected. Some heuristic results and conjectures are provided in [10] in the Markovian setting. The study of both the FLLN and FCLT for the non-Markovian model under this assumption is left for future work, since new challenges arise to prove the convergences.

Let $\hat{X}_k^n := \sqrt{n}\big(\bar{X}_k^n - \bar{X}_k^*\big)$ for $k = 1, \ldots, K$, and write $\hat{X}^n = \big(\hat{X}_1^n, \ldots, \hat{X}_K^n\big)$. Note that we center the process $\hat{X}_k^n$ by its equilibrium point. Then it is clear that if $\alpha_k = 1/\lambda_k$, $\delta_k(X^n(t)) = \delta_k(\hat{X}^n(t))$, $t \geq 0$. We choose this specific value of $\alpha_k = 1/\lambda_k$ to establish the following FCLT. Note that since all the service times are i.i.d. with mean one, the value $\lambda_k$ is the (offered) load at each station. Thus, the routing criterion chooses the station with minimum ratio of the current state and the steady state. When the $\lambda_k$ are equal for all $k$, the routing policy becomes the so-called ‘joining the shortest queue’ (JSQ) policy. Thus, the routing policy can be regarded as a ‘weighted’ JSQ policy with the weights being the reciprocal of the offered load.

Let $\hat{A}_k^n(t) := (A_k^n(t) - \lambda_k^n t)/\sqrt{n}$ for $t \geq 0$ and $k = 1, \ldots, K$, and let $\hat{A}_0^n(t) = A_0^n(t)/\sqrt{n}$ for $t \geq 0$. We make the following assumption for these scaled arrival processes.

**Assumption 2.2.** *The following hold for the arrival processes:*

(i) *In addition to the conditions in Assumption 2.1, for each $k = 1, \ldots, K$ there exists $\hat{\lambda}_k \in \mathbb{R}$ such that $\hat{\lambda}_k^n := \sqrt{n}\big(\lambda_k^n/n - \lambda_k\big) \to \hat{\lambda}_k$ as $n \to \infty$.*

(ii) *There exist $K$ mutually independent Brownian motions $(\hat{A}_1, \ldots, \hat{A}_K)$ such that*

$$\big(\hat{A}_1^n, \ldots, \hat{A}_K^n\big) \Rightarrow (\hat{A}_1, \ldots, \hat{A}_K) \quad \text{in } (D^K, J_1) \text{ as } n \to \infty,$$

*where $\hat{A}_k \overset{\mathrm{d}}{=} c_k \hat{B}_k(t)$ for the variance coefficient $c_k > 0$ and a standard Brownian motion $\hat{B}_k$ (mutually independent over $k$).*

(iii) *$\hat{A}_0^n \Rightarrow \lambda_0 e$ in $(D, J_1)$ as $n \to \infty$, where $e(t) \equiv t$ for $t \geq 0$. Moreover, associating measures on [0, T] to the non-negative, non-decreasing, càdlàg functions $\hat{A}_0^n$ and $\lambda_0 e$, we assume the total variation distance between $\mathrm{d}\hat{A}_0^n$ and $\lambda_0\, \mathrm{d}e$ converges weakly to 0 as $n \to \infty$ for every $T > 0$.*

In the second condition, when the arrival processes $A_k$ are mutually independent renewal processes with the interarrival times having mean $\lambda_k^{-1}$ and variance $\sigma_k^2$, if $A_k^n$ is defined by scaling the interarrival times by $n^{-1}$, then the limit is given by $\hat{A}_k(t) = \sqrt{\lambda_k^3 \sigma_k^2}\, \hat{B}_k(t)$ for some standard Brownian motion $\hat{B}_k(t)$ (see, e.g., [26, Chapter 13.7]), $k = 1, \ldots, K$, $t \geq 0$.

We also make the following assumption on the initial condition.

**Assumption 2.3.** *There exists a random vector $(\hat{X}_1(0), \ldots, \hat{X}_K(0)) \in \mathbb{R}^K$ such that*

$$\big(\hat{X}_1^n(0), \ldots, \hat{X}_K^n(0)\big) \Rightarrow (\hat{X}_1(0), \ldots, \hat{X}_K(0)) \quad \text{in } \mathbb{R}_+^K \text{ as } n \to \infty.$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

Under this condition, given the scaling of $\hat{X}_k$, it is also clear that $\bar{X}_k(0) = \bar{X}_k^* = \lambda_k$ for each $k = 1, \dots, K$.

For the FCLT in Theorem 2.2, we also assume that $F_0(t) = F_{\mathrm{e}}(t)$, where $F_{\mathrm{e}}(t) = \int_0^t F^{\mathrm{c}}(s)\,\mathrm{d}s$ for $t \geq 0$ is the equilibrium (stationary excess) distribution of $F$. Observe that for the fluid limit $\bar{X}(t)$ in (2.3), if $F_0 = F_{\mathrm{e}}$ and $\bar{X}(0) = \bar{X}^*$, then $\bar{X}(t) = \bar{X}^*$ for all $t \geq 0$. Recall that we have assumed that the mean for the c.d.f. $F$ is one. In the scaling limits, we have the following three mutually independent driving noises:

(i) some $K$-dimensional, independent Gaussian processes $\hat{X}_{\cdot,0} := (\hat{X}_{1,0}, \dots, \hat{X}_{K,0})'$;

(ii) another $K$-dimensional, independent Gaussian process, $\hat{X}_{\cdot,1} := (\hat{X}_{1,1}, \dots, \hat{X}_{K,1})'$; and

(iii) the $K$-dimensional, independent Brownian motions $(\hat{A}_1, \dots, \hat{A}_K)$ with strictly positive variance rates $(c_1, \dots, c_K)$ from Assumption 2.2.

The processes $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ are independent continuous Gaussian processes, independent of $\hat{A}_k$, with mean zero and covariance functions, for $t, t' \geq 0$,

$$\mathrm{Cov}(\hat{X}_{k,0}(t), \hat{X}_{k,0}(t')) = \lambda_0\big(F_{\mathrm{e}}^{\mathrm{c}}(t \vee t') - F_{\mathrm{e}}^{\mathrm{c}}(t)F_{\mathrm{e}}^{\mathrm{c}}(t')\big), \tag{2.4}$$

$$\mathrm{Cov}(\hat{X}_{k,1}(t), \hat{X}_{k,1}(t')) = \lambda_k \int_0^{t \wedge t'} \big(F^{\mathrm{c}}(t \vee t' - s) - F^{\mathrm{c}}(t-s)F^{\mathrm{c}}(t'-s)\big)\,\mathrm{d}s. \tag{2.5}$$

In addition, the processes $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ are independent of $\hat{X}_{k',0}$ and $\hat{X}_{k',1}$, as well as $\hat{A}_{k'}$, for every $k' \neq k$. The process $\int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_k(s)$ is also a continuous Gaussian process with covariance function, for $k = 1, \dots, K$,

$$\mathrm{Cov}\bigg(\int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_k(s), \int_0^{t'} F^{\mathrm{c}}(t'-s)\,\mathrm{d}\hat{A}_k(s)\bigg) = c_k \int_0^{t \wedge t'} F^{\mathrm{c}}(t-s)F^{\mathrm{c}}(t'-s)\,\mathrm{d}s.$$

As we shall see later in (3.2) and (3.5), respectively, $\hat{X}_{k,0}$ and $\hat{X}_{k,1}$ can be represented as a time-changed, Brownian bridge driven by another Brownian motion, and as a time-changed Kiefer process driven by a Brownian sheet, independent of Brownian motions.

