Stochastic Systems
2016, Vol. 6, No. 2, 519–600
DOI: 10.1214/15-SSY206

HEAVY-TRAFFIC LIMITS FOR A FORK-JOIN NETWORKIN THE HALFIN-WHITT REGIME

By Hongyuan Lu and Guodong Pang

The Pennsylvania State University

We study a fork-join network with a single class of jobs, which are forked into a fixed number of parallel tasks upon arrival to be processed at the corresponding multi-server stations. After service completion, each task will join a buffer associated with the service station waiting for synchronization, called "unsynchronized queue". The synchronization rule requires that all tasks from the same job must be completed, referred to as "non-exchangeable synchronization". Once synchronized, jobs will leave the system immediately. Service times of the parallel tasks of each job can be correlated and form a sequence of i.i.d. random vectors with a general continuous joint distribution function. We study the joint dynamics of the queueing and service processes at all stations and the associated unsynchronized queueing processes.

The main mathematical challenge lies in the "resequencing" of arrival orders after service completion at each station. As in Lu and Pang (2015) for the infinite-server fork-join network model, the dynamics of all the aforementioned processes can be represented via a multiparameter sequential empirical process driven by the service vectors for the parallel tasks of each job. We consider the system in the Halfin-Whitt regime, and prove a functional law of large number and a functional central limit theorem for queueing and synchronization processes. In this regime, although the delay for service at each station is asymptotically negligible, the delay for synchronization is of the same order as the service times.

1. Introduction. We consider a fundamental fork-join network with a single class of jobs that will fork into a fixed number of parallel tasks upon their arrival, and then join after service completion. Each parallel task is processed at a multi-server station under the first-come-first-serve (FCFS) and non-idling service discipline, and will join a buffer waiting for synchronization associated with the station after service completion. This buffer is called "unsynchronized queue" or "waiting buffer for synchronization". Tasks are

Received October 2015.

MSC 2010 subject classifications: 60F17, 60H20, 60K25, 60K30, 90B15, 90B22.

Keywords and phrases: Fork-join networks, non-exchangeable synchronization, resequencing, Halfin-Whitt (QED) regime, functional law of large numbers (FLLN), functional central limit theorem (FCLT), multiparameter sequential empirical process, generalized multiparameter Kiefer process.

only synchronized if all the parallel tasks of the same job are completed, called "non-exchangeable synchronization" (NES) [3, 60, 61, 33]. After synchronization, jobs will leave the system immediately (the synchronization time is irrelevant in our model). Figure 1 depicts such a network model. Unlike classical queueing models, there are two types of delays in this fork-join network: delay for service and delay for synchronization. The objective of this paper is to study the delay for synchronization when each service station is operating in the Halfin-Whitt (Quality-and-Efficiencty-Driven, QED) regime [18]. In this regime, the job arrival rate and the number of servers in each service station get large appropriately while fixing service time distributions so that each station is asymptotically critically loaded, achieving both high quality (low delay) and high efficiency (high utilization).

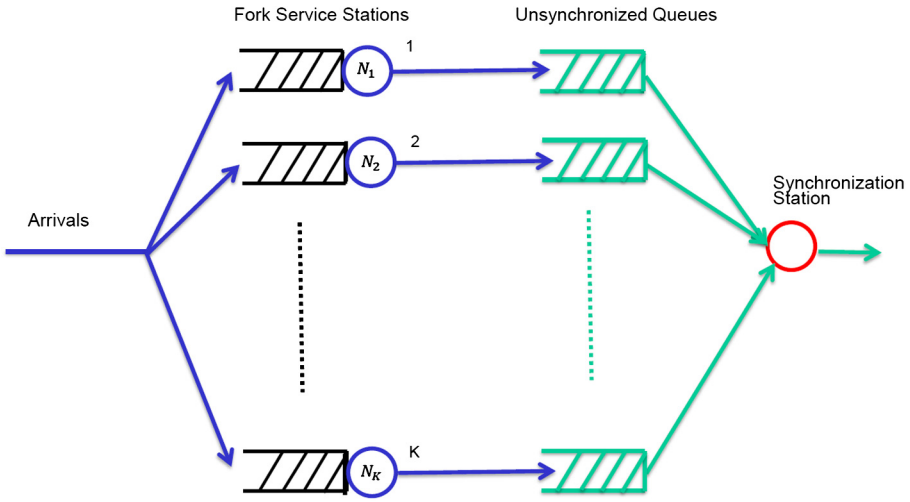

FIG 1. *A fundamental fork-join network with finite servers.*

Fork-join networks with NES are used in many applications, including healthcare systems, parallel computing, MapReducing scheduling (e.g., large-scale parallel Web search), disassembly and reassembly systems in manufacturing and so on. In patient flows of hospitals [2, 3, 22, 60, 61], the treatment and discharge processes are typical examples of fork-join networks with NES: a patient must have all test results ready before a doctor examination and these tests are conducted in different units/laboratories and can never be mixed; a patient, after the discharge decision is made, must wait for necessary procedures, pharmacy, transportation, etc., before being physically discharged. In MapReduce scheduling [11, 31, 54, 57], jobs are processed in two phases: in the map phase, a large-scale data input (e.g.,

Web processing data) is distributed into individual computation nodes, and each node processes one block of input data, and after the execution of all blocks of the same data input, they will be joined as an output in the reduce phase. In addition, fork-join networks with NES are also used in manufacturing and inventory systems [5, 6, 17, 27, 28, 40, 41, 49, 48, 53, 56], military operations [26, 59] and law reinforcement [30].

The main mathematical challenge in analyzing the multi-server fork-join network with NES is the resequencing of arrival orders after service completion at each service station due to the randomness of service times. Exact analysis of this model is prohibitively difficult since it is necessary to track the service completion times of all the parallel tasks of each job, which will require an infinite dimensional state space. Many efforts have been made to study the resequencing problems in the literature using the max-plus recursions [21, 4, 12]. Here we develop a completely new approach to study the resequencing problem in the fork-join networks with NES asymptotically when each station is operating in the Halfin-Whitt regime.

In [33], we have studied a fork-join network with NES as described above where each service station is operating in the quality-driven (QD) regime (equivalent to having infinite numbers of servers at each station asymptotically). The approach developed in [33] solves the resequencing problem when the number of servers at each station is infinite (no delay for service). However, it cannot be extended directly to resolve the resequencing problem when the number of servers at each station is finite. As shown in [51, 50, 24, 25], the queueing process for service itself in $G/GI/N$ queues in the Halfin–Whitt regime already present substantial difficulties. In our model, the delay for service also affects the resequencing of tasks after service completion at each station, and as a consequence, the queueing processes for synchronization. This complexity requires further development of the methodology in [33].

In this paper we aim to solve the resequencing problem when all service stations have multiple servers, operating in the Halfin-Whitt regime. Since the service times for the parallel tasks for a job are correlated, the service completion processes of the parallel tasks are dependent, which causes a substantial amount of difficulties in the analysis of the resequencing of the parallel tasks and the synchronization process, as well as the service dynamics at all parallel stations jointly. In our approach, the key is a representation of the service processes, the unsynchronized queueing processes and the synchronized process via functionals of a multiparameter sequential empirical process driven by the service vectors for the parallel tasks as well as the arrival process and the initial quantities. With this representation, we first

show a functional law of large numbers (FLLN), Theorem 3.1, for these processes assuming that the system starts from empty when the arrival rate is allowed to be time dependent. The fluid limit of the synchronized process is an integral of the minimum of the fluid entering service processes at all stations with respect to the joint service time distribution function. Numerical examples are provided to illustrate the fluid approximations in §3.1. We then prove a functional central limit theorem (FCLT), Theorem 4.1, for these processes when the arrival rate is constant in the Halfin-Whitt regime and when the number of parallel tasks is equal to two, under some stationarity conditions on the initial quantities. The limits of the diffusion-scaled processes are the unique solution to a set of stochastic integral equations driven by the corresponding multiparameter Kiefer process, the arrival limit process and the limiting initial quantities. One important term in the limits of the synchronized process and the unsynchronized queues is an integral of the limit of the diffusion-scaled minimum of "entering service" processes at both stations with respect to the joint service time distribution.

Our results show that when all service stations operate in the Halfin-Whitt regime and the arrival rate is scaled as $O(n)$, the numbers of tasks in the service stations and the numbers of tasks waiting for synchronization are of the same order, $O(n)$. This implies that waiting times for synchronization are $O(1)$, although waiting times for service are $O(1/\sqrt{n})$. This is an extremely important insight for the management of multi-server fork-join networks with NES in the Halfin-Whitt regime. An intuitive interpretation is that in steady state, for jobs whose tasks are waiting in the associated buffer(s) for synchronization, their other parallel tasks must be already in service with probability one asymptotically. Therefore, in order to minimize the response time - the time duration from the arrival time to synchronization, we conjecture that one must prioritize tasks in each service station dynamically to reduce the waiting time for synchronization to a smaller order.

1.1. *Literature review and comparisons.* Many studies on fork-join networks with synchronization constraints focus on service stations with a single server; see, e.g., [15, 16, 5, 6, 21, 4, 12, 27, 28, 40, 41, 49, 48, 53, 56] and references therein. We remark that in the single-server models with FCFS discipline, the NES constraint is equivalent to the exchangeable synchronization (ES), and thus, resequencing is not the mathematical challenge. A single-class single-server fork-join network with feedback is recently studied in Atar et al. [3], where resequencing becomes the main mathematical challenge due to task feedback. In the conventional heavy-traffic regime,

they show that the system dynamics under NES and ES constraints become asymptotically equivalent under a dynamic priority routing policy.

Very limited work has been done for fork-join networks with multi-server service stations under the NES constraint. Ko and Serfozo [27] studied a single-class multi-server fork-join model with NES as depicted in Figure 1, where the arrival process is Poisson and service times are independent exponential, but their focus is on obtaining approximations for the steady-state system response time. In [10] an exact simulation algorithm is provided for the same Markovian model. Recently, in [8], an exact sampling algorithm is developed to simulate the stationary distribution for a multi-server fork-join model with NES that has renewal arrivals and i.i.d. service vectors. Zaied [60] studied multiclass fork-join networks with NES, which have time-inhomogeneous Poisson arrivals and infinite-server service stations, focusing on the calculation of mean offered load functions. In addition to the work in [3], Zviran [61] also studied the fork-join network with NES in Figure 1 with exponential service times under the conventional heavy-traffic regime, and proved that the system dynamics under NES and ES are asymptotically equivalent under the FCFS discipline.

To the best of our knowledge, our work is the first to study (non-Markovian) multi-server fork-join networks with NES in the Halfin-Whitt regime. As mentioned earlier, we have developed an approach to study such networks in the QD regime in [33]. Specifically, we have shown that the service processes and the queueing processes for synchronization can be represented with a multiparameter sequential empirical process driven by the service vectors for the parallel tasks of jobs, and as a consequence, can be approximated by a multidimensional Gaussian process as a functional of the corresponding multiparameter generalized Kiefer process driven by the service vectors. In [34] and [35], we have further developed the model and methodology in [33]. We have studied the infinite-server fork-join network model with NES, where both the arrival and service processes are modulated by a finite-state continuous-time Markov chain in [34]. For the infinite-server fork-join network model with NES, when the service vectors of the parallel tasks satisfy the strong-mixing (α-mixing) condition, a new approach has been developed to prove the weak convergence of the aforementioned queueing processes in [35]. In that model, the service component of the limit process is driven by a multiparameter generalized Kiefer process accounting for the sequential correlations among the service vectors. In addition, we have also studied in the fork-join network model with disruptive services, which results in an additional jump component in the limit process, requiring the Skorohod M_1 topology for the weak convergence.

Our approach in this paper is based on the conjecture that the system dynamics (queueing, service, waiting for synchronization, and synchronization) in the multi-server fork-join network model with NES can be represented via the corresponding infinite-server model dynamics, which is studied in [33]. However, to prove this conjecture, it requires novel methods to take into account the multidimensionality and the dependence of the service dynamics at all the service stations. Our approach is much relevant to the recent development in the study of $G/GI/N$ queues in the Halfin-Whitt regime. In particular, Reed [51] proved an FCLT for the diffusion-scaled process counting the number of jobs in the $G/GI/N$ queues in the Halfin-Whitt regime, under certain conditions on the initial quantities. He developed a novel approach to represent the finite-server model dynamics via the corresponding infinite-server model dynamics, generalizing the approach for an infinite-server model developed by Krichagina and Puhalskii [29]. Puhalskii and Reed [50] extended that approach to allow for more general initial conditions and non-critical loading, proving the convergence of finite-dimensional distributions for the process counting the number of jobs in the $G/GI/N$ queues. Our work generalizes the methodology in Reed [51] for many-server queues to the multi-server fork-join network model with NES in the Halfin-Whitt regime.

1.2. *Notation.* Throughout the paper, the following notation will be used. $\mathbb{R}$ and $\mathbb{R}_+$ ($\mathbb{R}^d$ and $\mathbb{R}^d_+$, respectively) denote sets of real and real non-negative numbers (d-dimensional vectors, respectively, $d \geq 2$). $\mathbb{Z}_+$ is the set of non-negative integers. $\mathbb{N}$ denotes the set of natural numbers. For $a, b \in \mathbb{R}$, we denote $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$. For $x \in \mathbb{R}$, let $x^+ := \max\{x, 0\}$ and $x^- := -\min\{x, 0\}$. For any $x \in \mathbb{R}_+$, $\lfloor x \rfloor$ is used to denote the largest integer less than or equal to x. We use bold letter to denote a vector, e.g., $\boldsymbol{x} := (x_1, ..., x_N) \in \mathbb{R}^N$. $\mathbf{0}$ denotes the vector whose components are all 0. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, we denote $\boldsymbol{x} \leq \boldsymbol{y}$, $\boldsymbol{x} \geq \boldsymbol{y}$ and $\boldsymbol{x} > \boldsymbol{y}$ in the componentwise sense, and let $\boldsymbol{x} \wedge \boldsymbol{y} = (x_1 \wedge y_1, ..., x_N \wedge y_N)$. We use $\mathbf{1}(A)$ to denote the indicator function of a set A. The abbreviation *a.s.* means almost surely. For any univariate distribution function $F(\cdot)$, we denote $F^c(\cdot) = 1 - F(\cdot)$. For $\boldsymbol{\alpha} \in \mathbb{R}^2_+$ and $\alpha \in \mathbb{R}_+$, we call $\Delta_{\boldsymbol{\alpha}}(\delta)$ (*resp.* $\Delta_\alpha(\delta)$) is a δ-grid of $[0, \alpha_1] \times [0, \alpha_2]$ (*resp.* $[0, \alpha]$), if $\Delta_{\boldsymbol{\alpha}}(\delta)$ (*resp.* $\Delta_\alpha(\delta)$) is a finite partition of $[0, \alpha_1] \times [0, \alpha_2]$ (*resp.* $[0, \alpha]$), where each element of the partition is the rectangle $[s_1, t_1) \times [s_2, t_2)$ (*resp.* $[s, t)$), satisfying $0 \leq s_k < t_k < \alpha_k$ for $k = 1, 2$ (*resp.* $0 \leq s < t$), and $\min_{k=1,2}(t_k - s_k) \geq \delta$ (*resp.* $t - s \geq \delta$). For two real-valued functions f and g, we write $f(x) = O(g(x))$ if $\limsup_{x \to \infty} |f(x)/g(x)| < \infty$.

All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, P)$. For any two complete separable metric spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, we

denote $\mathcal{S}_1 \times \mathcal{S}_2$ as their product space, endowed with the maximum metric, i.e., the maximum of two metrics on $\mathcal{S}_1$ and $\mathcal{S}_2$. $\mathcal{S}^k$ is used to represent k-fold product space of any complete and separable metric space $\mathcal{S}$ for $k \in \mathbb{N}$. For a complete separable metric space $\mathcal{S}$, $\mathbb{D}([0,\infty),\mathcal{S})$ denotes the space of all $\mathcal{S}$-valued càdlàg functions on $[0,\infty)$, and is endowed with the Skorohod J_1 topology (see, e.g., [7, 14, 58]). Denote $\mathbb{D} \equiv \mathbb{D}([0,\infty),\mathbb{R})$. The space $\mathbb{D}([0,\infty),\mathbb{D})$, denoted as $\mathbb{D}_{\mathbb{D}}$, is endowed with the Skorohod J_1 topology, that is, both inside and outside $\mathbb{D}$ spaces are endowed with the Skorohod J_1 topology. For a complete separable metric space $\mathcal{S}$, the space $\mathbb{D}([0,\infty)^2,\mathcal{S})$ is the space of all $\mathcal{S}$-valued "continuous from above with limits from below" functions on $[0,\infty)^2$, and is endowed with the same metric as defined by [19]. $\mathbb{D}_2 \equiv \mathbb{D}([0,1]^2,\mathbb{R})$ is denoted as the space of all "continuous from above with limits from below" functions on the unit square $[0,1]^2$ in the sense of Neuhaus [39], and is endowed with the same metric $d_{\mathbb{D}_2}$ as in [39]. Weak convergence of probability measures μ_n to μ will be denoted as $\mu_n \Rightarrow \mu$. For a sequence of processes $\{\mathcal{X}^n : n \geq 1\}$ and a process $\mathcal{X}$, we use notation $\mathcal{X}^n \overset{d_f}{\Rightarrow} \mathcal{X}$ to denote the convergence in finite-dimensional distributions of $\mathcal{X}^n$ to $\mathcal{X}$.

1.3. *Organization of the paper.* The paper is organized as follows. In §2, we provide a detailed description of the model. In §3, we present the FLLN for the system dynamics, and provide numerical examples in §3.1. The FLLN is proved in §5. We state the FCLT for the system dynamics in §4 and provide its proof in §6. We make some concluding remarks in §7. Some additional proofs are collected in the Appendix.

2. The multi-server fork-join network model. In this section, we present a detailed description of our model. We consider a fork-join network with a single class of jobs, and each job is forked into K ($K > 1$) parallel tasks. Each task is processed in a service station with finite servers under the non-idling FCFS discipline. Namely, a newly arriving task immediately gets served if there is an idle server in that station, and joins the back of the queue otherwise, and the task waiting for the longest in the queue enters service as soon as a server in that station becomes available. After service completion, each task will join a waiting buffer for synchronization associated with each service station, and when all tasks of the same job are completed, they will be synchronized and leave the system. Here we assume that the synchronization process takes zero amount of time.

Let $A := \{A(t) : t \geq 0\}$ be the arrival process of jobs after time 0. Let τ_i be the arrival time of the i^{th} job, $i \in \mathbb{N}$, that is, $A(t) = \max\{i \geq 1 : \tau_i \leq t\}$

for $t > 0$ and $A(0) = 0$. Let N_k be the number of servers at service station k, $k = 1, ..., K$. Each job brings in a K-dimensional service vector, representing the service time at each service station, which can be correlated. Let $\boldsymbol{\eta}^i := (\eta_1^i, ..., \eta_K^i)$ be the service vector of the job that arrives at time τ_i, $i \in \mathbb{N}$, where η_k^i is the service time at the k^{th} service station. We assume that the sequence $\{\boldsymbol{\eta}^i : i \geq 1\}$ is i.i.d., and let the joint distribution function of $\boldsymbol{\eta}^i$ be $F(\boldsymbol{x}) = F(x_1, ..., x_K)$ for $x_k \geq 0$, $k = 1, ..., K$. Let $F^c(\boldsymbol{x}) := P(\eta_1^i > x_1, ..., \eta_K^i > x_K)$, for $x_1, ..., x_K \geq 0$. Their marginal distributions are $F_k(\cdot)$ with mean $1/\mu_k \in (0, \infty)$, for $k = 1, ..., K$. Let $\eta_m^i := \max\{\eta_1^i, ..., \eta_K^i\}$ and $F_m(x) := P(\eta_m^i \leq x) = P(\eta_j^i \leq x, \forall j) = F(x, ..., x)$ for $x \geq 0$. (Throughout this paper, we use "m" to index quantities and processes associated with the maximum.) We make a regularity assumption on the service time distributions for the parallel tasks. It is worth noting that in [50] and [51], the service time distribution is allowed to be general for $G/GI/N$ queues. Here we require the continuity of the joint distribution function F, which is necessary for Proposition 4.1 and the proof of the weak convergence in (6.55), and thus the weak convergence in Theorems 3.1 and 4.1. As a consequence, all the limits in the fluid and diffusion scales are continuous.

ASSUMPTION 1. *The joint distribution function $F(\boldsymbol{x})$ of the service time vector $\boldsymbol{\eta}^i$, $i \in \mathbb{N}$, is continuous.*

State Descriptors. Let $X_k := \{X_k(t) : t \geq 0\}$ be the process counting the number of tasks at the service station k, and $Y_k := \{Y_k(t) : t \geq 0\}$ be the process counting the number of tasks in the waiting buffer for synchronization (unsynchronized queue) after service completion at service station k, $k = 1, ..., K$. Denote $\boldsymbol{X} := (X_1, ..., X_K)$ and $\boldsymbol{Y} := (Y_1, ..., Y_K)$. Let $S := \{S(t) : t \geq 0\}$ be the process counting the number of synchronized jobs by each time $t \geq 0$. In addition, let $Q_k := \{Q_k(t) : t \geq 0\}$ and $B_k := \{B_k(t) : k \geq 0\}$ be the processes representing the queue length and the number of tasks in service at station k, respectively, $k = 1, ..., K$. Let $D_k := \{D_k(t) : t \geq 0\}$ be the cumulative service completion (departure) process at service station k, $k = 1, ..., K$. Denote $\boldsymbol{Q} := (Q_1, ..., Q_K)$, $\boldsymbol{B} := (B_1, ..., B_K)$, and $\boldsymbol{D} := (D_1, ..., D_K)$.

A Sequence of Systems. We consider a sequence of the above fork-join networks, indexed by superscript n and let $n \to \infty$. We assume that each service station is operating in the many-server heavy-traffic asymptotic regimes, where the arrival rate of jobs and the number of servers get large appropriately while the service time distributions are fixed. In establishing the FLLN, we allow the arrival rate to be time-dependent. In establishing the FCLT, we will assume that each service station is operating in the Halfin-

Whitt (QED) regime, so that it is critically loaded with a constant arrival rate (see Assumption 4 for the precise definition). For any process $\mathcal{X}$, we use $\mathcal{X}^n$ to represent the associated process in the sequence of the fork-join networks.

Some Fundamental Flow Balance Equations. For each service station k, $k = 1, ..., K$, and for each $t \geq 0$, we have the following flow conservation equations:

$$X_k^n(t) = B_k^n(t) + Q_k^n(t), \tag{2.1}$$

$$X_k^n(t) = X_k^n(0) + A^n(t) - D_k^n(t), \tag{2.2}$$

$$Y_k^n(t) = Y_k^n(0) + D_k^n(t) - S^n(t). \tag{2.3}$$

The non-idling condition implies that for each $k = 1, ..., K$ and $t \geq 0$,

$$B_k^n(t) = X_k^n(t) \wedge N_k^n, \quad Q_k^n(t) = (X_k^n(t) - N_k^n)^+. \tag{2.4}$$

In addition, we have the following flow balance equation, for each $k, k' = 1, ..., K$, $k \neq k'$, and $t \geq 0$,

$$X_k^n(t) + Y_k^n(t) = X_{k'}^n(t) + Y_{k'}^n(t), \tag{2.5}$$

that is, the total numbers of tasks in each service station and its associated waiting buffer for synchronization are equal at all time, and are equal to the total number of jobs in the system.

3. Fluid limit. In this section, we present the fluid limit for the fork-join network. We assume that the system starts from empty and allow the arrival rate to be time-dependent.

ASSUMPTION 2. *There exists a continuous nondecreasing deterministic real-valued function $\bar{a}(t)$ on $[0, \infty)$ with $\bar{a}(0) = 0$ such that*

$$\bar{A}^n(t) := n^{-1} A^n(t) \Rightarrow \bar{a}(t) \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty. \tag{3.1}$$

We also make the following assumption on the numbers of servers.

ASSUMPTION 3. *For $k = 1, ..., K$, $\bar{N}_k^n := N_k^n / n \to N_k > 0$ as $n \to \infty$.*

Under the empty initial condition, we can write the processes $X_k^n(t)$, $Y_k^n(t)$, $k = 1, ..., K$, and $S^n(t)$ as

$$X_k^n(t) = \sum_{i=1}^{A^n(t)} \mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i > t), \tag{3.2}$$

$$Y_k^n(t) = \sum_{i=1}^{A^n(t)} \mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i \le t,\ \tau_i^n + w_{k'}^{n,i} + \eta_{k'}^i > t, \text{ for some } k' \ne k), \tag{3.3}$$

$$S^n(t) = \sum_{i=1}^{A^n(t)} \mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i \le t,\ \forall k = 1, ..., K), \tag{3.4}$$

for $t \ge 0$, where $w_k^{n,i}$ is the waiting time of the i^{th} arrival at station k, $i \in \mathbb{N}$.

In addition, for $k = 1, ..., K$, let $E_k^n(t)$ be the number of tasks that have entered service at station k by time t, $t \ge 0$, and set $E_k^n := \{E_k^n(t) : t \ge 0\}$. Denote $\boldsymbol{E}^n := (E_1^n, ..., E_K^n)$. For each service station $k = 1, .., K$, we also have the balance equation

$$E_k^n(t) = A^n(t) - Q_k^n(t) = A^n(t) - (X_k^n(t) - N_k^n)^+, \quad t \ge 0.$$

Define the fluid-scaled processes $\bar{\mathcal{X}}^n := n^{-1}\mathcal{X}^n$ for $\mathcal{X}^n = \boldsymbol{X}^n, \boldsymbol{Y}^n, S^n, \boldsymbol{E}^n, \boldsymbol{Q}^n, \boldsymbol{B}^n, \boldsymbol{D}^n$. We now state the FLLN for the fluid-scaled processes.

THEOREM 3.1. *Under Assumptions 1-3,*

$$(\bar{A}^n, \bar{\boldsymbol{X}}^n, \bar{\boldsymbol{Y}}^n, \bar{S}^n, \bar{\boldsymbol{E}}^n, \bar{\boldsymbol{Q}}^n, \bar{\boldsymbol{B}}^n, \bar{\boldsymbol{D}}^n) \Rightarrow (\bar{a}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}}, \bar{S}, \bar{\boldsymbol{E}}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{B}}, \bar{\boldsymbol{D}})$$

in $\mathbb{D}^{6K+2}$ *as* $n \to \infty$, *where the limits are all deterministic continuous functions:* $\bar{a}$ *is the limit in (3.1),* $(\bar{\boldsymbol{E}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{Y}}, \bar{S})$ *is the unique solution to the following: for* $t \ge 0$ *and* $k = 1, ..., K$,

$$\bar{X}_k(t) = \int_0^t F_k^c(t-s)d\bar{a}(s) + \int_0^t (\bar{X}_k(t-s) - N_k)^+ dF_k(s), \tag{3.5}$$

$$\bar{E}_k(t) = \bar{a}(t) - (\bar{X}_k(t) - N_k)^+, \tag{3.6}$$

$$\bar{S}(t) = \int_0^t ... \int_0^t \left(\min_{k=1,...,K} \left\{ \bar{E}_k(t - s_k) \right\} \right) dF(s_1, ..., s_K), \tag{3.7}$$

$$\bar{Y}_k(t) = \int_0^t F_k(t-s)d\bar{a}(s) - \int_0^t (\bar{X}_k(t-s) - N_k)^+ dF_k(s) - \bar{S}(t), \tag{3.8}$$

and the limits $\bar{\boldsymbol{Q}}$, $\bar{\boldsymbol{B}}$ *and* $\bar{\boldsymbol{D}}$ *satisfy*

$$\bar{Q}_k(t) = (\bar{X}_k(t) - N_k)^+,\ \bar{B}_k(t) = \bar{X}_k(t) \wedge N_k,\ \bar{D}_k(t) = \bar{a}(t) - \bar{X}_k(t). \tag{3.9}$$

It is easy to check that for each $k = 1, ..., K$, the limit $\bar{X}_k(t)$ also satisfies the following equation:

$$\bar{X}_k(t) = \bar{a}(t) - \int_0^t \bar{E}_k(t-s)dF_k(s), \quad t \geq 0. \tag{3.10}$$

We remark that the fluid limit $\bar{X}_k$ for each $k = 1, ..., K$ depends only on the marginal distribution F_k, while the fluid limits $\bar{Y}_k$, $k = 1, ..., K$, and $\bar{S}$ depend on the joint distribution F. Specifically, each entering service fluid limit $\bar{E}_k(t)$ depends on the marginal distribution F_k, and the fluid limit $\bar{S}$ is a multivariate integral of the minimum of the entering service fluid limits with respect to the joint distribution function F. Since $\bar{Y}_k(t) = \bar{D}_k(t) - \bar{S}(t) = \bar{a}(t) - \bar{X}_k(t) - \bar{S}(t)$ for $t \geq 0$ and $k = 1, ..., K$, it is a functional of both F_k and F. However, as the FCLT (Theorem 4.1) below shows, the limits for all these processes in the diffusion scale will depend on the joint distribution F.

When $\bar{a}(t) = \int_0^t \lambda(s)ds$ and the service times are exponential (independent or dependent), where $\lambda(\cdot)$ is a positive function, for each $k = 1, ..., K$, the fluid limit $\bar{X}_k$ in (3.5) and (3.10) becomes an ordinary differential equation (ODE) [42], but the fluid limit $\bar{Y}_k$ in (3.8) does not have an ODE representation due to the dependence of the fluid limit $\bar{S}$ upon the minimum of the entering service fluid limits of all the parallel stations.

When the arrival rate is constant and each service station is underloaded or critically loaded, we give a corollary on the steady states of the fluid limits. The proof follows from a direct calculation and is omitted. It is evident that correlation among service times of parallel tasks only affects the steady state of $\bar{\boldsymbol{Y}}$ but not that of $\bar{\boldsymbol{X}}$.

COROLLARY 3.1. *Under Assumptions 1-3, if the arrival rate is constant,* $\bar{a}(t) = \lambda t$, *for* λ *satisfying* $0 < \lambda \leq N_k\mu_k$ *for all* $k = 1, ..., K$,

$$(\bar{\boldsymbol{X}}(t), \bar{\boldsymbol{Y}}(t), \bar{\boldsymbol{Q}}(t), \bar{\boldsymbol{B}}(t)) \to (\bar{\boldsymbol{X}}(\infty), \bar{\boldsymbol{Y}}(\infty), \bar{\boldsymbol{Q}}(\infty), \bar{\boldsymbol{B}}(\infty)) \quad as \quad t \to \infty,$$

and

$$\frac{1}{t}(\bar{\boldsymbol{D}}(t), \bar{\boldsymbol{E}}(t), \bar{S}(t)) \to \boldsymbol{\lambda} := (\lambda, ..., \lambda) \quad as \quad t \to \infty,$$

where

$$\bar{X}_k(\infty) = \bar{B}_k(\infty) = \lambda E[\eta_k^1] = \lambda/\mu_k, \ \bar{Y}_k(\infty) = \lambda(E[\eta_m^1] - E[\eta_k^1]), \ \bar{Q}_k(\infty) = 0.$$

3.1. *Numerical Examples.* We give two numerical examples to show the effectiveness of fluid approximations comparing with simulations, when $K = 2$. We let the arrival process be Poisson with time-varying rate $\lambda(t) =$

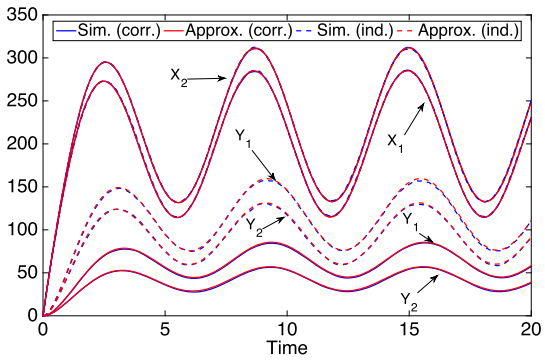

FIG 2. *Comparison of fluid approximations with simulations when the service times have the Marshall-Olkin bivariate exponential distributions with correlation* $\rho = 0$ *and* $\rho = 0.5$. *The figure shows the simulated values of* X_i *and* Y_i, $i = 1, 2$, *and their corresponding fluid approximations (solid lines for* $\rho = 0.5$ *and dashed lines for* $\rho = 0$*). The values for the* X_i *are the same when* $\rho = 0$ *and* $\rho = 0.5$, *while the values of* Y_i *when* $\rho = 0.5$ *are smaller than those when* $\rho = 0$.