**Definition 2.1.** (*Weak solution to a system of stochastic Volterra integral equations.*) We shall consider a weak solution to equation (2.6), that is, on some filtered probability space $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0})$, a continuous, adapted, $K$-dimensional process $\hat{\mathcal{X}} = (\hat{\mathcal{X}}_1, \dots, \hat{\mathcal{X}}_K)'$, an independent, $K$-dimensional Gaussian process $\hat{\mathcal{X}}_{\cdot,0} := (\hat{\mathcal{X}}_{1,0}, \dots, \hat{\mathcal{X}}_{K,0})'$, another independent, $K$-dimensional Gaussian process $\hat{\mathcal{X}}_{\cdot,1} := (\hat{\mathcal{X}}_{1,1}, \dots, \hat{\mathcal{X}}_{K,1})'$ determined by the covariance functions (2.4), and an independent, $K$-dimensional Brownian motion $\hat{\mathcal{A}}_{\cdot}$ with some variance rates $(c_1, \dots, c_K)$ satisfy almost surely, for $t \geq 0$,

$$\begin{aligned}\hat{\mathcal{X}}_k(t) = \hat{\mathcal{X}}_k(0)F_{\mathrm{e}}^{\mathrm{c}}(t) + \hat{\lambda}_k F_{\mathrm{e}}(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))F^{\mathrm{c}}(t-s)\,\mathrm{d}s \\ + \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{\mathcal{A}}_k(s) + \hat{\mathcal{X}}_{k,0}(t) + \hat{\mathcal{X}}_{k,1}(t).\end{aligned} \tag{2.6}$$

**Theorem 2.2** (FCLT.) *Under Assumptions 2.1, 2.2, and 2.3, we have*

$$\big(\hat{X}_1^n, \dots, \hat{X}_K^n\big) \Rightarrow (\hat{X}_1, \dots, \hat{X}_K) \quad \text{in } (D^K, J_1) \text{ as } n \to \infty,$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

*where the limit* $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_K)$*, together with the independent, continuous Gaussian processes* $(\hat{X}_{1,0}, \ldots, \hat{X}_{K,0})$*,* $(\hat{X}_{1,1}, \ldots, \hat{X}_{K,1})$*, and some Brownian motions* $(\hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_K)$ *(having the same law as that of* $(\hat{A}_1, \ldots, \hat{A}_K)$ *in Assumption* 2.2(ii)*, is a weak solution, unique in distribution, to the system of stochastic Volterra integral equations in* (2.6)*.*

For the convenience of notation, we write the limit $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_K)$ as

$$\begin{aligned}\hat{X}_k(t) &= \hat{X}_k(0)F_{\mathrm{e}}^{\mathrm{c}}(t) + \hat{\lambda}_k F_{\mathrm{e}}(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))F^{\mathrm{c}}(t-s)\,\mathrm{d}s \\ &\quad + \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{\mathcal{A}}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t). \end{aligned} \tag{2.7}$$

**Remark 2.2.** If we denote the last three Gaussian processes in (2.7) by

$$\hat{Y}_k(t) = \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{\mathcal{A}}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t)$$

and assume $F$ has a density $f$ and that $F(0) = 0$, then we can write

$$\begin{aligned}\hat{X}_k(t) &= \hat{X}_k(0)F_{\mathrm{e}}^{\mathrm{c}}(t) + \hat{\lambda}_k F_{\mathrm{e}}(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))F^{\mathrm{c}}(t-s)\,\mathrm{d}s + \hat{Y}_k(t) \\ &= \int_0^t \left( -\hat{X}_k(0)F^{\mathrm{c}}(s) + \hat{\lambda}_k F^{\mathrm{c}}(s) + \lambda_0 \delta_k(\hat{X}(s)) - \lambda_0 \int_0^s \delta_k(\hat{X}(s)) f(s-u)\,\mathrm{d}u \right) \mathrm{d}s \\ &\quad + \hat{Y}_k(t) \end{aligned} \tag{2.8}$$

for every $t \geq 0$. Observe that in the ‘drift’, there is not only a discontinuous term $\lambda_0 \delta_k(\hat{X}(t))$, but also a memory of the process $\hat{X}(t)$ in the term $\lambda_0 \int_0^t \delta_k(\hat{X}(s)) f(t-s)\,\mathrm{d}s$.

**Remark 2.3.** (*Diffusion with discontinuous drift.*) When the c.d.f. $F$ is given by $F(t) = 1 - \mathrm{e}^{-t}$, $t \geq 0$, it can be shown that the stochastic equation in (2.8) reduces to the diffusion with discontinuous drift studied in [5, 18] (with some modification on the coefficients and also in the drift due to the absence of queue thresholds; see also the conjecture in [10]), that is,

$$\mathrm{d}\hat{X}_k(t) = (\hat{\lambda}_k + \lambda_0 \delta_k(\hat{X}(t)) - \hat{X}_k(t))\,\mathrm{d}t + \sqrt{\lambda_k + c_k^2}\,\mathrm{d}\tilde{B}_k(t) \tag{2.9}$$

for standard Brownian motion $(\tilde{B}_1, \ldots, \tilde{B}_K)$. The proof of this property is given in Appendix A.

**Remark 2.4.** It would be interesting to characterize the stationary distribution of the limit process. Without the flexible stream, the limit for each queue is a Gaussian process (Ornstein–Uhlenbeck process in the Markovian setting), and the stationary distribution is Gaussian, whose variance formula can be given explicitly (see, e.g., [16, 22]). However, with the flexible stream, the components of the multidimensional limit process are coupled via the term with ranks, which makes it impossible to characterize the stationary distribution of the limit explicitly. In the Markov setting, an Euler–Maruyama method (e.g. [19]) may potentially be used to approximate the stationary distribution of the diffusion with ranks in the drift in (2.9). However, for the non-Markov model, it is an open problem to develop methods to approximate the stationary distribution of the Gaussian-driven multidimensional Volterra-type integral equation with ranks as given in (2.8).

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

**Remark 2.5.** (*Atlas models.*) We remark that in the completely symmetric case, i.e. $\lambda_k$ are all equal, we have that the $\hat{\lambda}_k$ are also all equal, and assuming that $\hat{X}_k(0)$ have the same distribution for all $k$, then the limit process $\hat{X}_k(t)$ in (2.7) becomes symmetric over $k$ with the ranks $\delta_k(\hat{X}(t))$ only ranking the coordinates without weights at each time $t$. When $\alpha_1 = \cdots = \alpha_K$, this resembles the first-order Atlas model in stochastic portfolio theory [8], which is a diffusion with both the drift and variance coefficients depending only on ranking, driven by a Brownian motion.

## 3. Existence and uniqueness of a weak solution to the limiting stochastic Volterra integral equation with ranks

In this section we prove that the stochastic Volterra integral equation in (2.7) has a unique weak solution that has continuous paths.