$200 + 120\sin(t)$, $t \geq 0$. The numbers of servers in stations 1 and 2 are $N_1 = 300$ and $N_2 = 340$, respectively. In the first numerical example, the service times of the two parallel tasks are assumed to have a bivariate Marshall-Olkin exponential distribution [37]. A bivariate Marshall-Olkin exponential distribution function $F(x, y)$ for the random vector (η_1, η_2) can be written as $F^c(x, y) := P(\eta_1 > x, \eta_2 > y) = \exp(-\mu_1 x - \mu_2 y - \mu_{12}(x \vee y))$, $x, y \geq 0$, where three parameters μ_1, μ_2, μ_{12} are such that the two marginals are exponential with rates $\mu_1 + \mu_{12}$ and $\mu_2 + \mu_{12}$ and their correlation $\rho = \mu_{12}/(\mu_1 + \mu_2 + \mu_{12}) \in [0, 1]$. We denote $MO(\lambda_1, \lambda_2, \rho)$ for a bivariate Marshall-Olkin exponential distribution, where λ_1 and λ_2 are the rates for the marginals, and ρ is the correlation parameter, for which the parameters $\mu_1 = (\lambda_1 - \rho\lambda_2)/(1 + \rho)$, $\mu_2 = (\lambda_2 - \rho\lambda_1)/(1 + \rho)$ and $\mu_{12} = (\rho(\lambda_1 + \lambda_2))/(1 + \rho)$. For our first numerical example, we set the service times to be $MO(1, 0.9, \rho)$ such that the service times of the two parallel tasks have exponential marginals with means 1 and 10/9 in stations 1 and 2, respectively, and their correlation is ρ. The numerical results with $\rho = 0$ and $\rho = 0.5$ are provided in Figure 2, marked with "ind." and "corr.", respectively. In the second numerical example, we let the service times of the two parallel tasks have a bivariate Marshall-Olkin hyperexponential distribution [44], which is a mixture of two independent bivariate Marshall-Olkin exponential distributions. Specifically, we take a mixture of $MO(4/5, 1, \rho_1)$ with probability 0.4 and $MO(6/5, 27/32, \rho_2)$ with probability 0.6, such that the two parallel service times have hyperexponential marginals with the same means as the first example. By setting $\rho_1 = \rho_2 = 0$,

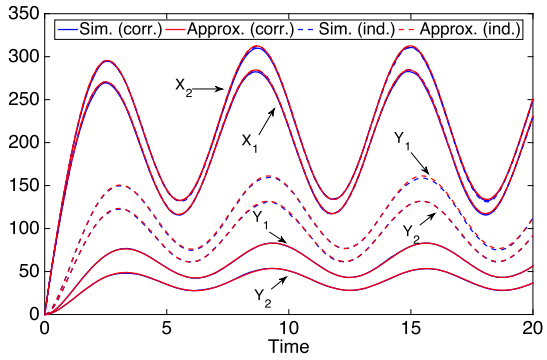

FIG 3. *Comparison of fluid approximations with simulations when the service times have the Marshall-Olkin bivariate hyperexponential distributions with correlation* $\rho = 0$ *and* $\rho = 0.5$ *(solid lines for* $\rho = 0.5$ *and dashed lines for* $\rho = 0$*). The figure shows the simulated values of* X_i *and* Y_i*,* $i = 1, 2$*, and their corresponding fluid approximations. The same observations can be made as in Figure 2.*

we have two independent parallel service times, and by setting $\rho_1 = 0.7$ and $\rho_2 = 521/1232$, we get the correlation between the two parallel service times to be 0.5. In Figure 3, we show the numerical results with $\rho = 0$ ("ind.") and $\rho = 0.5$ ("corr."). To calculate the simulated values, we simulated the system up to time 20 with 500 independent replications starting with an empty system. We make two remarks from numerical results. First, the fluid approximations match very well with the simulated results. Second, the positive correlation among parallel service times does not affect $\bar{X}_k$, but reduces $\bar{Y}_k$, for $k = 1, 2$. The maximum relative errors of the simulated values and the corresponding numerical solutions of the fluid models over the time interval $[0, 20]$ are less than 3% in Figures 2 and 3.

4. FCLT in the Halfin-Whitt regime. In this section, we study the fork-join network with NES in the Halfin-Whitt regime, which requires that each service station operates in a critically loaded regime asymptotically. Specifically, we assume the following. Let λ^n be the arrival rate of jobs such that $\bar{\lambda}^n := \lambda^n/n \to \lambda > 0$ as $n \to \infty$, and set $N_k^n := nN_k$, where $N_k \in \mathbb{N}$, and $\rho_k^n := \lambda^n/(\mu_k N_k^n)$ for each $k = 1, ..., K$.

ASSUMPTION 4. *For each* $k = 1, ..., K$*,* $\lambda = N_k\mu_k$ *and* $\sqrt{n}(1 - \rho_k^n) \to \beta_k > 0$*, as* $n \to \infty$*.*

The arrival processes $A^n = \{A^n(t) : t \geq 0\}$ satisfy an FCLT.

ASSUMPTION 5. *There exists a stochastic process $\hat{A}$ with continuous sample paths satisfying*

$$\hat{A}^n(t) := \frac{A^n(t) - \lambda^n t}{\sqrt{n}} \Rightarrow \hat{A}(t) \quad \textit{in} \quad \mathbb{D} \quad \textit{as} \quad n \to \infty. \tag{4.1}$$

It follows from (4.1) that we have the associated FLLN:

$$\bar{A}^n(t) \Rightarrow \lambda t \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty.$$

We now describe the non-empty initial conditions. Due to the complexity caused by initial conditions, we focus on the case of $K = 2$, but our approach can be extended to $K > 2$. For convenience, we use the notation k' to denote its counterpart, i.e., $k' = 1$ ($k' = 2$, respectively) if $k = 2$ ($k = 1$, respectively), for $k = 1, 2$. At time $0-$, there are $X_k^n(0)$ tasks at service station k, and $Y_k^n(0)$ tasks in its associated waiting buffer for synchronization, for $k = 1, 2$. Let $\boldsymbol{X}^n(0) := (X_1^n(0), X_2^n(0))$ and $\boldsymbol{Y}^n(0) := (Y_1^n(0), Y_2^n(0))$. Recall the flow balance equation (2.5). At time $0-$,

$$X_k^n(0) + Y_k^n(0) = X_{k'}^n(0) + Y_{k'}^n(0), \quad k = 1, 2,$$

which is equal to the number of jobs in the system. Note that $X_k^n(0) \geq Y_{k'}^n(0)$ for each $k = 1, 2$, since tasks in the waiting buffer associated with station k' for synchronization must be in station k, either in service or in queue. Let $B_k^n(0) := \min(X_k^n(0), N_k^n)$ and $Q_k^n(0) := (X_k^n(0) - N_k^n)^+$ be the number of tasks in service (busy servers) and the queue length at station k at time $0-$, respectively, $k = 1, 2$. We also assume that $Y_{k'}^n(0) \leq B_k^n(0)$ for $k = 1, 2$. This is not a restrictive assumption, because in the Halfin-Whitt regime, waiting times for service at each station are $O(1/\sqrt{n})$ and service times are $O(1)$, and jobs that have completed tasks in one station and joined its waiting buffer for synchronization have their associated tasks receiving service in the other station with probability one asymptotically.

Let $J^n(0) := \min_{k=1,2}\{B_k^n(0) - Y_{k'}^n(0)\}$ be the number of jobs whose both tasks are in service at time $0-$. Then $Z_k^n(0) := B_k^n(0) - Y_{k'}^n(0) - J^n(0)$ represents the number of jobs in the system at time $0-$ whose task k is in service but whose task k' is in queue waiting for service, $k = 1, 2$. Let $I^n(0) := Q_1^n(0) \wedge Q_2^n(0)$ be the number of jobs (if any) whose both tasks are in queue at their service stations at time $0-$. Then $R_k^n(0) := Q_k^n(0) - I^n(0)$ represents the number of jobs (if any) whose task k is waiting in queue for service while whose task k' is in service, $k = 1, 2$. (Note that our assumption above implies that if a job is waiting in queue at station k, its parallel task can be either in queue or in service at station k'.) By our definition, we

can see that $Z_k^n(0) = R_{k'}^n(0)$, $k = 1, 2$. Set $\boldsymbol{R}^n(0) := (R_1^n(0), R_2^n(0))$ and $\boldsymbol{Z}^n(0) := (Z_1^n(0), Z_2^n(0))$. We also obtain a decomposition for $X_k^n(0)$:

$$X_k^n(0) = B_k^n(0) + Q_k^n(0) = Y_{k'}^n(0) + J^n(0) + Z_k^n(0) + I^n(0) + R_k^n(0) \tag{4.2}$$

for $k = 1, 2$.

We let $\{\tilde{w}_k^{n,i} : i = 1, ..., Q_k^n(0)\}$ be the sequence of remaining waiting times of the tasks in station k at time $0-$, $k = 1, 2$. It is in the order of their positions in queue: $\tilde{w}_k^{n,1}$ is the remaining waiting time of the task in the front of the queue while $\tilde{w}_k^{n,Q_k^n(0)}$ is that for the task in the end of the queue at station k at time $0-$, $k = 1, 2$. Let $\{\tilde{\eta}_k^i : i = 1, ..., B_k^n(0)\}$ be the sequence of remaining service times of the tasks in station k at time $0-$, for $k = 1, 2$. Let $\{\eta_k^{i,Q} : i = 1, ..., Q_k^n(0)\}$ be the sequence of service times of the tasks in station k that are in queue at time $0-$, $k = 1, 2$. Without abuse of notation, we use $\{\tilde{\eta}_k^{i,Y_k} : i = 1, ..., Y_{k'}^n(0)\}$, $\{\tilde{\eta}_k^{i,J} : i = 1, ..., J^n(0)\}$ and $\{\tilde{\eta}_k^{i,Z} : i = 1, ..., Z_k^n(0)\}$, which are partitioning subsets of $\{\tilde{\eta}_k^i : i = 1, ..., B_k^n(0)\}$, to represent the remaining service times of the tasks in station k at time $0-$ corresponding to the quantities $Y_{k'}^n(0)$, $J^n(0)$ and $Z_k^n(0)$, respectively, $k = 1, 2$. Similarly, we use $\{\tilde{w}_k^{n,i,I} : i = 1, ..., I^n(0)\}$ and $\{\tilde{w}_k^{n,i,R} : i = 1, ..., R_k^n(0)\}$, which are partitioning subsets of $\{\tilde{w}_k^{n,i} : i = 1, ..., Q_k^n(0)\}$, to represent the remaining waiting times of the tasks in station k at time $0-$ corresponding to the quantities $I^n(0)$ and $R_k^n(0)$, respectively, $k = 1, 2$. Finally, we use $\{\eta_k^{i,I} : i = 1, ..., I^n(0)\}$ and $\{\eta_k^{i,R} : i = 1, ..., R_k^n(0)\}$, which are partitioning subsets of $\{\eta_k^{i,Q} : i = 1, ..., Q_k^n(0)\}$, to represent the service times of the tasks in station k corresponding to the quantities $I^n(0)$ and $R_k^n(0)$ in queue at time $0-$, respectively, $k = 1, 2$. We assume that these initial quantities are independent of the arrival process A^n and the service times of new arrivals after time 0.

We can now give a representation for the processes $\boldsymbol{X}^n$, $\boldsymbol{Y}^n$ and S^n: for $t \geq 0$ and $k = 1, 2$,

$$\begin{aligned} X_k^n(t) &= \sum_{i=1}^{B_k^n(0)} \mathbf{1}(\tilde{\eta}_k^i > t) + \sum_{i=1}^{Q_k^n(0)} \mathbf{1}(\tilde{w}_k^{n,i} + \eta_k^{i,Q} > t) \\ &\quad + \sum_{i=1}^{A^n(t)} \mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i > t), \end{aligned} \tag{4.3}$$

$$S^n(t) = \sum_{i=1}^{Y_2^n(0)} \mathbf{1}(\tilde{\eta}_1^{i,Y_1} \leq t) + \sum_{i=1}^{Y_1^n(0)} \mathbf{1}(\tilde{\eta}_2^{i,Y_2} \leq t) + \sum_{i=1}^{J^n(0)} \mathbf{1}(\tilde{\eta}_j^{i,J} \leq t, \forall j) \tag{4.4}$$

$$+\sum_{i=1}^{Z_1^n(0)}\mathbf{1}(\tilde{\eta}_1^{i,Z}\le t,\tilde{w}_2^{n,i,R}+\eta_2^{i,R}\le t)+\sum_{i=1}^{Z_2^n(0)}\mathbf{1}(\tilde{w}_1^{n,i,R}+\eta_1^{i,R}\le t,\tilde{\eta}_2^{i,Z}\le t)$$
$$+\sum_{i=1}^{I^n(0)}\mathbf{1}(\tilde{w}_j^{n,i,I}+\eta_j^{i,I}\le t,\forall j)+\sum_{i=1}^{A^n(t)}\mathbf{1}(\tau_i^n+w_j^{n,i}+\eta_j^i\le t,\forall j),$$

and

$$Y_k^n(t) = Y_k^n(0)+X_k^n(0)+A^n(t)-X_k^n(t)-S^n(t). \tag{4.5}$$

We use the convention that $\sum_{i=1}^0\equiv 0$ throughout the paper.

We impose the following assumptions on the initial quantities.

ASSUMPTION 6. *There exists* $(\bar{Y}_1(0),\bar{Y}_2(0))\in\mathbb{R}_+^2$ *such that*

$$(\bar{\boldsymbol{X}}^n(0),\bar{\boldsymbol{Y}}^n(0)):=n^{-1}(\boldsymbol{X}^n(0),\boldsymbol{Y}^n(0))\Rightarrow(\bar{\boldsymbol{X}}(0),\bar{\boldsymbol{Y}}(0))$$

in $\mathbb{R}^4$ *as* $n\to\infty$, *where* $\bar{\boldsymbol{X}}(0):=(N_1,N_2)$ *and* $\bar{\boldsymbol{Y}}(0):=(\bar{Y}_1(0),\bar{Y}_2(0))$. *There exist random vectors* $\hat{\boldsymbol{X}}(0):=(\hat{X}_1(0),\hat{X}_2(0))\in\mathbb{R}^2$ *and* $\hat{\boldsymbol{Y}}(0):=(\hat{Y}_1(0),\hat{Y}_2(0))\in\mathbb{R}^2$ *such that*

$$(\hat{\boldsymbol{X}}^n(0),\hat{\boldsymbol{Y}}^n(0)):=\sqrt{n}(\bar{\boldsymbol{X}}^n(0)-\bar{\boldsymbol{X}}(0),\bar{\boldsymbol{Y}}^n(0)-\bar{\boldsymbol{Y}}(0))\Rightarrow(\hat{\boldsymbol{X}}(0),\hat{\boldsymbol{Y}}(0))$$

in $\mathbb{R}^4$ *as* $n\to\infty$.