**Theorem 3.1.** *The stochastic Volterra integral equation* (2.7) *has a unique weak solution in* $\mathbb{C}(\mathbb{R}_+, \mathbb{R}^K)$*. Moreover, almost surely,*

$$\int_0^\infty \sum_{k,\ell=1,\, k\neq\ell}^{K} \mathbf{1}_{\{\alpha_k \hat{X}_k(t)=\alpha_\ell \hat{X}_\ell(t)\}}\, \mathrm{d}t = 0, \tag{3.1}$$

*where* $\{\alpha_k\}_{1\le k\le K}$ *are the associated weights in* (2.1)*.*

Before proceeding to the proof, we make the following observations on the Gaussian processes $\hat{X}_{\cdot,0} := (\hat{X}_{1,0}, \ldots, \hat{X}_{K,0})'$ and $\hat{X}_{\cdot,1} := (\hat{X}_{1,1}, \ldots, \hat{X}_{K,1})'$. The process $\hat{X}_{k,0}$ is equivalent in distribution to a time-changed Brownian bridge $\hat{W}^0_k$:

$$\hat{X}_{k,0}(t) = \lambda_k^{1/2} W^0(F_{\mathrm{e}}(t)) = \lambda_k^{1/2} \hat{W}^0_k(1-\mathrm{e}^{-t}). \tag{3.2}$$

Recall that the Brownian bridge $W^0_k$ is the unique strong solution to the one-dimensional stochastic differential equation (SDE) [15]

$$\mathrm{d}\hat{W}^0_k(t) = -\frac{\hat{W}^0_k(t)}{1-t}\,\mathrm{d}t + \mathrm{d}\breve{B}_k(t),$$

where $\breve{B}_k(t)$ is an independent standard Brownian motion. Thus we can write

$$\hat{X}_{k,0}(t) = \lambda_k^{1/2}\left(-\int_0^{1-\mathrm{e}^{-t}} \frac{W^0_k(s)}{1-s}\,\mathrm{d}s + \breve{B}_k(1-\mathrm{e}^{-t})\right) = -\int_0^t \hat{X}_{k,0}(s)\,\mathrm{d}s + \lambda_k^{1/2}\breve{B}_k(1-\mathrm{e}^{-t}) \tag{3.3}$$

for standard Brownian motions $\breve{B}_k(t)$, independent of the $\hat{B}_k(t)$ in Assumption 2.2(ii). The second equality follows from a calculation using the change of variables for the integrable term.

Next, the Gaussian process $\hat{X}_{k,1}$ is equivalent in distribution to a time-changed Kiefer process, $\hat{\mathcal{K}}_k(t, x)$, that is,

$$\hat{X}_{k,1}(t) = -\int_0^t \int_0^t \mathbf{1}_{s+x\le t}\,\mathrm{d}\hat{\mathcal{K}}_k(\lambda_k s, F(x)) = -\int_0^t \int_0^t \mathbf{1}_{s+x\le t}\,\mathrm{d}\hat{\mathcal{K}}_k(\lambda_k s, 1-\mathrm{e}^{-x}). \tag{3.4}$$

Recall that, similar to a Brownian bridge, the Kiefer process $\hat{\mathcal{K}}_k(t, x)$ can be written in terms of Brownian sheets $\hat{W}_k(t, x)$, i.e. $\hat{\mathcal{K}}_k(t, x)$ is the unique strong solution to the SDE [24]

$$\mathcal{K}(t, x) = -\int_0^x \frac{\mathcal{K}(t, y)}{1-y}\,\mathrm{d}y + \hat{W}_k(t, x).$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

So we have

$$\begin{aligned}\hat{X}_{k,1}(t) &= -\int_0^t \int_0^t \mathbf{1}_{s+x\le t}\, \mathrm{d}_{s,x}\bigg(-\int_0^{1-\mathrm{e}^{-x}} \frac{\mathcal{K}(\lambda_k s,\, y)}{1-y}\, \mathrm{d}y + \hat{W}_k(\lambda_k s,\, 1-\mathrm{e}^{-x})\bigg)\\ &= -\int_0^t \hat{X}_{k,1}(s)\, \mathrm{d}s + \int_0^t \int_0^t \mathbf{1}_{s+x\le t}\, \mathrm{d}\hat{W}_k(\lambda_k s,\, 1-\mathrm{e}^{-x}), \end{aligned} \tag{3.5}$$

where the $\hat{W}_k$ are mutually independent Brownian sheets, which are also independent of the standard Brownian motions $\hat{B}_k$ in (3.3) and $B_k$ in Assumption 2.2(ii).

*Proof of Theorem 3.1.*

*Step 1: Construction.* In order to construct the solution, let us analyze the solution to the stochastic Volterra integral equation (2.7) and derive equation (3.8) first. Consider a probability space $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge 0})$ under which the independent Brownian motions $(\hat{A}_1, \ldots, \hat{A}_K)$ in Assumption 2.2, independent Brownian bridges $(\hat{W}^0_1, \ldots, \hat{W}^0_K)$ in (3.2), and independent Kiefer processes $(\hat{\mathcal{K}}_1, \ldots, \hat{\mathcal{K}}_K)$ in (3.4) are defined. Assume for a moment that the solution $\hat{X}(\cdot)$ in (2.7) exists and, for $k = 1, \ldots, K$, let us define

$$\hat{R}_k(t) := \hat{\lambda}_k t + c_k \hat{B}_k(t) + \lambda_0 \int_0^t \delta_k(\hat{X}(s))\, \mathrm{d}s, \tag{3.6}$$

$$\hat{S}_k(t) := \hat{X}_k(0) F^{\mathrm{c}}_{\mathrm{e}}(t) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t). \tag{3.7}$$

Write $\hat{R} = (\hat{R}_1, \ldots, \hat{R}_K)'$ and $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_K)'$. Then $\hat{R}$ is a semimartingale with respect to the filtration $\{\mathcal{F}_t\}_{t\ge 0}$, which is a Brownian motion with a random drift. The process $\hat{S}_k$ is a continuous Gaussian process that starts at $\hat{X}_k(0)$ for $k = 1, \ldots, K$. Recall the representations of $\hat{X}_{k,0}(t)$ and $\hat{X}_{k,1}(t)$ in (3.2) and (3.4), respectively, using Brownian bridge $\hat{W}^0_k(t)$ and Kiefer process $\hat{\mathcal{K}}_k(t, x)$ (through the Brownian sheet $\hat{W}_k$). Also, let $\hat{X}_{\cdot,0} = (\hat{X}_{1,0}, \ldots, \hat{X}_{K,0})'$ and $\hat{X}_{\cdot,1} = (\hat{X}_{1,1}, \ldots, \hat{X}_{K,1})'$. We similarly define, for $\hat{W}^0$, $\hat{\mathcal{K}}$ and $\hat{W}$. Note that since $\hat{B}(t)$, $\hat{X}_{\cdot,0}(t)$, and $\hat{X}_{\cdot,1}(t)$ are mutually independent, we also have the mutual independence between $\hat{B}(t)$, $\hat{W}^0$, and $\hat{W}$. Then, with (3.6) and (3.7), the stochastic integral equation (2.7) can be rewritten as

$$\hat{X}_k(t) = \int_0^t F^{\mathrm{c}}(t-s)\, \mathrm{d}\hat{R}_k(s) + \hat{S}_k(t) \tag{3.8}$$

for $k = 1, \ldots, K$, $t \ge 0$, where the semimartingale $\hat{R}$ depends on $\hat{X}$ as in (3.6) and the continuous Gaussian process $\hat{S}$ is independent of $\hat{B}$.

We shall first construct a weak solution. To construct a solution to this stochastic equation, we use the change of measure in such a way that the semimartingale $\hat{R}_k(s)$ becomes a Brownian motion under a new measure.