This assumption implies that the associated fluid-scaled initial quantities

$$(\bar{J}^n(0),\bar{\boldsymbol{Z}}^n(0),\bar{I}^n(0),\bar{\boldsymbol{R}}^n(0)):=n^{-1}(J^n(0),\boldsymbol{Z}^n(0),I^n(0),\boldsymbol{R}^n(0))\Rightarrow(\bar{J}(0),\bar{\boldsymbol{Z}}(0),\bar{I}(0),\bar{\boldsymbol{R}}(0))$$

in $\mathbb{R}^6$ as $n\to\infty$, where

$$\bar{J}(0):=N_1-\bar{Y}_2(0)=N_2-\bar{Y}_1(0),\quad \bar{\boldsymbol{Z}}(0):=(\bar{Z}_1(0),\bar{Z}_2(0)):=(0,0),$$
$$\bar{I}(0):=0,\quad \bar{\boldsymbol{R}}(0):=(0,0).$$

Define the associated diffusion-scaled quantities $(\hat{J}^n(0),\hat{\boldsymbol{Z}}^n(0),\hat{I}^n(0),\hat{\boldsymbol{R}}^n(0))$ by

$$\hat{J}^n(0):=\frac{J^n(0)-n\bar{J}(0)}{\sqrt{n}},\quad \hat{Z}_k^n(0):=\frac{Z_k^n(0)}{\sqrt{n}},$$
$$\hat{I}^n(0):=\frac{I^n(0)}{\sqrt{n}},\quad \hat{R}_k^n(0):=\frac{R_k^n(0)}{\sqrt{n}},\quad k=1,2.$$

Then Assumption 6 implies that

$$\left(\hat{J}^n(0), \hat{\boldsymbol{Z}}^n(0), \hat{I}^n(0), \hat{\boldsymbol{R}}^n(0)\right) \Rightarrow \left(\hat{J}(0), \hat{\boldsymbol{Z}}(0), \hat{I}(0), \hat{\boldsymbol{R}}(0)\right)$$

in $\mathbb{R}^6$ as $n \to \infty$, where

$$\hat{J}(0) := \min_{k=1,2}\{-(\hat{X}_k(0))^- - \hat{Y}_{k'}(0)\}, \quad \hat{I}(0) := \min_{k=1,2}(\hat{X}_k(0))^+,$$
$$\hat{Z}_k(0) := -(\hat{X}_k(0))^- - \hat{Y}_{k'}(0) - \hat{J}(0), \quad \hat{R}_k(0) := (\hat{X}_k(0))^+ - \hat{I}(0), \quad k = 1, 2.$$

Let

$$F_{k,e}(t) := \frac{1}{E[\eta_k^1]} \int_0^t F_k^c(s)ds, \quad t \geq 0,$$

be the equilibrium distribution associated with F_k, $k = 1, 2$.

ASSUMPTION 7. *For $k = 1, 2$, $\{\tilde{\eta}_k^i : i \in \mathbb{N}\}$ is a sequence of i.i.d. random variables with distribution $F_{k,e}$ and for each $i \in \mathbb{N}$, $\tilde{\eta}_1^i$ and $\tilde{\eta}_2^i$ are independent. $\{\eta_k^{i,Q} : i \in \mathbb{N}\}$ is a sequence of i.i.d. random variables with distribution F_k for each $i \in \mathbb{N}$ and $k = 1, 2$. $\{(\eta_1^{i,I}, \eta_2^{i,I}) : i \in \mathbb{N}\}$ is a sequence of i.i.d. random vectors with a joint distribution $F(\cdot, \cdot)$. $\{(\eta_k^{i,R}, \tilde{\eta}_{k'}^{i,Z}) : i \in \mathbb{N}\}$ is a sequence of i.i.d. random vectors with independent components, $k = 1, 2$.*

Note that in Assumption 6, we have assumed that the system starts from stationarity (in the fluid-scale steady state). Here in Assumption 7, the remaining service times of the tasks in service are assumed to have the associated equilibrium distributions, and they are also assumed to be independent for the two tasks. For jobs with both tasks in queue, we assume that their service vectors have a joint distribution $F(\cdot, \cdot)$, as the new arrivals. For jobs with one task in service and the other in queue, we assume that the task in service has the associated equilibrium distribution, the task in queue has the same marginal distributions as the new arrivals, and both tasks are independent.

Finally, we also make an assumption for the residual waiting times $\{\tilde{w}_k^{n,i} : i = 1, ..., Q_k^n(0)\}$, $k = 1, 2$.

ASSUMPTION 8. *The residual waiting times of the tasks in queue $\{\tilde{w}_k^{n,i} : i = 1, ..., Q_k^n(0)\}$, $k = 1, 2$, converge to zero a.s. as $n \to \infty$.*

We define the diffusion-scaled processes $\hat{\boldsymbol{X}}^n := (\hat{X}_1^n, \hat{X}_2^n)$, $\hat{\boldsymbol{Y}}^n := (\hat{Y}_1^n, \hat{Y}_2^n)$ and $\hat{S}^n$ by

(4.6)

$$\hat{X}_k^n(t) := \frac{X_k^n(t) - N_k^n}{\sqrt{n}}, \quad \hat{Y}_k^n(t) := \frac{Y_k^n(t) - \tilde{Y}_k^n(t)}{\sqrt{n}}, \quad \hat{S}^n(t) := \frac{S^n(t) - \tilde{S}^n(t)}{\sqrt{n}},$$

for $k = 1, 2$, and $t \geq 0$, where

$$\tilde{S}^n(t) := n\bar{S}^0(t) + \bar{\lambda}^n \int_0^t \int_0^t ((t - s_1) \wedge (t - s_2))\, dF(s_1, s_2), \tag{4.7}$$

$$\bar{S}^0(t) := \bar{Y}_2(0)F_{1,e}(t) + \bar{Y}_1(0)F_{2,e}(t) + \bar{J}(0)F_{1,e}(t)F_{2,e}(t), \tag{4.8}$$

$$\tilde{Y}_k^n(t) := n\bar{Y}_k(0) + \lambda^n t - \tilde{S}^n(t). \tag{4.9}$$

From the balance equation for Y_k^n in (4.5), we can rewrite $\hat{Y}_k^n$ as

$$\hat{Y}_k^n(t) = \hat{Y}_k^n(0) + \hat{X}_k^n(0) + \hat{A}^n(t) - \hat{X}_k^n(t) - \hat{S}^n(t), \quad t \geq 0, \tag{4.10}$$

for $k = 1, 2$.

Recall $E_k^n(t)$ is defined as the cumulative number of tasks entering service by time $t \geq 0$ at station k, $k = 1, 2$, assuming the system starts empty in §3. Without abuse of notation, in §4 and §6 related to the FCLT, we let $E_k^n(t)$ be the number of *new arrivals* after time 0 whose task k has entered service by time $t \geq 0$ at station k, $k = 1, 2$.

Define the diffusion-scaled processes $(\hat{\boldsymbol{E}}^n, \hat{\boldsymbol{Q}}^n, \hat{\boldsymbol{B}}^n, \hat{\boldsymbol{D}}^n)$, $\hat{\boldsymbol{E}}^n := (\hat{E}_1^n, \hat{E}_2^n)$, $\hat{\boldsymbol{Q}}^n := (\hat{Q}_1^n, \hat{Q}_2^n)$, $\hat{\boldsymbol{B}}^n := (\hat{B}_1^n, \hat{B}_2^n)$ and $\hat{\boldsymbol{D}}^n := (\hat{D}_1^n, \hat{D}_2^n)$, by

$$\begin{aligned} \hat{E}_k^n(t) &:= \frac{E_k^n(t) - \lambda^n t}{\sqrt{n}}, \quad \hat{Q}_k^n(t) := (\hat{X}_k^n(t))^+, \\ \hat{B}_k^n(t) &:= -(\hat{X}_k^n(t))^-, \quad \hat{D}_k^n(t) := \hat{X}_k^n(0) + \hat{A}^n(t) - \hat{X}_k^n(t), \end{aligned} \tag{4.11}$$

for $k = 1, 2$ and $k \geq 0$. For $s_1, s_2 \geq 0$, let

$$\begin{aligned} \hat{\mathcal{E}}^n(s_1, s_2) &:= \frac{1}{\sqrt{n}} \left((E_1^n(s_1) \wedge E_2^n(s_2)) - \lambda^n(s_1 \wedge s_2) \right) \\ &= \left(\hat{E}_1^n(s_1) + (\lambda^n/\sqrt{n})(s_1 - s_1 \wedge s_2) \right) \\ &\qquad \wedge \left(\hat{E}_2^n(s_2) + (\lambda^n/\sqrt{n})(s_2 - s_1 \wedge s_2) \right). \end{aligned} \tag{4.12}$$

Before we present the FCLT for the fork-join network with NES in the Halfin-Whitt regime, we provide some preliminaries for the limit processes. The limit processes will be functionals of a generalized multiparameter Kiefer process, as a limit of the multiparameter sequential empirical process driven by the service time vectors of new arrivals. Define the multiparameter sequential empirical processes $\hat{K}^n := \{\hat{K}^n(t_1, t_2, \boldsymbol{x}) : t_1 \geq 0, t_2 \geq 0, \boldsymbol{x} \in \mathbb{R}_+^2\}$ by

$$\hat{K}^n(t_1, t_2, \boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt_1 \rfloor \wedge \lfloor nt_2 \rfloor} (\mathbf{1}(\boldsymbol{\eta}^i \leq \boldsymbol{x}) - F(\boldsymbol{x})). \tag{4.13}$$

We prove the convergence of $\hat{K}^n$ in the space $\mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R}))$ endowed with a generalized Skorohod J_1 topology defined in [19] in Proposition 4.1. This proposition generalizes Lemma 3.1 of [29] to the multiparameter setting and Theorem 3.1 in [33], and its proof is provided in §6.1.

The processes $\hat{K}^n$ and their limit $\hat{K}$ are much relevant to the vast literature in empirical processes and Gaussian random fields (see, e.g., [55] and [1]). It is worth noting that the time domain (t_1, t_2) of the processes $\hat{K}^n$ and $\hat{K}$ are two-dimensional, unlike the standard sequential empirical processes studied in the literature. This unique feature arises from the fork-join network model, in order to provide representations for the system dynamics and characterize the limit processes (see (4.24)–(4.25) in Theorem 4.1 and (5.6)). Sequential empirical processes have played an important role in studying many-server queueing models, as first observed by Krichagina and Puhalskii [29], and further developed in [51, 50, 38, 43, 45, 46, 47]. It is also worth noting that in these papers, the weak convergence of the associated sequential empirical processes in the space $\mathbb{D}([0,\infty), \mathbb{D})$ with the Skorohod J_1 topology is required as first observed in [29]. For the fork-join networks with NES, the weak convergence in the space $\mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R}))$ in a generalized Skorohod J_1 topology is required in the proofs of the FCLT.

PROPOSITION 4.1. *Under Assumption 1,*

$$\hat{K}^n(t_1, t_2, \boldsymbol{x}) \Rightarrow \hat{K}(t_1, t_2, \boldsymbol{x}) \tag{4.14}$$

in $\mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R}))$ *as* $n \to \infty$, *where* $\hat{K}(t_1, t_2, \boldsymbol{x})$ *is a continuous Gaussian random field, called a generalized multiparameter Kiefer process, with mean* $E[\hat{K}(t_1, t_2, \boldsymbol{x})] = 0$ *and covariance function*

$$\begin{aligned} &Cov(\hat{K}(s_1, s_2, \boldsymbol{x}), \hat{K}(t_1, t_2, \boldsymbol{y})) \\ &\quad = (s_1 \wedge s_2 \wedge t_1 \wedge t_2)(F(\boldsymbol{x} \wedge \boldsymbol{y}) - F(\boldsymbol{x})F(\boldsymbol{y})), \end{aligned} \tag{4.15}$$

for $s_k, t_k \geq 0$, $k = 1, 2$, *and* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2_+$.

We define the processes $\hat{W}_k := \{\hat{W}_k(t) : t \geq 0\}$, $\hat{W}^c_k := \{\hat{W}^c_k(t) : t \geq 0\}$ and $\hat{W} := \{\hat{W}(t) : t \geq 0\}$ as integral functionals of $\hat{K}$: for $t \geq 0$, $k = 1, 2$,

$$\hat{W}_k(t) := \int_0^t \int_0^t \int_{\mathbb{R}^2_+} \mathbf{1}(s_k + x_k \leq t) d\hat{K}(\lambda s_1, \lambda s_2, \boldsymbol{x}), \tag{4.16}$$

$$\hat{W}(t) := \int_0^t \int_0^t \int_{\mathbb{R}^2_+} \mathbf{1}(s_j + x_j \leq t, \forall j) d\hat{K}(\lambda s_1, \lambda s_2, \boldsymbol{x}), \tag{4.17}$$

and

$$\begin{aligned}\hat{W}_k^c(t) &:= \hat{W}_k(t) - \hat{W}(t)\\ &= \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_k + x_k \le t, s_{k'} + x_{k'} > t) d\hat{K}(\lambda s_1, \lambda s_2, \boldsymbol{x}),\end{aligned} \tag{4.18}$$

where the integrals are defined in the sense of mean-square limits (see the precise definition in §6.2).

PROPOSITION 4.2. *The processes $\hat{W}_k$, $\hat{W}_k^c$ and $\hat{W}$ are well-defined continuous Gaussian processes with mean zero, and for $0 \le s < t$ and $k = 1, 2$,*

$$E[(\hat{W}_k(t) - \hat{W}_k(s))^2] = \lambda \int_0^t (F_k(t-u) - F_k(s-u))(1 - F_k(t-u) + F_k(s-u)) du,$$

$$\begin{aligned}E[(\hat{W}(t) - \hat{W}(s))^2] &= \lambda \int_0^t \int_0^t [\Delta F((s - s_1, s - s_2); (t - s_1, t - s_2))]\\ &\quad \times [1 - \Delta F((s - s_1, s - s_2); (t - s_1, t - s_2))] d(s_1 \wedge s_2),\end{aligned} \tag{4.19}$$

$$\begin{aligned}E[(\hat{W}_k^c(t) - \hat{W}_k^c(s))^2] &= E[(\hat{W}_k(t) - \hat{W}_k(s))^2] + E[(\hat{W}(t) - \hat{W}(s))^2]\\ &\quad - 2\lambda \int_0^t \int_0^t [F(t - s_1, t - s_2) - F_{k,k'}(s - s_k, t - s_{k'})\\ &\quad + (F_k(t - s_k) - F_k(s - s_k))\\ &\qquad \times (F(s - s_1, s - s_2) - F(t - s_1, t - s_2))] d(s_1 \wedge s_2),\end{aligned}$$

and covariance functions

$$\begin{aligned}&Cov(\hat{W}_k(t), \hat{W}_{k'}(t))\\ &= \lambda \int_0^t \int_0^t [F(t - s_1, t - s_2) - F_k(t - s_k) F_{k'}(t - s_{k'})] d(s_1 \wedge s_2),\end{aligned}$$

$$\begin{aligned}&Cov(\hat{W}_k(t), \hat{W}_{k'}^c(t))\\ &= \lambda \int_0^t \int_0^t [F_k(t - s_k) F(t - s_1, t - s_2) - F_k(t - s_k) F_{k'}(t - s_{k'})] d(s_1 \wedge s_2),\end{aligned}$$

$$\begin{aligned}&Cov(\hat{W}_k(t), \hat{W}(t))\\ &= \lambda \int_0^t \int_0^t [F(t - s_1, t - s_2) - F_k(t - s_k) F(t - s_1, t - s_2)] d(s_1 \wedge s_2),\end{aligned}$$

$$Cov(\hat{W}_k^c(t), \hat{W}(t))$$
$$= \lambda \int_0^t \int_0^t [(F(t-s_1, t-s_2))^2 - F_k(t-s_k)F(t-s_1, t-s_2)]d(s_1 \wedge s_2),$$

where $F_{k,k'}(x,y) := P(\eta_k^i \le x, \eta_{k'}^i \le y)$ *for* $x, y \in \mathbb{R}_+$, *and*

$$\Delta F(\boldsymbol{x}; \boldsymbol{y}) := F(y_1, y_2) - F(x_1, y_2) - F(y_1, x_2) + F(x_1, x_2),$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^2$, $\boldsymbol{x} \le \boldsymbol{y}$.