We consider a $K$-dimensional standard Brownian motion $\widetilde{\beta} := (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_K)$ and the independent Gaussian process $\hat{S}$ in (3.7) constructed from the independent Brownian bridges $\hat{W}^0_{\cdot}$ and the independent Kiefer processes $\hat{\mathcal{K}}_{\cdot}$ through the Brownian sheets, independent of $\widetilde{\beta}$, on a filtered probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \{\widetilde{\mathcal{F}}(t)\}_{t\ge 0}, \widetilde{\mathbb{P}})$, and we define $\xi := (\xi_1, \ldots, \xi_K)$, $M := (M_1, \ldots, M_K)$, with

$$\xi_k(t) := \hat{S}_k(t) + \int_0^t F^{\mathrm{c}}(t-s) c_k\, \mathrm{d}\widetilde{\beta}_k(s), \qquad M_k(t) := \widetilde{\beta}_k(t) - \int_0^t \bigg(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\bigg)\, \mathrm{d}s, \tag{3.9}$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

with the indicator function $\delta_k$ in (2.2) for $k = 1, \ldots, K$ and $t \geq 0$,

$$Z(t) = \exp\left(\sum_{k=1}^{K} \int_0^t \left(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\right) \mathrm{d}\widetilde{\beta}_k(s) - \frac{1}{2}\sum_{k=1}^{K} \int_0^t \left(\frac{\hat{\lambda}_k}{c_k} + \frac{\lambda_0}{c_k}\delta_k(\xi(s))\right)^2 \mathrm{d}s\right).$$

Here, $M$ is a $K$-dimensional, drifted Brownian motion with at most linearly growing drifts.

Then the stochastic exponential $Z$ is a continuous martingale under the probability measure $\widetilde{\mathbb{P}}$, and hence, for a fixed $T > 0$, we define a new probability measure $\widetilde{\mathbb{Q}}$ by

$$\left.\frac{\mathrm{d}\,\widetilde{\mathbb{Q}}}{\mathrm{d}\,\widetilde{\mathbb{P}}}\right|_{\widetilde{\mathcal{F}}(T)} := Z(T).$$

Applying the Girsanov theorem ([15, Theorem 3.5.1] for Brownian motions; see also, e.g., [20, Proposition 1.6] for Brownian sheets), we see that $M$ is a $K$-dimensional, standard Brownian motion, independent of $\hat{S}$, under the probability measure $\widetilde{\mathbb{Q}}$. Here, by the Girsanov theorem, we remove the drifts of the drifted Brownian motion but we do not shift the Brownian sheets that drive the independent Gaussian processes $\hat{S}$. Thus, under $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \{\widetilde{\mathcal{F}}(t)\}_{t\geq 0}, \widetilde{\mathbb{Q}})$, the adapted continuous process $\xi$, the continuous Gaussian process $\hat{S}$, and the Brownian motion $M$ satisfy the equation

$$\xi_k(t) = \hat{S}_k(t) + \int_0^t F^{\mathrm{c}}(t-s)c_k\,\mathrm{d}M_k(s) + \int_0^t \hat{\lambda}_k F^{\mathrm{c}}(t-s)\,\mathrm{d}s + \lambda_0 \int_0^t F^{\mathrm{c}}(t-s)\delta_k(\xi(s))\,\mathrm{d}s$$

for $k = 1, \ldots, K$, $0 \leq t \leq T$. Thus, $\xi$, $M$, and $\hat{S}$ in (3.9) satisfy the system (2.7) of stochastic Volterra integral equations for $0 \leq t \leq T$ under the new measure $\widetilde{\mathbb{Q}}$. Since $T > 0$ is arbitrary, by the above construction there is a weak solution for $t \geq 0$ to the system (2.7) of the stochastic Volterra integral equations.

*Step 2: Uniqueness.* The joint distribution of the weak solution $(\hat{X}, \hat{A}, \hat{X}_{\cdot,0}, \hat{X}_{\cdot,1})$ to (2.7) is uniquely determined by the Girsanov change of measure as in the proof of [15, Proposition 5.3.10], i.e. we firstly localize the problem by defining the sequence of the first passage times for $\hat{X}$ to the sphere of integer radii, centered at the origin; secondly we apply the Girsanov change of measure with those stopping times; and then we take the limits. Note that the Gaussian processes $(\hat{X}_{\cdot,0}, \hat{X}_{\cdot,1})$ are independent of the Brownian motion $\hat{A}$. When the initial values $\hat{X}(0)$ are randomized, we may determine the distribution as in [15, Corollary 5.3.11].

*Step 3: Proof of* (3.1). To show (3.1), we first show that for the processes $\xi$ in (3.9) under $\widetilde{\mathbb{P}}$,

$$\int_0^T \sum_{k,\ell=1,\,k\neq\ell}^{K} \mathbf{1}_{\{\alpha_k\xi_k(t)=\alpha_\ell\xi_\ell(t)\}}\,\mathrm{d}t = 0, \tag{3.10}$$

and then we once again apply the Girsanov theorem to show that (3.10) holds under $\widetilde{\mathbb{Q}}$ as in Step 1, and hence the weak solution $\hat{X}$ satisfies (3.1). Thus, it suffices to show (3.10) under $\widetilde{\mathbb{P}}$ where $\hat{S}$ and $\widetilde{\beta}$ are independent. Note that since the tail probability $F^{\mathrm{c}}$ and positive constants $c_k$ are deterministic, each integral $\int_0^t F^{\mathrm{c}}(t-s)c_k\,\mathrm{d}\widetilde{\beta}_k(s)$ is normally distributed with mean 0 and variance $\int_0^t (F^{\mathrm{c}}(t-s)c_k)^2\,\mathrm{d}s$ for each $k = 1, \ldots, K$, independent of $\hat{S}$ under $\widetilde{\mathbb{P}}$.

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

It follows that, for every $k \neq \ell$ and $t \geq 0$, and for every fixed $(\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$,

$$\widetilde{\mathbb{P}}\left(\alpha_k\left(\theta_k + \int_0^t F^{\mathrm{c}}(t-s)c_k \, \mathrm{d}\widetilde{\beta}_k(s)\right) = \alpha_\ell\left(\theta_\ell + \int_0^t F^{\mathrm{c}}(t-s)c_\ell \, \mathrm{d}\widetilde{\beta}_\ell(s)\right)\right) = 0,$$

and hence, by the tower property of the conditional probability and by the independence, we have, for any $k \neq \ell$ and $t \in [0, T]$,

$$\begin{aligned}
&\widetilde{\mathbb{P}}(\alpha_k\xi_k(t) = \alpha_\ell\xi_\ell(t)) \\
&= \widetilde{\mathbb{E}}\left[\widetilde{\mathbb{P}}\left(\alpha_k\left(\hat{S}_k(t) + \int_0^t F^{\mathrm{c}}(t-s)c_k \, \mathrm{d}\widetilde{\beta}_k(s)\right) = \alpha_\ell\left(\hat{S}_\ell(t) + \int_0^t F^{\mathrm{c}}(t-s)c_\ell \, \mathrm{d}\widetilde{\beta}_\ell(s)\right) \mid \hat{S}(t)\right)\right] \\
&= \widetilde{\mathbb{E}}\left[\widetilde{\mathbb{P}}\left(\alpha_k\left(\theta_k + \int_0^t F^{\mathrm{c}}(t-s)c_k \, \mathrm{d}\widetilde{\beta}_k(s)\right)\right.\right. \\
&\qquad \left.\left. = \alpha_\ell\left(\theta_\ell + \int_0^t F^{\mathrm{c}}(t-s)c_\ell \, \mathrm{d}\widetilde{\beta}_\ell(s)\right)\right) \mid \theta_k = \hat{S}_k(t), \ \theta_\ell = \hat{S}_\ell(t)\right] = 0.
\end{aligned}$$