In addition, let $\hat{U} := \{\hat{U}(\boldsymbol{t}) : \boldsymbol{t} \in \mathbb{R}_+^2\}$ be a continuous two-parameter Gaussian process with mean zero and covariance function:

$$\begin{aligned} &Cov(\hat{U}(\boldsymbol{s}), \hat{U}(\boldsymbol{t})) \\ &\quad = (F_{1,e}(s_1 \wedge t_1)F_{2,e}(s_2 \wedge t_2) - F_{1,e}(s_1)F_{2,e}(s_2)F_{1,e}(t_1)F_{2,e}(t_2)), \end{aligned} \tag{4.20}$$

for $\boldsymbol{s} := (s_1, s_2) \in \mathbb{R}_+^2$ and $\boldsymbol{t} := (t_1, t_2) \in \mathbb{R}_+^2$. Define $\hat{U}_k := \{\hat{U}_k(t) : t \ge 0\}$, for $k = 1, 2$, by

$$\hat{U}_1(t) := \hat{U}(t, \infty), \quad \hat{U}_2(t) := \hat{U}(\infty, t), \quad t \ge 0, \tag{4.21}$$

and without abuse of notation, we denote $\hat{U}(t) = \hat{U}(t,t)$, $t \ge 0$. Note that the processes $\hat{W}_k$, $\hat{W}_k^c$ and $\hat{W}$ are independent with $\hat{U}$, as well as $\hat{U}_k$, $k = 1, 2$.

We are now ready to state the FCLT.

THEOREM 4.1. *Under Assumptions 1 and 4-8,*

$$\left(\hat{A}^n, \hat{\boldsymbol{X}}^n, \hat{\boldsymbol{Y}}^n, \hat{S}^n, \hat{\boldsymbol{E}}^n, \hat{\boldsymbol{Q}}^n, \hat{\boldsymbol{B}}^n, \hat{\boldsymbol{D}}^n\right) \Rightarrow \left(\hat{A}, \hat{\boldsymbol{X}}, \hat{\boldsymbol{Y}}, \hat{S}, \hat{\boldsymbol{E}}, \hat{\boldsymbol{Q}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{D}}\right) \tag{4.22}$$

in $\mathbb{D}^{14}$ *as* $n \to \infty$, *where* $\hat{A}$ *is in (4.1),* $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{Y}}$ *and* $\hat{S}$ *are the unique solutions to the following set of stochastic integral equations: for* $t \ge 0$ *and* $k = 1, 2$,

$$\begin{aligned} \hat{X}_k(t) = \hat{X}_k^0(t) &- N_k\beta_k F_{k,e}(t) - \bar{J}(0)^{1/2}\hat{U}_k(t) \\ &- \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t)) - \hat{W}_k(t) \\ &+ \int_0^t (\hat{X}_k(t-s))^+ dF_k(s) + \int_0^t F_k^c(t-s)d\hat{A}(s), \end{aligned} \tag{4.23}$$

$$\begin{aligned} \hat{Y}_k(t) = \hat{Y}_k^0(t) &+ N_k\beta_k F_{k,e}(t) - \bar{Y}_k(0)^{1/2}\hat{B}_{0,k'}(F_{k',e}(t)) \\ &+ \bar{J}(0)^{1/2}(\hat{U}_k(t) - \hat{U}(t)) + \hat{W}_k^c(t) - \hat{\Psi}(t) \\ &- \int_0^t (\hat{X}_k(t-s))^+ dF_k(s) + \int_0^t F_k(t-s)d\hat{A}(s), \end{aligned} \tag{4.24}$$

$$\hat{S}(t) = \hat{S}^0(t) + \bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)) + \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)) + \bar{J}(0)^{1/2}\hat{U}(t) + \hat{W}(t) + \hat{\Psi}(t), \tag{4.25}$$

and $\hat{\boldsymbol{E}}^n$, $\hat{\boldsymbol{Q}}^n$, $\hat{\boldsymbol{B}}^n$ *and* $\hat{\boldsymbol{D}}^n$ *are given as follows:*

$$\hat{E}_k(t) = \hat{A}(t) - (\hat{X}_k(t))^+, \quad \hat{D}_k(t) = \hat{X}_k(0) + \hat{A}(t) - \hat{X}_k(t), \tag{4.26}$$
$$\hat{Q}_k(t) = (\hat{X}_k(t))^+, \quad \hat{B}_k(t) = -(\hat{X}_k(t))^-,$$

where

$$\hat{X}_k^0(t) := \hat{X}_k(0)F_{k,e}^c(t) + (\hat{X}_k(0))^+(F_k^c(t) - F_{k,e}^c(t)), \tag{4.27}$$

$$\hat{S}^0(t) := \sum_{k=1}^{2}(\hat{Y}_{k'}(0)F_{k,e}(t) + \hat{Z}_{k'}(0)F_k(t)F_{k',e}(t)) + \hat{J}(0)F_{1,e}(t)F_{2,e}(t) + \hat{I}(0)F_m(t), \tag{4.28}$$

$$\hat{Y}_k^0(t) := \hat{Y}_k(0) + \hat{X}_k(0)F_{k,e}(t) + (\hat{X}_k(0))^+(F_k(t) - F_{k,e}(t)) - \hat{S}^0(t), \tag{4.29}$$

the processes $\hat{B}_{0,k} := \{\hat{B}_{0,k}(t) : t \geq 0\}$, $k = 1, 2$, *are independent standard Brownian bridges, the process* $\hat{U}$ *is a continuous two-parameter Gaussian process defined above with the processes* $\hat{U}_1$ *and* $\hat{U}_2$ *defined in* (4.21), *and the processes* $\hat{W}_k$, $\hat{W}_k^c$ *and* $\hat{W}$ *are defined in* (4.16), (4.18) *and* (4.17), *and* $\hat{B}_{0,k}$ *is independent of* $\hat{U}$ *and* $\hat{W}_k$, $\hat{W}_k^c$ *and* $\hat{W}$, *and the process* $\hat{\Psi} := \{\hat{\Psi}(t) : t \geq 0\}$ *defined by*

$$\hat{\Psi}(t) := \int_0^t \int_0^t \hat{\mathcal{E}}(t - s_1, t - s_2)dF(s_1, s_2), \tag{4.30}$$

is a well-defined continuous process, where, for $s_1, s_2 \geq 0$,

$$\hat{\mathcal{E}}(s_1, s_2) := \hat{E}_1(s_1)\mathbf{1}(s_1 < s_2) + \hat{E}_2(s_2)\mathbf{1}(s_2 < s_1) + (\hat{E}_1(s_1) \wedge \hat{E}_2(s_2))\mathbf{1}(s_1 = s_2). \tag{4.31}$$

It is worth noting that we have generalized the methodology in Reed [51] for $G/GI/N$ queues to non-Markovian multi-server fork-join networks with NES. The limit processes are shown to be of convolution type, driven by Gaussian random fields. We remark that the limit processes $\hat{X}_k$, $k = 1, 2$, have the same structure as the unique solution to an integral convolution equation, as shown in Reed [51], but are also different because they are both driven by the same generalized multiparameter Kiefer process $\hat{K}$ defined in Proposition 4.1. These two limiting processes $\hat{X}_k$, $k = 1, 2$, are correlated because of the correlated service times of the parallel tasks of each job, which

is captured by the process $\hat{K}$, as well as the same arrival limit process $\hat{A}$. In fact, these two processes $\hat{K}$ and $\hat{A}$ as well as the limits associated with the initial quantities are the driving stochastic components of all the limit processes in (4.23)–(4.26).

5. Proof of fluid limit. In this section, we prove Theorem 3.1. For conciseness, we only show the case when $K = 2$. The argument can be easily generalized to the fork-join system with $K > 2$ parallel service stations. We first give a representation for the fluid-scaled processes $\bar{\boldsymbol{X}}^n$, $\bar{\boldsymbol{Y}}^n$ and $\bar{S}^n$. Recall that the systems are assumed to start from empty in Theorem 3.1.

LEMMA 5.1. *The processes* $\boldsymbol{X}^n$ *in* (3.2), $\boldsymbol{Y}^n$ *in* (3.3) *and* S^n *in* (3.4) *can be represented as*

$$X_k^n(t) = M_k^n(t) + \int_0^t F_k^c(t-s)dA^n(s) + \int_0^t (X_k^n(t-s) - N_k^n)^+ dF_k(s), \tag{5.1}$$

$$\begin{aligned} Y_k^n(t) = {} & \int_0^t F_k(t-s)dA^n(s) - \int_0^t (X_k^n(t-s) - N_k^n)^+ dF_k(s) \\ & - \int_0^t \int_0^t (E_1^n(t-s_1) \wedge E_2^n(t-s_2))\, dF(s_1, s_2) - M_k^n(t) - V^n(t), \end{aligned} \tag{5.2}$$

$$S^n(t) = V^n(t) + \int_0^t \int_0^t (E_1^n(t-s_1) \wedge E_2^n(t-s_2))\, dF(s_1, s_2), \tag{5.3}$$

for $k = 1, 2$, *and* $t \geq 0$, *where*

$$M_k^n(t) := \sum_{i=1}^{A^n(t)} \big(\mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i > t) - F_k^c(t - \tau_i^n - w_k^{n,i})\big), \tag{5.4}$$

$$\begin{aligned} V^n(t) := \sum_{i=1}^{A^n(t)} \big(& \mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i \leq t,\ k = 1, 2) \\ & - F(t - \tau_i^n - w_1^{n,i}, t - \tau_i^n - w_2^{n,i})\big). \end{aligned} \tag{5.5}$$

PROOF. The representation for X_k^n in (5.1) follows from Proposition 2.1 in [51]. We first prove (5.3) holds. By (3.4) and (5.5), we have

$$S^n(t) = V^n(t) + \sum_{i=1}^{A^n(t)} F(t - \tau_i^n - w_1^{n,i}, t - \tau_i^n - w_2^{n,i}), \quad t \geq 0.$$

Observe that, for $t \geq 0$,

$$\begin{aligned}
&\sum_{i=1}^{A^n(t)} F(t-\tau_i^n - w_1^{n,i}, t-\tau_i^n - w_2^{n,i}) \\
&= \sum_{i=1}^{A^n(t)} \int_0^t \int_0^t \mathbf{1}(s_k \leq t-\tau_i^n - w_k^{n,i},\ k=1,2) dF(s_1,s_2) \\
&= \sum_{i=1}^{A^n(t)} \int_0^t \int_0^t \mathbf{1}(\tau_i^n + w_k^{n,i} \leq t-s_k,\ k=1,2) dF(s_1,s_2) \\
&= \int_0^t \int_0^t \sum_{i=1}^{A^n(t)} \mathbf{1}(\tau_i^n + w_k^{n,i} \leq t-s_k,\ k=1,2) dF(s_1,s_2) \\
&= \int_0^t \int_0^t (E_1^n(t-s_1) \wedge E_2^n(t-s_2))\, dF(s_1,s_2).
\end{aligned}$$

Thus, we have derived (5.3). Finally, (5.2) follows from (2.3), (5.1) and (5.3). □

By Theorem 4.1 in [51], we have the following lemma for the convergence of $(\bar{\boldsymbol{X}}^n, \bar{\boldsymbol{E}}^n)$.

LEMMA 5.2. *Under Assumptions 1-3,*

$$(\bar{\boldsymbol{X}}^n, \bar{\boldsymbol{E}}^n) \Rightarrow (\bar{\boldsymbol{X}}, \bar{\boldsymbol{E}})$$

in $\mathbb{D}^4$ *as* $n \to \infty$, *where* $\bar{\boldsymbol{X}}$ *and* $\bar{\boldsymbol{E}}$ *are the unique solutions to* (3.5) *and* (3.6), *respectively.*

We next prove the convergence of $\bar{S}^n$. We observe that the process V^n in (5.5) can be represented as follows:

$$V^n(t) = n \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \leq t, \forall j) d\bar{K}^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}), \tag{5.6}$$

for $t \geq 0$, where the process $\bar{K}^n := \{\bar{K}^n(t_1,t_2,\boldsymbol{x}) : t_1 \geq 0, t_2 \geq 0, \boldsymbol{x} \in \mathbb{R}_+^2\}$ is defined by

$$\bar{K}^n(t_1,t_2,\boldsymbol{x}) := \frac{1}{\sqrt{n}} \hat{K}^n(t_1,t_2,\boldsymbol{x}), \quad t_1,t_2 \in \mathbb{R}_+,\ \boldsymbol{x} \in \mathbb{R}_+^2,$$

where $\hat{K}^n(t_1,t_2,\boldsymbol{x})$ is defined in (4.13). The integral in (5.6) is well-defined as a Stieltjes integral. The following lemma follows directly from Proposition 4.1.

LEMMA 5.3. *Under Assumption 1,*

$$\bar{K}^n \Rightarrow 0 \quad in \quad \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R})) \quad as \quad n \to \infty.$$

Note that the processes $\bar{K}^n(t_1, t_2, \boldsymbol{x})$ have the following decomposition: for $t_1 \geq 0$, $t_2 \geq 0$ and $\boldsymbol{x} \in \mathbb{R}^2_+$,

$$\bar{K}^n(t_1, t_2, \boldsymbol{x}) = \bar{K}^n_1(t_1, t_2, \boldsymbol{x}) + \bar{K}^n_2(t_1, t_2, \boldsymbol{x}),$$

where

(5.7)
$$\bar{K}^n_1(t_1, t_2, \boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{\lfloor nt_1 \rfloor \wedge \lfloor nt_2 \rfloor} \left(\mathbf{1}(\boldsymbol{\eta}^i \leq \boldsymbol{x}) - \int_0^{x_1} \int_0^{x_2} \frac{\mathbf{1}(\boldsymbol{\eta}^i > \boldsymbol{u})}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \right),$$

$$\bar{K}^n_2(t_1, t_2, \boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{\lfloor nt_1 \rfloor \wedge \lfloor nt_2 \rfloor} \left(\int_0^{x_1} \int_0^{x_2} \frac{\mathbf{1}(\boldsymbol{\eta}^i > \boldsymbol{u}) - F^c(\boldsymbol{u})}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \right). \tag{5.8}$$

We then decompose V^n into two processes, $G^n := \{G^n(t) : t \geq 0\}$ and $H^n := \{H^n(t) : t \geq 0\}$ as follows:

$$V^n(t) = H^n(t) + G^n(t), \quad t \geq 0, \tag{5.9}$$

where

$$H^n(t) := n \int_0^t \int_0^t \int_{\mathbb{R}^2_+} \mathbf{1}(s_j + x_j \leq t, \forall j) d\bar{K}^n_1(\bar{E}^n_1(s_1), \bar{E}^n_2(s_2), \boldsymbol{x}), \tag{5.10}$$

$$G^n(t) := n \int_0^t \int_0^t \int_{\mathbb{R}^2_+} \mathbf{1}(s_j + x_j \leq t, \forall j) d\bar{K}^n_2(\bar{E}^n_1(s_1), \bar{E}^n_2(s_2), \boldsymbol{x}). \tag{5.11}$$

Define the fluid-scaled processes $\bar{H}^n := n^{-1}H^n$ and $\bar{G}^n := n^{-1}G^n$. We next show the convergence of the processes $\bar{H}^n$ and $\bar{G}^n$.