Here, $\widetilde{\mathbb{E}}$ represents the expectation under $\widetilde{\mathbb{P}}$. Thus, we obtain (3.10) under $\widetilde{\mathbb{P}}$, because

$$\widetilde{\mathbb{E}}\left[\int_0^T \sum_{k,\ell=1,\, k\neq\ell}^K \mathbf{1}_{\{\alpha_k\xi_k(t)=\alpha_\ell\xi_\ell(t)\}} \, \mathrm{d}t\right] = \int_0^T \sum_{k,\ell=1,\, k\neq\ell}^K \widetilde{\mathbb{P}}\big(\alpha_k\xi_k(t) = \alpha_\ell\xi_\ell(t)\big) \, \mathrm{d}t = 0.$$

By the reasoning in the previous paragraph, we claim the property (3.1). □

**Corollary 3.1.** *Recall the conic set $\mathcal{R}_k$ defined from the indicator function $\delta_k$ in (2.2) and $\delta_k(\cdot) = \mathbf{1}_{\mathcal{R}_k}(\cdot)$ for $k = 1, \ldots, K$. We denote the closure of $\mathcal{R}_k$ by $\overline{\mathcal{R}_k}$ for every $k$. The stochastic equation*

$$\begin{aligned}
\hat{X}_k(t) = \hat{X}_k(0)F^{\mathrm{c}}_{\mathrm{e}}(t) + \hat{\lambda}_k F_{\mathrm{e}}(t) + \lambda_0 \int_0^t \mathbf{1}_{\overline{\mathcal{R}_k}}(\hat{X}(s))F^{\mathrm{c}}(t-s) \, \mathrm{d}s \\
+ \int_0^t F^{\mathrm{c}}(t-s) \, \mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t), \quad k = 1, \ldots, K
\end{aligned}$$

*has a unique weak solution in $C(\mathbb{R}_+, \mathbb{R}^K)$.*

*Proof.* Thanks to (3.1), the integrals $\int_0^\cdot \mathbf{1}_{\overline{\mathcal{R}_k}}(\hat{X}(s))F^{\mathrm{c}}(t-s) \, \mathrm{d}s$ and $\int_0^\cdot \delta_k(\hat{X}(s))F^{\mathrm{c}}(t-s) \, \mathrm{d}s$ are the same almost surely. □

## 4. Proof for the convergence to the limit

We have the representation

$$\begin{aligned}
\hat{X}^n_k(t) = \hat{X}^n_k(0)F^{\mathrm{c}}_0(t) + \hat{\lambda}^n_k \int_0^t F^{\mathrm{c}}(t-s) \, \mathrm{d}s + \int_0^t \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(t-s) \, \mathrm{d}\hat{A}^n_0(s) \\
+ \int_0^t F^{\mathrm{c}}(t-s) \, \mathrm{d}\hat{A}^n_k(s) + \hat{X}^n_{k,0}(t) + \hat{X}^n_{k,1}(t) + \hat{X}^n_{k,2}(t), \qquad (4.1)
\end{aligned}$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

where

$$\hat{X}^n_{k,0}(t) := \frac{1}{\sqrt{n}} \sum_{j=1}^{X^n_k(0)} \Big(\mathbf{1}_{\eta^0_{k,j}>t} - F^{\mathrm{c}}_{\mathrm{e}}(t)\Big) = -\frac{1}{\sqrt{n}} \sum_{j=1}^{X^n_k(0)} \Big(\mathbf{1}_{\eta^0_{k,j}\le t} - F_{\mathrm{e}}(t)\Big),$$

$$\hat{X}^n_{k,1}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_k(t)} \Big(\mathbf{1}_{\tau^n_{k,i}+\eta_{k,i}>t} - F^{\mathrm{c}}\big(t-\tau^n_{k,i}\big)\Big),$$

$$\hat{X}^n_{k,2}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n_0(t)} \delta_k\big(X^n\big(\tau^n_{0,i}-\big)\big)\Big(\mathbf{1}_{\tau^n_{0,i}+\eta_{0,i}>t} - F^{\mathrm{c}}\big(t-\tau^n_{0,i}\big)\Big).$$

Note that this assumption of $F_0 = F_{\mathrm{e}}$ is essential since we are centering the process $\bar{X}^n_k$ by its equilibrium $\bar{X}^*_k$, and in the derivation of the representation of $X^n_k(t)$, the term $-\sqrt{n}\lambda_k F_0(t)$ cancels out the term $\sqrt{n}\lambda_k \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}s$ only if $F_0 = F_{\mathrm{e}}$.

The joint convergence in the following lemma follows directly from the existing results for each component in [16, 22] for the heavy-traffic analysis of G/GI/∞ queues and the mutual independence of the corresponding limits.

**Lemma 4.1.** *Under Assumptions 2.1, 2.2, and 2.3,*

$$\bigg(\int_0^{\cdot} F^{\mathrm{c}}(\cdot-s)\,\mathrm{d}\hat{A}^n_k(s),\, \hat{X}^n_{k,0},\, \hat{X}^n_{k,1}\bigg)_{k=1,\dots,K} \Rightarrow \bigg(\int_0^{\cdot} F^{\mathrm{c}}(\cdot-s)\,\mathrm{d}\hat{A}_k(s),\, \hat{X}_{k,0},\, \hat{X}_{k,1}\bigg)_{k=1,\dots,K}$$

*in* $(D^{3K}, J_1)$ *as* $n\to\infty$, *where the limits* $\hat{X}_{k,0}$ *and* $\hat{X}_{k,1}$ *are given in Theorem 2.2.*

**Lemma 4.2.** *Under Assumptions 2.1 and 2.2*(i) *and* (iii), $\hat{X}^n_{k,2} \Rightarrow 0$ *in* $(D, J_1)$ *as* $n\to\infty$.

*Proof.* For each $t$, by conditioning on the filtration generated by the arrival process $A^n_0$, we have

$$\mathbb{E}\big[\big(\hat{X}^n_{k,2}(t)\big)^2\big] = \frac{1}{n}\mathbb{E}\Bigg[\sum_{i=1}^{A^n_0(t)} \delta_k\big(X^n\big(\tau^n_{0,i}-\big)\big)F\big(t-\tau^n_{0,i}\big)F^{\mathrm{c}}\big(t-\tau^n_{0,i}\big)\Bigg]$$

$$\le \frac{1}{\sqrt{n}}\mathbb{E}\bigg[\int_0^t F(t-s)F^{\mathrm{c}}(t-s)\,\mathrm{d}\frac{A^n_0(s)}{\sqrt{n}}\bigg] \to 0 \quad \text{as } n\to\infty,$$

where the convergence follows from Assumption 2.2 that $A^n_0/\sqrt{n} \Rightarrow \lambda_0 e$ in $D$.

Next, we consider the increment, for $t, u\ge 0$,

$$\begin{aligned}&\big|\hat{X}^n_{k,2}(t+u) - \hat{X}^n_{k,2}(t)\big|\\ &\le \frac{1}{\sqrt{n}}\sum_{i=1}^{A^n_0(t)} \mathbf{1}_{t<\tau^n_{0,i}+\eta_{0,i}\le t+u} + \frac{1}{\sqrt{n}}\sum_{i=1}^{A^n_0(t)}\big(F\big(t+u-\tau^n_{0,i}\big) - F\big(t-\tau^n_{0,i}\big)\big)\\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=A^n_0(t)+1}^{A^n_0(t+u)} \delta_k\big(X^n\big(\tau^n_{0,i}-\big)\big)\big|\mathbf{1}_{\tau^n_{0,i}+\eta_{0,i}>t+u} - F^{\mathrm{c}}\big(t+u-\tau^n_{0,i}\big)\big|.\end{aligned} \tag{4.2}$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

For the first term, since it is non-decreasing in $u$, we have, for $\delta > 0$ and $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{u\in[0,\delta]} \frac{1}{\sqrt{n}} \sum_{i=1}^{A_0^n(t)} \mathbf{1}_{t<\tau_{0,i}^n+\eta_{0,i}\le t+u} > \frac{\varepsilon}{3}\bigg) \\
&\le \frac{9}{\varepsilon^2}\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n}} \sum_{i=1}^{A_0^n(t)} \mathbf{1}_{t<\tau_{0,i}^n+\eta_{0,i}\le t+\delta}\bigg)^2\Bigg] \\
&\le \frac{18}{\varepsilon^2}\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n}} \sum_{i=1}^{A_0^n(t)} \mathbf{1}_{t<\tau_{0,i}^n+\eta_{0,i}\le t+\delta} - \big(F\big(t+\delta-\tau_{0,i}^n\big) - F\big(t-\tau_{0,i}^n\big)\big)\bigg)^2\Bigg] \\
&\quad + \frac{18}{\varepsilon^2}\mathbb{E}\Bigg[\bigg(\frac{1}{\sqrt{n}} \sum_{i=1}^{A_0^n(t)} \big(F\big(t+\delta-\tau_{0,i}^n\big) - F\big(t-\tau_{0,i}^n\big)\big)\bigg)^2\Bigg] \\
&= \frac{18}{\varepsilon^2}\mathbb{E}\bigg[\frac{1}{\sqrt{n}}\int_0^t (F(t+\delta-s)-F(t-s))(1-(F(t+\delta-s)-F(t-s)))\,\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\bigg] \\
&\quad + \frac{18}{\varepsilon^2}\mathbb{E}\bigg[\bigg(\int_0^t (F(t+\delta-s)-F(t-s))\,\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\bigg)^2\bigg].
\end{aligned}$$

Here, the first term converges to zero as $n \to \infty$, and the second term satisfies

$$\begin{aligned}
&\frac{1}{\delta}\limsup_{N\to\infty}\sup_{t\in[0,T]}\mathbb{E}\bigg[\bigg(\int_0^t (F(t+\delta-s)-F(t-s))\,\mathrm{d}\frac{A_0^n(s)}{\sqrt{n}}\bigg)^2\bigg] \\
&\quad \le \frac{1}{\delta}\sup_{t\in[0,T]}\lambda_0^2\bigg(\int_0^t (F(t+\delta-s)-F(t-s))\,\mathrm{d}s\bigg)^2 \\
&\quad = \frac{1}{\delta}\sup_{t\in[0,T]}\lambda_0^2\bigg(\int_t^{t+\delta} F(s)\,\mathrm{d}s - \int_0^\delta F(s)\,\mathrm{d}s\bigg)^2 \le \lambda_0^2\delta, \qquad (4.3)
\end{aligned}$$

which converges to zero as $\delta \to 0$.

The second term in (4.2) satisfies (4.3). The third term in (4.2) can be bounded by $\big(A_0^n(t+u) - A_0^n(t)\big)/\sqrt{n}$, which is non-decreasing in $u$, so that the supremum over $u \in [0, \delta]$ is bounded by $\big(A_0^n(t+\delta) - A_0^n(t)\big)/\sqrt{n}$. Then, by the convergence of $A_0^n/\sqrt{n} \Rightarrow \lambda_0 e$ in $D$, we obtain that, for small enough $\delta$,

$$\limsup_{N\to\infty}\mathbb{P}\bigg(\sup_{u\in[0,\delta]}\frac{1}{\sqrt{n}}\big(A_0^n(t+u)-A_0^n(t)\big) > \frac{\varepsilon}{3}\bigg) = 0.$$

Thus, by [4, corollary on p. 83], we have shown that

$$\frac{1}{\delta}\limsup_{N\to\infty}\sup_{t\in[0,T]}\mathbb{P}\bigg(\sup_{u\in[0,\delta]}\big|\hat{X}_{k,2}^n(t+u)-\hat{X}_{k,2}^n(t)\big| > \frac{\varepsilon}{3}\bigg) \to 0 \quad \text{as } \delta \to 0.$$

This completes the proof. □

*Proof of Theorem* 2.2. We first observe that (4.1) can be rewritten as

$$\begin{aligned}\hat{Z}^n_k(t) &:= \hat{X}^n_k(t) - \int_0^t \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}^n_0(s)\\ &= \hat{X}^n_k(0)F^{\mathrm{c}}_0(t) + \hat{\lambda}^n_k \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}s\\ &\quad + \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}^n_k(s) + \hat{X}^n_{k,0}(t) + \hat{X}^n_{k,1}(t) + \hat{X}^n_{k,2}(t) \qquad (4.4)\end{aligned}$$

for $k = 1, \ldots, K$, $0 \le t \le T$, and we write $\hat{Z}^n := (\hat{Z}^n_1, \ldots, \hat{Z}^n_k)$. Because of the tightness of the sequence $(\hat{X}^n_k(0), \hat{\lambda}^n_k)_{n\ge 1}$ in $\mathbb{R}^2$ and the tightness of the sequence

$$\big((\hat{A}^n_0(\cdot), (\hat{A}^n_k(\cdot), \hat{X}^n_{k,0}(\cdot), \hat{X}^n_{k,1}(\cdot), \hat{X}^n_{k,2}(\cdot)),\ k = 1, \ldots, K)\big)_{n\ge 1}$$

in $\big(D^{4K+1}_{[0,T]}, J_1\big)$, we claim the tightness of the sequence

$$\big(\hat{X}^n(\cdot), \hat{Z}^n(\cdot), \hat{A}^n_0(\cdot)\big)_{n\ge 1} = \big(\hat{X}^n_1(\cdot), \ldots, \hat{X}^n_K(\cdot), \hat{Z}^n_1(\cdot), \ldots, \hat{Z}^n_K(\cdot), \hat{A}^n_0(\cdot)\big)_{n\ge 1} \qquad (4.5)$$

in $\big(D^{2K+1}_{[0,T]}, J_1\big)$.

Now let us take a weak limit point $\big(\hat{X}^\infty(\cdot), \hat{Z}^\infty(\cdot), \hat{A}^\infty_0(\cdot)\big)$ of the sequence (4.5) in $\big(D^{2K+1}_{[0,T]}, J_1\big)$. Without loss of generality, we may assume that the whole sequence may converge weakly to this limit point. By the Skorokhod representation theorem for the separable metric space $\big(D^{2K+1}_{[0,T]}, J_1\big)$, we may take almost sure convergence

$$\lim_{n\to\infty} \big(\hat{X}^n(t), \hat{Z}^n(t), \hat{A}^n_0(t)\big) = \big(\hat{X}^\infty(t), \hat{Z}^\infty(t), \hat{A}^\infty_0(t)\big) \qquad (4.6)$$

for all but countably many $t$ on $[0, T]$ by extending the probability space if necessary. The filtration for the corresponding probability space is taken to be the one generated by all these processes.