LEMMA 5.4. *Under Assumptions 1-3,*

$$(\bar{H}^n, \bar{G}^n) \Rightarrow (0, 0) \quad in \quad \mathbb{D}^2 \quad as \quad n \to \infty.$$

We first prove the convergence $\bar{H}^n \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$ in Lemma 5.4. Let $\hat{\tau}^{n,i}_j$ be the time at which task j of job i enters service after time 0, i.e.,

$$\hat{\tau}^{n,i}_j = \inf\{t \geq 0 : E^n_j(t) \geq i\}, \quad j = 1, 2.$$

We denote $\tilde{F}$ to be the distribution function of $\max_j(\hat{\tau}_j^{n,i}+\eta_j^i)$ conditional on $\hat{\tau}_1^{n,i}$ and $\hat{\tau}_2^{n,i}$, i.e., $\tilde{F}(t) := F(t-\hat{\tau}_1^{n,i}, t-\hat{\tau}_2^{n,i})$ for $t \geq 0$. Note that $\tilde{F}$ depends on n and i and we omit n and i below for conciseness. Let $\tilde{F}^c := 1-\tilde{F}$. Define, for $t \geq 0$,

$$H^{n,i}(t) := \mathbf{1}(\eta_j^i \leq t-\hat{\tau}_j^{n,i}, \forall j) - \int_0^{\eta_1^i \wedge (t-\hat{\tau}_1^{n,i})^+} \int_0^{\eta_2^i \wedge (t-\hat{\tau}_2^{n,i})^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}), \tag{5.12}$$

$$\tilde{H}^{n,i}(t) := \mathbf{1}(\eta_j^i \leq t-\hat{\tau}_j^{n,i}, \forall j) - \int_0^t \frac{\mathbf{1}\big(\max_j(\hat{\tau}_j^{n,i}+\eta_j^i) > u\big)}{\tilde{F}^c(u)} d\tilde{F}(u), \tag{5.13}$$

and, for each $\kappa \geq 1$,

$$H_\kappa^n(t) := \sum_{i=1}^{E_1^n(t)\wedge E_2^n(t)\wedge\kappa} H^{n,i}(t). \tag{5.14}$$

Denote $H^{n,i} := \{H^{n,i}(t) : t \geq 0\}$, $\tilde{H}^{n,i} := \{\tilde{H}^{n,i}(t) : t \geq 0\}$ and $H_\kappa^n := \{H_\kappa^n(t) : t \geq 0\}$. Let $\{\xi_i^n := \tau_i^n - \tau_{i-1}^n,\ i \geq 1\}$ be the interarrival times between the $(i-1)^{\text{th}}$ and i^{th} jobs arriving to the system. Define the filtration $\mathcal{H}^n := \{\mathcal{H}_t^n : t \geq 0\}$ by

$$\begin{aligned}\mathcal{H}_t^n := \sigma\Big(\mathbf{1}(\eta_j^i \leq s-\hat{\tau}_j^{n,i}, \forall j), s \leq t, i = 1, ..., E_1^n(t) \wedge E_2^n(t)\Big) \\ \vee \sigma\left(E_j^n(s), s \leq t, \forall j\right) \vee \sigma\left(\xi_i^n, i \geq 1\right) \vee \mathcal{N},\end{aligned} \tag{5.15}$$

where $\mathcal{N}$ includes all the null sets. It is easy to verify that $\mathcal{H}^n$ is actually a filtration and satisfies the usual conditions [23]. We first state the martingale property of $H^{n,i}$ and $\tilde{H}^{n,i}$ in Lemma 5.5, whose proof can be found in §7.

LEMMA 5.5. *Under Assumptions 1-3, the processes $H^{n,i}$ and $\tilde{H}^{n,i}$ are $\mathcal{H}^n$-martingales.*

Next, we will show the process $\tilde{H}^{n,i}$ is an martingale with respect to the filtration $\mathcal{H}^n$.

LEMMA 5.6. *Under Assumptions 1-3, for each $\kappa \geq 1$, the process H_κ^n is an $\mathcal{H}^n$-square-integrable martingale with predictable quadratic variation process*

$$\langle H_\kappa^n \rangle(t) = \sum_{i=1}^{E_1^n(t)\wedge E_2^n(t)\wedge\kappa} \int_0^{\eta_1^i \wedge (t-\hat{\tau}_1^{n,i})^+} \int_0^{\eta_2^i \wedge (t-\hat{\tau}_2^{n,i})^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}), \quad t \geq 0.$$

PROOF OF LEMMA 5.6. By the definition of H^n_κ in (5.14) and Lemma 5.5, H^n_κ is $\mathcal{H}^n$-adapted and an $\mathcal{H}^n$-martingale. Note that, for each $t \geq 0$,

$$|H^{n,i}(t)| \leq 1 + \int_0^{\eta_1^i} \int_0^{\eta_2^i} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}), \quad a.s.$$

By Lemma 4.3 in [33], we have $E\left[|H^{n,i}(t)|^2\right] < \infty$, for $t \geq 0$.

It is easy to check that the second terms (without the minus) on the RHS of (5.12) and (5.13) are predictable with respect to the filtration $\mathcal{H}^n$, and thus are compensators for the point process $\{\mathbf{1}(\max_{j=1,2}(\hat{\tau}_j^{n,i} + \eta_j^i) \leq t) : t \geq 0\}$. By the uniqueness of Doob-Mayer decomposition (see, e.g., Theorem 4.10 in [23]) in the sense of indistinguishability, $H^{n,i}$ and $\tilde{H}^{n,i}$ are indistinguishable and we can write

$$H^n_\kappa(t) = \sum_{i=1}^{E_1^n(t) \wedge E_2^n(t) \wedge \kappa} \tilde{H}^{n,i}(t), \quad t \geq 0.$$

Thus, it suffices to prove the following two claims:

(i) The predictable quadratic variation process of $\tilde{H}^{n,i}$ is given by

$$\langle \tilde{H}^{n,i} \rangle(t) = \int_0^t \frac{\mathbf{1}\big(\max_j(\hat{\tau}_j^{n,i} + \eta_j^i) > u\big)}{\tilde{F}^c(u)} d\tilde{F}(u).$$

(ii) The martingales $H^{n,i}$ and $H^{n,j}$ are orthogonal for $i \neq j$, i.e., the product $H^{n,i}H^{n,j}$ is an $\mathcal{H}^n$-martingale, or equivalently, the predictable quadratic covariation $\langle H^{n,i}, H^{n,j} \rangle(t) = 0$ for $t \geq 0$ (see Proposition 4.15 of Chapter I in Jacod and Shiryaev [20]).

The proof of claim (i) follows a similar argument as part 2 of Lemma A.1 in [51]. We provide the details here for completeness. Since the second term on the RHS of (5.13) is $\mathcal{H}^n$-predictable, by Proposition 1 of Chapter 3.4 in Liptser and Shiryaev [32], the $\mathcal{H}^n$-predictable measure of the jumps of the process $\{\mathbf{1}(\hat{\tau}_j^{n,i} + \eta_j^i \leq t, \forall j) : t \geq 0\}$ is

$$\nu^{n,i}([0,t], C) = \{1 \in C\} \int_0^t \mathbf{1}\big(0 < u \leq \max_j(\hat{\tau}_j^{n,i} + \eta_j^i)\big) \frac{d\tilde{F}(u)}{\tilde{F}^c(u)}, \quad t \geq 0,$$

where C is a Borel set in $\mathbb{R}$, and thus, the predictable quadratic-variation process of $\tilde{H}^{n,i}$ is (see, e.g., Problem 11 of Chapter 4.1 in Liptser and Shiryaev [32])

$$\langle \tilde{H}^{n,i} \rangle(t) = \int_0^t \int_{\mathbb{R}} x^2 \nu^{n,i}(du, dx) - \sum_{0<u\leq t} \left(\int_{\mathbb{R}} x \nu^{n,i}(\{u\}, dx) \right)^2$$

$$= \int_0^t \mathbf{1}\big(0 < u \le \max_j(\hat{\tau}_j^{n,i} + \eta_j^i)\big) \frac{d\tilde{F}(u)}{\tilde{F}^c(u)}$$
$$- \sum_{0<u\le t} \mathbf{1}\big(0 < u \le \max_j(\hat{\tau}_j^{n,i} + \eta_j^i)\big) \left(\frac{\Delta\tilde{F}(u)}{\tilde{F}^c(u)}\right)^2$$
$$= \int_0^t \frac{\mathbf{1}\big(\max_j(\hat{\tau}_j^{n,i} + \eta_j^i) > u\big)}{\tilde{F}^c(u)} d\tilde{F}(u), \quad t \ge 0.$$

This completes the proof of claim (i).

We now focus on the proof of claim (ii), i.e., the martingale property for $H^{n,i}H^{n,j}$. It is sufficient to show, for $s < t$, $j < i$,

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right] = 0, \tag{5.16}$$

and

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} \le s) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right] = H^{n,i}(s)H^{n,j}(s). \tag{5.17}$$

We first prove (5.16). Note that $\hat{\tau}_k^{n,i}$ is an $\mathcal{H}^n$-stopping time since $\sigma(E_k^n(s), s \le t) \subset \mathcal{H}_t^n$ for each $t \ge 0$, $k = 1, 2$. This implies $\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}$ is also a stopping time with respect to $\mathcal{H}^n$, and $\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}$ is well-defined. We then have

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n]$$
$$= \mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E\left[E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right]|\mathcal{H}_s^n\right].$$

Note that

$$E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right] \tag{5.18}$$
$$= \mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right]$$
$$+ \mathbf{1}(\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right]$$
$$+ \mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right]$$
$$+ \mathbf{1}(\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i}\vee\hat{\tau}_2^{n,i}}\right].$$

Since

$$\mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j})$$

and

$$\begin{aligned}&\mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t) \\ &\quad = \mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t \wedge (\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}))\end{aligned}$$

are $\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}$-measurable, it follows that

$$\begin{aligned}&\mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\quad = E\left[\mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\quad = \mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t) E\left[H^{n,i}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right],\end{aligned}$$

and thus, the martingale property of $H^{n,i}$ and Doob's stopping theorem (see, e.g., Theorem 1.39 in Chapter I of [23]) imply that

$$E\left[H^{n,i}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] = E\left[H^{n,i}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i})\right] = 0.$$

Thus, the first term on the RHS of (5.18) is equal to 0. We now consider the second term on the RHS of (5.18). It can be decomposed into two terms as follows:

$$\begin{aligned}&\mathbf{1}(\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\quad = \mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\qquad - \mathbf{1}(\eta_1^j \le \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right].\end{aligned}$$

We only need to show the first term on the RHS of the above is equal to 0. Since $\mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j})$ and $\mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t) = H^{n,j}(t \wedge \hat{\tau}_2^{n,i})$ are both $\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}$-measurable,

$$\begin{aligned}&\mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\quad = E\left[\mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\ &\quad = \mathbf{1}(\eta_2^j \le \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t) E\left[H^{n,i}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right].\end{aligned}$$

Since $H^{n,i}$ is an $\mathcal{H}^n$-martingale, by Doob's stopping theorem (see, e.g., Theorem 1.39 in Chapter I of [23]),

$$E\left[H^{n,i}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] = E\left[H^{n,i}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i})\right] = 0.$$

Thus, we obtain that the second term on the RHS of (5.18) is equal to 0. The proof that the third term on the RHS of (5.18) is 0 is analogous to that for the second term, and is omitted.

We now consider the last term on the RHS of (5.18). On the event $\{\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}\}$, we have that task k of job j has finished service after task k of job i arrives, and so the service time vector of job j has no effect on $\hat{\tau}_k^{n,i}$, the time at which task k of job i enters service, for $k = 1, 2$. Thus, $\boldsymbol{\eta}^j$ and $\hat{\tau}_k^{n,i}$ are independent on the event $\{\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}\}$, for $k = 1, 2$. More precisely, there exist random variables $\check{\tau}_1^{n,i}$ and $\check{\tau}_2^{n,i}$, which are Borel functions of ξ_r^n, $r \geq 1$, $\boldsymbol{\eta}^p$, $p \geq 1, p \neq i, p \neq j$, such that $\{\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}\} = \{\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}\}$ and $\{\hat{\tau}_k^{n,i} = \check{\tau}_k^{n,i}, k = 1, 2\}$ on either event. Thus, applying Lemma 3.6 in [29] and using the fact that $\boldsymbol{\eta}^i$ and $\boldsymbol{\eta}^j$ are independent of $\check{\tau}_k^{n,i}$ and $\hat{\tau}_k^{n,j}$, $k = 1, 2$, we have

$$\begin{aligned}
&\mathbf{1}(\eta_1^j > \hat{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] \\
&= \mathbf{1}(\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) \\
&\qquad \times \frac{E\left[\mathbf{1}(\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) \check{H}^{n,i}(t) H^{n,j}(t) | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\right]}{P(\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j} | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k)},
\end{aligned}$$

where $\check{H}^{n,i}$ denotes $H^{n,i}$ with $\check{\tau}_k^{n,i}$ substituted for $\hat{\tau}_k^{n,i}$ for $k = 1, 2$. Furthermore, since $\boldsymbol{\eta}^i$ is independent of $\check{\tau}_k^{n,i}$, $\boldsymbol{\eta}^j$ and $\hat{\tau}_k^{n,j}$, $k = 1, 2$, we have

$$\begin{aligned}
E\left[\mathbf{1}(\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) \check{H}^{n,i}(t) H^{n,j}(t) | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\right] \\
= E\big[\mathbf{1}(\eta_1^j > \check{\tau}_1^{n,i} - \hat{\tau}_1^{n,j}, \eta_2^j > \check{\tau}_2^{n,i} - \hat{\tau}_2^{n,j}) H^{n,j}(t) | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\big] \\
\times E\big[\check{H}^{n,i}(t) | \check{\tau}_k^{n,i}, \forall k\big].
\end{aligned} \tag{5.19}$$

By the definition of $\check{H}^{n,i}$ and the fact that $\boldsymbol{\eta}^i$ is independent of $\check{\tau}_k^{n,i}$, $k = 1, 2$, we note that $E[\check{H}^{n,i}(t) | \check{\tau}_k^{n,i}, \forall k] = 0$, which implies that the RHS of (5.19) is 0, and thus (5.16) holds.

We will now focus on the proof of (5.17). The proof proceeds similarly to that in Lemma 5.5. Let $\mathcal{R}_k^l$ be either $\leq$ or $>$, the relationship between two real numbers, for $k = 1, 2$, and $l = i, j$. We then have a decomposition for (5.17) by

$$\begin{aligned}
&\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} \leq s) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}_s^n\right] \\
&\quad = \sum_{\mathcal{R}_1^i, \mathcal{R}_2^i, \mathcal{R}_1^j, \mathcal{R}_2^j} \Big(\mathbf{1}(\eta_k^i \mathcal{R}_k^i (s - \hat{\tau}_k^{n,i}), s - \hat{\tau}_k^{n,i} \geq 0, \forall k)
\end{aligned}$$

$$\times \mathbf{1}(\eta_k^j \mathcal{R}_k^j(s-\hat{\tau}_k^{n,j}), \forall k) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\Big),$$

where the summation $\sum_{\mathcal{R}_1^i,\mathcal{R}_2^i,\mathcal{R}_1^j,\mathcal{R}_2^j}$ denotes the sum of all the cases for the relationships $\mathcal{R}_k^l$, for $l = i, j$ and $k = 1, 2$. In order to prove (5.17), it is enough to check for each $\mathcal{R}_k^i$ and $\mathcal{R}_k^j$, $k = 1, 2$,

$$\begin{aligned}
&\mathbf{1}(\eta_k^i \mathcal{R}_k^i(s-\hat{\tau}_k^{n,i}), s-\hat{\tau}_k^{n,i} \geq 0, \forall k)\\
&\quad \times \mathbf{1}(\eta_k^j \mathcal{R}_k^j(s-\hat{\tau}_k^{n,j}), \forall k) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&= \mathbf{1}(\eta_k^i \mathcal{R}_k^i(s-\hat{\tau}_k^{n,i}), s-\hat{\tau}_k^{n,i} \geq 0, \forall k)\mathbf{1}(\eta_k^j \mathcal{R}_k^j(s-\hat{\tau}_k^{n,j}), \forall k) H^{n,i}(s)H^{n,j}(s).
\end{aligned}$$

Here we only focus on proving the following two equations:

$$\begin{aligned}
(5.20)\quad &\mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \geq 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \geq 0, \eta_2^j \leq s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&= \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_1^{n,i} \geq 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \geq 0, \eta_2^j \leq s-\hat{\tau}_2^{n,j}) H^{n,i}(s)H^{n,j}(s),
\end{aligned}$$

and

$$\begin{aligned}
(5.21)\quad &\mathbf{1}(\eta_1^i > s-\hat{\tau}_1^{n,i} \geq 0, \eta_2^i > s-\hat{\tau}_2^{n,i} \geq 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s-\hat{\tau}_2^{n,j} \geq 0) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&= \mathbf{1}(\eta_1^i > s-\hat{\tau}_1^{n,i} \geq 0, \eta_2^i > s-\hat{\tau}_2^{n,i} \geq 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s-\hat{\tau}_2^{n,j} \geq 0) H^{n,i}(s)H^{n,j}(s),
\end{aligned}$$

and the proof of the other cases can be carried out similarly.

For (5.20), we first observe that

$$\begin{aligned}
&\mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \geq 0)\mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \geq 0, \eta_2^j \leq s-\hat{\tau}_2^{n,j})\\
&\quad = \left[\mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}) - \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i \leq s-\hat{\tau}_2^{n,i})\right]\\
&\qquad \times \left[\mathbf{1}(\eta_2^j \leq s-\hat{\tau}_2^{n,j}) - \mathbf{1}(\eta_1^j \leq s-\hat{\tau}_1^{n,j}, \eta_2^j \leq s-\hat{\tau}_2^{n,j})\right]\\
&\quad = \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_2^j \leq s-\hat{\tau}_2^{n,j})\\
&\qquad - \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i \leq s-\hat{\tau}_2^{n,i})\mathbf{1}(\eta_2^j \leq s-\hat{\tau}_2^{n,j})\\
&\qquad - \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_1^j \leq s-\hat{\tau}_1^{n,j}, \eta_2^j \leq s-\hat{\tau}_2^{n,j})\\
&\qquad + \mathbf{1}(\eta_1^i \leq s-\hat{\tau}_1^{n,i}, \eta_2^i \leq s-\hat{\tau}_2^{n,i})\mathbf{1}(\eta_1^j \leq s-\hat{\tau}_1^{n,j}, \eta_2^j \leq s-\hat{\tau}_2^{n,j}).
\end{aligned}$$

Thus, we obtain

(5.22)
$$\begin{aligned}
&\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \ge 0, \eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&= \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad - \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i \le s-\hat{\tau}_2^{n,i})\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad - \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_1^j \le s-\hat{\tau}_1^{n,j}, \eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad + \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i \le s-\hat{\tau}_2^{n,i})\\
&\qquad\qquad \times \mathbf{1}(\eta_1^j \le s-\hat{\tau}_1^{n,j}, \eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right].
\end{aligned}$$

Since $\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})$, $\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})$, $\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})H^{n,i}(t) = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})H^{n,i}(s)$ and $\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})H^{n,j}(t) = \mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})H^{n,j}(s)$ are $\mathcal{H}_s^n$-measurable, we then have

$$\begin{aligned}
&\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j}) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})\\
&\qquad \times E\left[\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})H^{n,i}(t)\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})\mathbf{1}(\eta_2^j \le s-\hat{\tau}_2^{n,j})H^{n,i}(s)H^{n,j}(s).
\end{aligned}$$

Similarly, we can obtain the corresponding results for the other terms on the RHS of (5.22), which completes the proof of (5.20).

Next, we focus on the proof of (5.21). Since

$$\begin{aligned}
&\{s-\hat{\tau}_1^{n,i} \ge 0, s-\hat{\tau}_2^{n,i} \ge 0, \eta_1^j > s-\hat{\tau}_1^{n,j}, \eta_2^j > s-\hat{\tau}_2^{n,j}\}\\
&\qquad\qquad \subset \{\eta_1^j > \hat{\tau}_1^{n,i}-\hat{\tau}_1^{n,j}, \eta_2^j > \hat{\tau}_2^{n,i}-\hat{\tau}_2^{n,j}\},
\end{aligned}$$

it follows as above that

$$\begin{aligned}
&\{s-\hat{\tau}_1^{n,i} \ge 0, s-\hat{\tau}_2^{n,i} \ge 0, \eta_1^j > s-\hat{\tau}_1^{n,j}, \eta_2^j > s-\hat{\tau}_2^{n,j}\}\\
&\quad = \{s-\check{\tau}_1^{n,i} \ge 0, s-\check{\tau}_2^{n,i} \ge 0, \eta_1^j > s-\hat{\tau}_1^{n,j}, \eta_2^j > s-\hat{\tau}_2^{n,j}\}
\end{aligned}$$

and $\{\hat{\tau}_k^{n,i} = \check{\tau}_k^{n,i}, \forall k\}$ on either event. Hence,

(5.23)
$$\begin{aligned}
&\mathbf{1}(\eta_1^i > s-\hat{\tau}_1^{n,i} \ge 0, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0)\\
&\qquad \times \mathbf{1}(\eta_1^j > s-\hat{\tau}_1^{n,j} \ge 0, \eta_2^j > s-\hat{\tau}_2^{n,j} \ge 0) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right]\\
&\quad = \mathbf{1}(\eta_1^i > s-\check{\tau}_1^{n,i} \ge 0, \eta_2^i > s-\check{\tau}_2^{n,i} \ge 0)
\end{aligned}$$

$$\times \mathbf{1}(\eta_1^j > s - \hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s - \hat{\tau}_2^{n,j} \geq 0) E\left[\check{H}^{n,i}(t) H^{n,j}(t) | \mathcal{H}_s^n\right],$$

where $\boldsymbol{\eta}^i$ and $\boldsymbol{\eta}^j$ are independent of $\check{\tau}_k^{n,i}$ and $\hat{\tau}_k^{n,j}$ for $k = 1, 2$. Analogous to the proof of Lemma 5.5, we have

$$\begin{aligned}
&\mathcal{H}_s^n \cap \{\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k\} \cap \{\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k\} \\
&\quad \subset (\sigma(\xi_r^n, r \geq 1, \boldsymbol{\eta}^p, p \geq 1, p \neq i, p \neq j) \vee \sigma(\check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k) \vee \mathcal{N}) \\
&\qquad \cap \{\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k\} \cap \{\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k\},
\end{aligned}$$

where $\mathcal{N}$ includes all null sets. Applying (5.23) and Lemma 3.6 in [29], we have

$$\begin{aligned}
(5.24)\quad &\mathbf{1}(\eta_1^i > s - \hat{\tau}_1^{n,i} \geq 0, \eta_2^i > s - \hat{\tau}_2^{n,i} \geq 0) \\
&\quad \times \mathbf{1}(\eta_1^j > s - \hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s - \hat{\tau}_2^{n,j} \geq 0) E\left[H^{n,i}(t) H^{n,j}(t) | \mathcal{H}_s^n\right] \\
&= \mathbf{1}(\eta_1^i > s - \check{\tau}_1^{n,i} \geq 0, \eta_2^i > s - \check{\tau}_2^{n,i} \geq 0) \\
&\quad \times \mathbf{1}(\eta_1^j > s - \hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s - \hat{\tau}_2^{n,j} \geq 0) \\
&\quad \times E\Big[\mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k) \mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) \\
&\qquad\qquad \times \check{H}^{n,i}(t) H^{n,j}(t) \big| \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\Big] \\
&\quad \div P\left(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\right).
\end{aligned}$$

Recall that $\boldsymbol{\eta}^i$ and $\boldsymbol{\eta}^j$ are independent of $\check{\tau}_k^{n,i}$ and $\hat{\tau}_k^{n,j}$ for $k = 1, 2$. We then obtain

$$\begin{aligned}
&E\Big[\mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k) \mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) \\
&\qquad\qquad \times \check{H}^{n,i}(t) H^{n,j}(t) | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\Big] \\
&= E\big[\mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k) \check{H}^{n,i}(t) | \check{\tau}_k^{n,i}, \forall k\big] \\
&\qquad \times E\big[\mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) H^{n,j}(t) | \hat{\tau}_k^{n,j}, \forall k\big],
\end{aligned}$$

and

$$\begin{aligned}
(5.25)\quad &P\left(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k | \check{\tau}_k^{n,i}, \hat{\tau}_k^{n,j}, \forall k\right) \\
&= P\left(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k | \check{\tau}_k^{n,i}, \forall k\right) \\
&\qquad \times P\left(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k | \hat{\tau}_k^{n,j}, \forall k\right).
\end{aligned}$$

By Lemma 3.6 of [29] and an analogous argument in the proof of Lemma 5.5, we have

$$\mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k) \frac{E\left[\mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k)\check{H}^{n,i}(t)|\check{\tau}_k^{n,i}, \forall k\right]}{P\left(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k|\check{\tau}_k^{n,i}, \forall k\right)}$$
$$= \mathbf{1}(\eta_k^i > s - \check{\tau}_k^{n,i} \geq 0, \forall k) E\left[\check{H}^{n,i}(t)|\mathcal{H}_s^n\right],$$

and

$$\mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) \frac{E\left[\mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k)H^{n,j}(t)|\hat{\tau}_k^{n,j}, \forall k\right]}{P\left(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k|\hat{\tau}_k^{n,j}, \forall k\right)}$$
$$= \mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) E[H^{n,j}(t)|\mathcal{H}_s^n].$$

Combining (5.24)-(5.25), and noting the fact that $\check{\tau}_k^{n,i} = \hat{\tau}_k^{n,i}$, $k = 1, 2$, on the event

$$\{s - \hat{\tau}_k^{n,i} \geq 0, \eta_k^j > s - \hat{\tau}_k^{n,j}, \forall k\} = \{s - \check{\tau}_k^{n,i} \geq 0, \eta_k^j > s - \hat{\tau}_k^{n,j}, \forall k\},$$

as well as the martingale property of $H^{n,i}$ and $H^{n,j}$, we obtain

$$\begin{aligned}
&\mathbf{1}(\eta_1^i > s - \hat{\tau}_1^{n,i} \geq 0, \eta_2^i > s - \hat{\tau}_2^{n,i} \geq 0) \\
&\quad \times \mathbf{1}(\eta_1^j > s - \hat{\tau}_1^{n,j} \geq 0, \eta_2^j > s - \hat{\tau}_2^{n,j} \geq 0) E\left[H^{n,i}(t)H^{n,j}(t)|\mathcal{H}_s^n\right] \\
&= \mathbf{1}(\eta_k^i > s - \hat{\tau}_k^{n,i} \geq 0, \forall k) E[H^{n,i}(t)|\mathcal{H}_s^n] \\
&\quad \times \mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) E[H^{n,j}(t)|\mathcal{H}_s^n] \\
&= \mathbf{1}(\eta_k^i > s - \hat{\tau}_k^{n,i} \geq 0, \forall k)\mathbf{1}(\eta_k^j > s - \hat{\tau}_k^{n,j} \geq 0, \forall k) H^{n,i}(s)H^{n,j}(s),
\end{aligned}$$

which completes the proof of (5.21). Thus, we have shown Lemma 5.6 holds. □

Proof of the convergence $\bar{H}^n \Rightarrow 0$ in Lemma 5.4. Fix $T > 0$. For each $\epsilon > 0$, by Lemma 5.6, we have that for each $\kappa \in \mathbb{N}$,

$$P\left(\sup_{0\leq t\leq T} |\bar{H}^n(t)| > \epsilon\right) \leq P\left(\bar{E}_1^n(T) \wedge \bar{E}_2^n(T) > \kappa\right) + P\left(\sup_{0\leq t\leq T} \left|\bar{H}_{\kappa n}^n(t)\right| > \epsilon\right).$$

Lemma 5.2 implies that the processes $(\bar{E}_1^n, \bar{E}_2^n)$ are stochastically bounded, and thus, for κ sufficiently large, the first term on the RHS of the inequality above goes to 0 as $n \to \infty$. We only need to show the second term converges

to 0 as $n \to \infty$. By the Lenglart-Rebolledo inequality [32], it follows that for any $\gamma > 0$,

$$P\left(\sup_{0\le t\le T} \left|\bar{H}^n_{\kappa n}(t)\right| > \epsilon \right) \le \frac{\gamma}{\epsilon^2} + P\left(\langle \bar{H}^n_{\kappa n}\rangle(T) > \gamma\right).$$

Recall from Lemma 5.6 that

$$\langle \bar{H}^n_{\kappa n}\rangle(T) = \frac{1}{n^2} \sum_{i=1}^{E^n_1(T)\wedge E^n_2(T)\wedge(\kappa n)} \int_0^{\eta^i_1\wedge(T-\hat{\tau}^{n,i}_1)^+} \int_0^{\eta^i_2\wedge(T-\hat{\tau}^{n,i}_2)^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}).$$

Hence,

$$\langle \bar{H}^n_{\kappa n}\rangle(T) \le \frac{1}{n^2} \sum_{i=1}^{E^n_1(T)\wedge E^n_2(T)} \int_0^{\eta^i_1} \int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}).$$

Note that, by Fubini's theorem,

$$E\left[\int_0^{\eta^i_1} \int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \right] = 1. \tag{5.26}$$

It follows by the FLLN that

$$\frac{1}{n^2} \sum_{i=1}^{\lfloor n\cdot\rfloor} \int_0^{\eta^i_1} \int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty.$$

Thus, by Lemma 5.2, the continuous mapping theorem [7] and the random time change theorem [7], we have

$$P\left(\langle \bar{H}^n_{\kappa n}\rangle(T) > \gamma\right) \to 0 \quad \text{as} \quad n \to \infty,$$

which completes the proof. □

Next, we will prove the convergence $\bar{G}^n \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$ in Lemma 5.4. We follow a similar argument in Lemma A.3 in [51] to prove the convergence of $\bar{G}^n$, but generalize that to the multiparameter setting.

We introduce a multiparameter process $\tilde{T}^n := \{\tilde{T}^n(t_1, t_2, \boldsymbol{x}) : t_1 \ge 0, t_2 \ge 0, \boldsymbol{x} \in \mathbb{R}^2_+\}$ defined by

$$\tilde{T}^n(t_1, t_2, \boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{E^n_1(t_1)\wedge E^n_2(t_2)} (\mathbf{1}(\boldsymbol{\eta}^i \ge \boldsymbol{x}) - F^c(\boldsymbol{x})), \tag{5.27}$$

for $t_1, t_2 \ge 0$, $\boldsymbol{x} \in \mathbb{R}^2_+$. Following a similar argument as in Lemma 5.3, we obtain the following lemma.