The limits of the right-hand side of (4.4), i.e. the limit of $\hat{Z}^n(\cdot) = (\hat{Z}^n_1(\cdot), \ldots, \hat{Z}^n_K(\cdot))$ can be represented as

$$\hat{Z}^\infty_k(t) = \hat{X}_k(0)F^{\mathrm{c}}_0(t) + \hat{\lambda}_k \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}s + \int_0^t F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_k(s) + \hat{X}_{k,0}(t) + \hat{X}_{k,1}(t) \qquad (4.7)$$

for $1 \le k \le K$, $0 \le t \le T$, and it is continuous on $[0, T]$. Thus, the almost sure convergence of $\hat{Z}^n$ to $\hat{Z}^\infty$ in $\big(D^K_{[0,T]}, J_1\big)$ is uniform on $[0, T]$. With the same reasoning, the almost sure convergence $\lim_{n\to\infty} \hat{A}^n_0(t) = \lambda_0 t$ is also uniform on $[0, T]$, that is,

$$\lim_{n\to\infty} \sup_{0\le t\le T} \big|\hat{A}^n_0(t) - \lambda_0 t\big| = 0.$$

Moreover, since the right-hand side of (4.7) is continuous almost surely in $t \in [0, T]$, so is $\hat{Z}^\infty_k$ for each $k$. Note that the integrand $\delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(\cdot - s)$ of $\int_0^\cdot \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(\cdot - s)\,\mathrm{d}\hat{A}^n_0(s)$ is bounded and $F^{\mathrm{c}}$ is differentiable with a bounded derivative. Then, the sequence $\big\{\int_0^t \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}^n_0(s),\ t \in [0, T], n \ge 1\big\}$ of absolutely continuous functions on $[0, T]$ is uniformly equicontinuous. Thus, the almost sure limit

$$\Phi_k(\cdot) := \lim_{n\to\infty} \int_0^\cdot \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(\cdot - s)\,\mathrm{d}\hat{A}^n_0(s) = \lambda_0 \lim_{n\to\infty} \int_0^\cdot \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(\cdot - s)\,\mathrm{d}s \qquad (4.8)$$

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

is uniformly converging by the Arzelà–Ascoli theorem, and hence it is also continuous in $t \in [0, T]$. Thus, the limit $\hat{X}_k^\infty(t)$ is continuous in $t$, because of (4.4) and the continuity of $\hat{Z}^\infty$. That is, $\hat{X}^\infty(t) = \hat{X}^\infty(t-)$ for every $t \in [0, T]$. Then we have the uniform convergence

$$\lim_{n\to\infty} \sup_{0\le t\le T} |\hat{X}^n(t) - \hat{X}^\infty(t)| = 0,$$

and $\hat{X}^\infty$ satisfies

$$\hat{X}_k^\infty(t) = \Phi_k(t) + \hat{Z}_k^\infty(t), \quad k = 1, \ldots, K, \ t \ge 0, \tag{4.9}$$

where $\Phi_k$ and $\hat{Z}_k^\infty$ are given in (4.8) and (4.7), respectively.

Now we claim that each $\Phi_k$ is absolutely continuous with respect to Lebesgue measure by an application of the Riesz representation theorem for bounded linear functionals. Hence, by a similar argument of the change of measure as in the proof of Theorem 3.1, an analogue of (3.1) holds for $\hat{X}^\infty$:

$$\int_0^T \sum_{k,\ell=1,k\ne\ell}^K \mathbf{1}_{\{\alpha_k \hat{X}_k^\infty(t) = \alpha_\ell \hat{X}_\ell^\infty(t)\}} \,\mathrm{d}t = \int_0^T \mathbf{1}_{\cup_{k=1}^K \partial\mathcal{R}_k}(\hat{X}^\infty(t)) \,\mathrm{d}t = 0. \tag{4.10}$$

Here, $\partial\mathcal{R}_k$ is the boundary of $\mathcal{R}_k$, i.e. $\partial\mathcal{R}_k = \{x \in \mathbb{R}_+^K : \alpha_k x_k = \alpha_\ell x_\ell \text{ for some } \ell\}$ for $k = 1, \ldots, K$.

Since $\delta_k(\cdot)$ is an indicator function of a conic set in (2.2), the almost sure convergence (4.6) of $\hat{X}^n$ implies that the almost convergence

$$\lim_{n\to\infty} \delta_k(\hat{X}^n(s-)) = \delta_k(\hat{X}^\infty(s-)) = \delta_k(\hat{X}^\infty(s))$$

holds if $\hat{X}^\infty(s)$ is not on the boundary $\partial\mathcal{R}_k$ of the set $\mathcal{R}_k$ for $k = 1, \ldots, K$. Thus, thanks to (4.10), the almost sure limit $\Phi_k(t)$ of $\int_0^t \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_0^n(s)$ in (4.4) is

$$\lim_{n\to\infty} \int_0^t \delta_k(\hat{X}^n(s-))F^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_0^n(s) = \lambda_0 \int_0^t \delta_k(\hat{X}^\infty(s-))F^{\mathrm{c}}(t-s)\,\mathrm{d}s, \quad 0 \le t \le T.$$

Hence, we claim that (4.9) is reduced to

$$\hat{Z}_k^\infty(t) = \hat{X}_k^\infty(t) - \lambda_0 \int_0^t \delta_k(\hat{X}^\infty(s))F^{\mathrm{c}}(t-s)\,\mathrm{d}s, \quad 0 \le t \le T, \ 1 \le k \le K.$$

In other words, because of the representation in (4.7), the weak limit point $\hat{X}^\infty(\cdot)$ satisfies the stochastic equation (2.7). Thanks to the weak uniqueness in Theorem 3.1, the distribution of $\hat{X}^\infty(\cdot)$ is uniquely determined. Therefore, $\hat{X}^n(\cdot)$ converges weakly to $\hat{X}(\cdot)$, which satisfies the stochastic equation (2.7) as $n \to \infty$.

## 5. Concluding remarks

In this paper we have studied the non-Markovian parallel infinite-server model with a flexible stream under the scaling of its arrival rate being of order $\sqrt{n}$. It will be interesting to investigate the case where the flexible stream has an arrival rate of order $n$, as also discussed in [10]. It will also be interesting to consider such non-Markovian parallel many-server models with or without abandonment in the Halfin–Whitt regime [12, 23]. The infinite-server model

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press

studied here can be used as an approximation for the many-server models in the underloaded regime, and can also be used as an approximation of the offered load process for the many-server models in the Halfin–Whitt regime. If the stationary distribution of the limiting process could be characterized or approximated, then it could be used for staffing decisions in each queue, as well as approximating the delay probabilities; see for example, the relevant work in [27, 28].

## Appendix A.

*Proof of Remark* 2.3. Recall the expression for $\hat{X}_k(t)$ in (2.8). We have

$$\begin{aligned} &-\hat{X}_k(0)F^{\mathrm{c}}(t)+\hat{\lambda}_kF^{\mathrm{c}}(t)+\lambda_0\delta_k(\hat{X}(t))-\lambda_0\int_0^t\delta_k(\hat{X}(s))f(t-s)\,\mathrm{d}s\\ &\quad=-\hat{X}_k(0)\mathrm{e}^{-t}+\hat{\lambda}_k\mathrm{e}^{-t}+\lambda_0\delta_k(\hat{X}(t))-\lambda_0\int_0^t\delta_k(\hat{X}(s))\mathrm{e}^{-(t-s)}\,\mathrm{d}s\\ &\quad=\hat{\lambda}_k+\lambda_0\delta_k(\hat{X}(t))-\left(\hat{X}_k(0)\mathrm{e}^{-t}+\hat{\lambda}_k(1-\mathrm{e}^{-t})+\lambda_0\int_0^t\delta_k(\hat{X}(s))\mathrm{e}^{-(t-s)}\,\mathrm{d}s\right). \qquad \text{(A.1)}\end{aligned}$$

Writing $\hat{X}_k^A(t)=\int_0^tF^{\mathrm{c}}(t-s)\,\mathrm{d}\hat{A}_k(s)=\int_0^t\mathrm{e}^{-(t-s)}c_k\,\mathrm{d}\hat{B}_k(t)$, we obtain

$$\hat{X}_k^A(t)=-\int_0^t\hat{X}_k^A(s)\,\mathrm{d}s+c_k\hat{B}_k(t). \qquad \text{(A.2)}$$

Recall the representations of $\hat{X}_{k,0}$ in (3.3) and $\hat{X}_{k,1}$ in (3.5).

For the stochastic terms in (A.2), (3.3), and (3.5), since they are mutually independent, we obtain that the covariance function of their summation at times $0\le t<t'$ is equal to

$$\begin{aligned}&\mathrm{Cov}(c_kB_k(t),\,c_kB_k(t'))+\mathrm{Cov}\big(\lambda_k^{1/2}\hat{B}_k(1-\mathrm{e}^{-t}),\,\lambda_k^{1/2}\hat{B}_k(1-\mathrm{e}^{-t'})\big)\\ &\quad+\mathrm{Cov}\left(\int_0^t\int_0^t1_{s+x\le t}\,\mathrm{d}\hat{W}_k(\lambda_ks,\,1-\mathrm{e}^{-x}),\,\int_0^{t'}\int_0^{t'}1_{s+x\le t'}\,\mathrm{d}\hat{W}_k(\lambda_ks,\,1-\mathrm{e}^{-x})\right)\\ &\qquad=c_k^2t+\lambda_k(1-\mathrm{e}^{-t})+\lambda_k(t-(1-\mathrm{e}^{-t}))=(\lambda_k+c_k^2)t.\end{aligned}$$

That is, the three stochastic terms in (A.2), (3.3), and (3.5) are equivalent in distribution to a Brownian motion with variance coefficient $\lambda_k+c_k^2$. In the case of the renewal arrival process, $c_k^2=\lambda_k^3\sigma_k^2=\lambda_k\mathrm{SCV}_k$, with $\mathrm{SCV}_k=\lambda_k^2\sigma_k^2$ being the squared coefficient of variation of the interarrival times, so the variance coefficient $\lambda_k+c_k^2=\lambda_k(1+\mathrm{SCV}_k)$. If, in addition, the arrival processes are Poisson, then $\mathrm{SCV}_k=1$ and the variance coefficient $\lambda_k+c_k^2=2\lambda_k$.

Finally, combining (A.1), (A.2), (3.3), and (3.5), we obtain the expression in (2.9). □

## Funding information

T. Ichiba was supported by NSF Grant DMS 2008427 and G. Pang was supported by NSF Grant DMS 2108683/2216765.

## Competing interests

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

## References

[1] Almada Monter, S. A., Shkolnikov, M. and Zhang, J. (2019). Dynamics of observables in rank-based models and performance of functionally generated portfolios. *Ann. Appl. Prob.* **29**, 2849–2883.

[2] Bass, R. F. and Pardoux, È. (1987). Uniqueness for diffusions with piecewise constant coefficients. *Prob. Theory Relat. Fields* **76**, 557–572.

[3] Berger, M. A. and Mizel, V. J. (1980). Volterra equations with Itô integrals I. *J. Integral Equ.* **2**, 187–245.

[4] Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley, Chichester.

[5] Chao, Y.-J. (2002). Weak convergence of a sequence of semimartingales to a diffusion with discontinuous drift and diffusion coefficients. *Queueing Systems* **42**, 153–188.

[6] Chen, H. and Ye, H.-Q. (2012). Asymptotic optimality of balanced routing. *Operat. Res.* **60**, 163–179.

[7] 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.

[8] Fernholz, E. R. (2002). *Stochastic Portfolio Theory*. Springer, New York.

[9] Filippov, A. F. (1988). *Differential Equations with Discontinuous Right-Hand Sides* (Math. Appl. (Soviet Ser.) **18**). Kluwer, Dordrecht.

[10] Fleming, P. J. and Simon, B. (1999). Heavy traffic approximations for a system of infinite servers with load balancing. *Prob. Eng. Inf. Sci.* **13**, 251–273.

[11] Goldsztajn, D., Borst, S. C., van Leeuwaarden, J. S., Mukherjee, D. and Whiting, P. A. (2022). Self-learning threshold-based load balancing. *INFORMS J. Computing* **34**, 39–54.

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

[13] Ichiba, T., Karatzas, I. and Shkolnikov, M. (2013). Strong solutions of stochastic equations with rank-based coefficients. *Prob. Theory Relat. Fields* **156**, 229–248.

[14] Ichiba, T., Papathanakos, V., Banner, A., Karatzas, I. and Fernholz, R. (2011). Hybrid atlas models. *Ann. Appl. Prob.* **21**, 609–644.

[15] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York.

[16] Krichagina, E. V. and Puhalskii, A. A. (1997). A heavy-traffic analysis of a closed queueing system with a $\mathrm{GI}/\infty$ service center. *Queueing Systems* **25**, 235–280.

[17] Krylov, N. V. (2004). On weak uniqueness for some diffusions with discontinuous coefficients. *Stoch. Process. Appl.* **113**, 37–64.

[18] Krylov, N. V. and Liptser, R. (2002). On diffusion approximation with discontinuous coefficients. *Stoch. Process. Appl.* **102**, 235–264.

[19] Leobacher, G. and Szölgyenyi, M. (2017). A strong order 1/2 method for multidimensional SDEs with discontinuous drift. *Prob. Theory Relat. Fields* **27**, 2383–2418.

[20] Nualart, D. and Pardoux, E. (1994). Markov field properties of solutions of white noise driven quasi-linear parabolic PDEs. *Stoch. Stoch. Reports* **48**, 17–44.

[21] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surv.* **4**, 193–267.

[22] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65**, 325–364.

[23] Reed, J. (2009). The $\mathrm{G}/\mathrm{GI}/N$ queue in the Halfin–Whitt regime. *Ann. Appl. Prob.* **19**, 2211–2269.

[24] Shorack, G. R. and Wellner, J. A. (2009). *Empirical Processes with Applications to Statistics*. SIAM, Philadelphia, PA.

[25] Viterbi, A. J. (1995). *CDMA: Principles of Spread Spectrum Communication*. Addison Wesley, Boston, MA.

[26] Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. Springer, New York.

[27] Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Res. Logistics* **54**, 476–484.

[28] Whitt, W. (2013). Offered load analysis for staffing. *Manuf. Serv. Operat. Manag.* **15**, 166–169.

[29] Yao, H. and Knessl, C. (2005). On the infinite server shortest queue problem: Symmetric case. *Stoch. Models* **21**, 101–132.

[30] Yao, H. and Knessl, C. (2006). On the infinite server shortest queue problem: Non-symmetric case. *Queueing Systems* **52**, 157–177.

https://doi.org/10.1017/jpr.2025.10046 Published online by Cambridge University Press