LEMMA 5.7. *Under Assumptions 1-3,*

$$\tilde{T}^n \Rightarrow 0 \quad in \quad \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R})) \quad as \quad n \to \infty.$$

We also define the mapping $\phi : \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R})) \to \mathbb{D}$ by

$$\phi(u)(t) := \int_0^t \int_0^t \frac{u(t-x_1, t-x_2, \boldsymbol{x})\mathbf{1}(F^c(\boldsymbol{x}) \geq \epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}),$$

for some $\epsilon \in (0,1)$ and for $t \geq 0$ and $u \in \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R}))$. The next lemma shows the continuity property of this mapping.

LEMMA 5.8. *Suppose $u_n, u \in \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R}))$ and u is continuous. If $u_n \to u$ as $n \to \infty$, then $\phi(u_n) \to \phi(u)$ as $n \to \infty$.*

PROOF. Since u is continuous, by the definition of ϕ, $\phi(u)(\cdot)$ is also continuous in space $\mathbb{D}$. To show the continuity of the mapping ϕ, it is sufficient to show that for $T > 0$,

$$\sup_{0 \leq t \leq T} |\phi(u_n)(t) - \phi(u)(t)| \to 0 \quad \text{as} \quad n \to \infty.$$

Denote the set $\mathcal{A} := \{\boldsymbol{x} \in \mathbb{R}_+^2 : F^c(\boldsymbol{x}) \geq \epsilon\}$ and let $C_{\mathcal{A}} := \int_{\mathcal{A}} dF(\boldsymbol{x}) > 0$ be a positive constant. Note that

$$\begin{aligned}
&\sup_{0 \leq t \leq T} |\phi(u_n)(t) - \phi(u)(t)| \\
&= \sup_{0 \leq t \leq T} \left| \int_0^t \int_0^t \tfrac{(u_n(t-x_1,t-x_2,\boldsymbol{x}) - u(t-x_1,t-x_2,\boldsymbol{x}))\mathbf{1}(F^c(\boldsymbol{x}) \geq \epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}) \right| \\
&\leq \sup_{\substack{0 \leq t_1, t_2 \leq T \\ \boldsymbol{x} \in \mathcal{A}}} |u_n(t_1,t_2,\boldsymbol{x}) - u(t_1,t_2,\boldsymbol{x})| \left| \int_0^T \int_0^T \frac{\mathbf{1}(F^c(\boldsymbol{x}) \geq \epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}) \right| \\
&\leq \sup_{\substack{0 \leq t_1, t_2 \leq T \\ \boldsymbol{x} \in \mathcal{A}}} |u_n(t_1,t_2,\boldsymbol{x}) - u(t_1,t_2,\boldsymbol{x})| \frac{C_{\mathcal{A}}}{\epsilon}.
\end{aligned}$$

The convergence of u_n and the continuity of u imply that the RHS of the above inequality goes to 0 as $n \to \infty$. Therefore, the continuity of the mapping ϕ follows. □

PROOF OF OF THE CONVERGENCE $\bar{G}^n \Rightarrow 0$ IN LEMMA 5.4. We first rewrite $\bar{G}^n$ in (5.11) as

$$\bar{G}^n(t) = \int_0^t \int_0^t \frac{\tilde{T}^n(t-x_1, t-x_2, \boldsymbol{x})}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}), \quad t \geq 0.$$

Fix $\epsilon \in (0,1)$. We decompose $\bar{G}^n$ as follows: for $t \geq 0$,

$$\bar{G}^n(t) = \bar{G}_1^{n,\epsilon}(t) + \bar{G}_2^{n,\epsilon}(t),$$

where

$$\bar{G}_1^{n,\epsilon}(t) := \int_0^t \int_0^t \frac{\tilde{T}^n(t-x_1, t-x_2, \boldsymbol{x})\mathbf{1}(F^c(\boldsymbol{x}) \geq \epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}),$$
$$\bar{G}_2^{n,\epsilon}(t) := \int_0^t \int_0^t \frac{\tilde{T}^n(t-x_1, t-x_2, \boldsymbol{x})\mathbf{1}(F^c(\boldsymbol{x}) < \epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}).$$

Now it suffices to prove the following two claims:

(i) $\bar{G}_1^{n,\epsilon}(t) \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$;
(ii) For each $\delta > 0$ and $T > 0$,

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} P\left(\sup_{0 \leq t \leq T} |\bar{G}_2^{n,\epsilon}(t)| > \delta \right) = 0.$$

By Lemmas 5.7 and 5.8, we can conclude (i) holds. We now focus on proving (ii). Without abuse of notation, we denote $\tilde{T}^n(t, \boldsymbol{x}) := \tilde{T}^n(t,t,\boldsymbol{x})$ for $t \geq 0$ and $\boldsymbol{x} \in \mathbb{R}_+^2$. Recall the definition of $\tilde{T}^n$ in (5.27). We obtain, for any $\kappa > 0$,

$$\begin{aligned}
&P\left(\sup_{0 \leq t \leq T} |\bar{G}_2^{n,\epsilon}(t)| > \delta \right) \\
&\leq P(\bar{E}_i^n(T) > \kappa T, \forall i) \\
&\quad + P\left(\int_0^t \int_0^t \frac{\mathbf{1}(F^c(\boldsymbol{x}) < \epsilon)}{F^c(\boldsymbol{x})} \sup_{0 \leq t_1, t_2 \leq \kappa T} |\tilde{T}^n(t_1, t_2, \boldsymbol{x})| dF(\boldsymbol{x}) > \delta \right) \\
&\leq P(\bar{E}_i^n(T) > \kappa T, \forall i) \\
&\quad + P\left(\int_0^t \int_0^t \frac{\mathbf{1}(F^c(\boldsymbol{x}) < \epsilon)}{F^c(\boldsymbol{x})} \left(\sup_{0 \leq t_1 \leq t_2 \leq \kappa T} |\tilde{T}^n(t_1, t_2, \boldsymbol{x})| \right.\right. \\
&\qquad\qquad \left.\left. + \sup_{0 \leq t_2 \leq t_1 \leq \kappa T} |\tilde{T}^n(t_1, t_2, \boldsymbol{x})| \right) dF(\boldsymbol{x}) > \delta \right) \\
&\leq P(\bar{E}_i^n(T) > \kappa T, \forall i) \\
&\quad + 2P\left(\int_0^t \int_0^t \frac{\mathbf{1}(F^c(\boldsymbol{x}) < \epsilon)}{F^c(\boldsymbol{x})} \sup_{0 \leq t \leq \kappa T} |\tilde{T}^n(t, \boldsymbol{x})| dF(\boldsymbol{x}) > \frac{\delta}{2} \right).
\end{aligned}$$

For κ sufficiently large, by Lemma 5.2, we have the first term on the RHS of the above inequality converges to 0 as $n \to \infty$. For the second term, we

proceed as that in Lemma 6.5 in [33] and can show this term also goes to 0 when $n \to \infty$ and $\epsilon \to 0$, which completes the proof. □

PROOF OF THEOREM 3.1. By Lemma 5.2, and the balance equations in (2.1), (2.2) and (2.4), we obtain the joint convergence of

$$(\bar{A}^n, \bar{\boldsymbol{X}}^n, \bar{\boldsymbol{E}}^n, \bar{\boldsymbol{Q}}^n, \bar{\boldsymbol{B}}^n, \bar{\boldsymbol{D}}^n) \Rightarrow (\bar{a}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{E}}, \bar{\boldsymbol{Q}}, \bar{\boldsymbol{B}}, \bar{\boldsymbol{D}})$$

in $\mathbb{D}^{5K+1}$ as $n \to \infty$, where the limits are given in (3.1), (3.5), (3.6) and (3.9).

Now to show the weak convergence of $\bar{S}^n$, by (5.3), it is sufficient to show

$$\bar{V}^n \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{5.28}$$

and

$$\begin{aligned} &\int_0^t \int_0^t \left(\bar{E}_1^n(t-s_1) \wedge \bar{E}_2^n(t-s_2)\right) dF(s_1,s_2) \\ &\qquad \Rightarrow \int_0^t \int_0^t \left(\bar{E}_1(t-s_1) \wedge \bar{E}_2(t-s_2)\right) dF(s_1,s_2) \end{aligned} \tag{5.29}$$

in $\mathbb{D}$ as $n \to \infty$. By the decomposition of V^n in (5.9), Lemma 5.4 and the continuous mapping theorem [7], we can conclude (5.28) holds.

To prove (5.29), we define a mapping $\psi : \mathbb{D}^2 \to \mathbb{D}$ by

$$\psi(x_1,x_2)(t) := \int_0^t \int_0^t (x_1(t-s_1) \wedge x_2(t-s_2))\, dF(s_1,s_2),$$

for $x_1, x_2 \in \mathbb{D}$ and $t \geq 0$. By the weak convergence of $\bar{\boldsymbol{E}}^n$, it suffices to show the mapping ψ is continuous at all continuity points in $\mathbb{D}^2$ and thus, applying the continuous mapping theorem [7], we can conclude the convergence in (5.29). We now prove the continuity property of the mapping ψ. Suppose $x_1^n, x_2^n \in \mathbb{D}$ satisfy

$$(x_1^n, x_2^n) \to (x_1, x_2) \quad \text{in} \quad \mathbb{D}^2 \quad \text{as} \quad n \to \infty,$$

where x_j is continuous in $\mathbb{D}$ for $j = 1, 2$. Recall that we endow the product metric space with the maximum metric of each component space. Since x_j is continuous, by the definition of ψ, $\psi(x_1,x_2)(\cdot)$ is also continuous in $\mathbb{D}$. To show the continuity of the mapping, it is sufficient to show that for $T > 0$,

$$\sup_{0 \leq t \leq T} |\psi(x_1^n, x_2^n)(t) - \psi(x_1,x_2)(t)| \to 0, \quad \text{as} \quad n \to \infty.$$

Note the fact that for $a, b \in \mathbb{R}$,

$$a \wedge b = \frac{1}{2}(a + b - |a - b|). \tag{5.30}$$

Now, for the fixed $T > 0$,

(5.31)

$$\begin{aligned}
&\sup_{0\le t\le T} |\psi(x_1^n, x_2^n)(t) - \psi(x_1, x_2)(t)| \\
&= \sup_{0\le t\le T} \left| \int_0^t \int_0^t [x_1^n(t - s_1) \wedge x_2^n(t - s_2) - x_1(t - s_1) \wedge x_2(t - s_2)] dF(s_1, s_2) \right| \\
&\le \sup_{0\le t\le T} \int_0^t \int_0^t |x_1^n(t - s_1) \wedge x_2^n(t - s_2) - x_1(t - s_1) \wedge x_2(t - s_2)| dF(s_1, s_2) \\
&\le F_m(T) \sup_{0\le s_1, s_2\le T} |x_1^n(s_1) \wedge x_2^n(s_2) - x_1(s_1) \wedge x_2(s_2)| \\
&= \frac{F_m(T)}{2} \sup_{0\le s_1, s_2\le T} \Big| x_1^n(s_1) + x_2^n(s_2) - |x_1^n(s_1) - x_2^n(s_2)| \\
&\qquad\qquad - (x_1(s_1) + x_2(s_2) - |x_1(s_1) - x_2(s_2)|) \Big| \\
&\le \frac{F_m(T)}{2} \Bigg(\sum_{i=1}^{2} \sup_{0\le s\le T} |x_i^n(s) - x_i(s)| \\
&\qquad\qquad + \sup_{0\le s_1, s_2\le T} \big| |x_1^n(s_1) - x_2^n(s_2)| - |x_1(s_1) - x_2(s_2)| \big| \Bigg).
\end{aligned}$$

Since

$$\sup_{0\le s_1, s_2\le T} \big| |x_1^n(s_1) - x_2^n(s_2)| - |x_1(s_1) - x_2(s_2)| \big| \le \sum_{i=1}^{2} \sup_{0\le s\le T} |x_i^n(s) - x_i(s)|,$$

we further obtain by (5.31),

$$\sup_{0\le t\le T} |\psi(x_1^n, x_2^n)(t) - \psi(x_1, x_2)(t)| \le F_m(T) \sum_{i=1}^{2} \sup_{0\le s\le T} |x_i^n(s) - x_i(s)|.$$

The convergence of x_1^n and x_2^n and the continuity of x_1 and x_2 imply the RHS of the above inequalty converges to 0 as $n \to \infty$, which completes the proof of the continuity of the mapping ψ.

Finally, the proof of the convergence of $\bar{\boldsymbol{Y}}^n$ follows from the balance equation (2.3) and the continuous mapping theorem [7]. The uniqueness of all these processes follows from the uniqueness of $\bar{X}_k$, $k = 1, 2$. □

6. Proof of FCLT. In this section, we prove Theorem 4.1. We start by proving Propositions 4.1 and 4.2 in §§6.1 and 6.2, respectively. We then give some preliminary results in §6.3. We prove the convergence of the processes $\hat{W}^n$, $\hat{W}^n_k$ and $\hat{W}^{n,c}_k$ in §6.4, $k = 1, 2$. The convergence of the initial quantities is proved in §6.5. We complete the proof of Theorem 4.1 in §6.6.

6.1. *Proof of Proposition 4.1.* We first define a multiparameter sequential empirical process $\breve{U}^n := \{\breve{U}^n(t_1, t_2, \boldsymbol{x}) : t_1 \geq 0, t_2 \geq 0, \boldsymbol{x} \in [0,1]^2\}$ by

$$\breve{U}^n(t_1, t_2, \boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt_1 \rfloor \wedge \lfloor nt_2 \rfloor} (\mathbf{1}(\boldsymbol{\xi}^i \leq \boldsymbol{x}) - H(\boldsymbol{x})), \tag{6.1}$$

for $t_1, t_2 \geq 0$ and $\boldsymbol{x} \in [0,1]^2$, where $\{\boldsymbol{\xi}^i := (\xi^i_1, \xi^i_2) : i \in \mathbb{N}\}$ is a sequence of i.i.d. random vectors with joint distribution function $H(\cdot)$ and uniform marginals over $[0,1]$. Define $\boldsymbol{F} : \mathbb{R}^2_+ \to [0,1]^2$ with $\boldsymbol{F}(\boldsymbol{x}) = (F_1(x_1), F_2(x_2))$. By Sklar's theorem [52], for any multivariate distribution function F, there exists a unique multivariate distribution function H (called "copula") with uniform marginals on $[0,1]$ such that $F(\boldsymbol{x}) = H(\boldsymbol{F}(\boldsymbol{x}))$ when the marginal distribution functions F_k, $k = 1, 2$, are continuous. Then, we can write

$$\hat{K}^n(t_1, t_2, \boldsymbol{x}) = \breve{U}^n(t_1, t_2, \boldsymbol{F}(\boldsymbol{x})), \quad t_1, t_2 \in \mathbb{R}_+, \ \boldsymbol{x} \in \mathbb{R}^2_+.$$

To prove Proposition 4.1, it suffices to show that

$$\breve{U}^n(t_1, t_2, \boldsymbol{x}) \Rightarrow \breve{U}(t_1, t_2, \boldsymbol{x}) \quad \text{in} \quad \mathbb{D}([0,\infty)^2, \mathbb{D}_2) \quad \text{as} \quad n \to \infty,$$

where $\breve{U}(t, \boldsymbol{x})$ is a continuous Gaussian random field with mean $E[\breve{U}(t_1, t_2, \boldsymbol{x})] = 0$ and covariance function

$$Cov(\breve{U}(s_1, s_2, \boldsymbol{x}), \breve{U}(t_1, t_2, \boldsymbol{y})) = (s_1 \wedge s_2 \wedge t_1 \wedge t_2)(H(\boldsymbol{x} \wedge \boldsymbol{y}) - H(\boldsymbol{x})H(\boldsymbol{y})).$$

We proceed by proving that the finite-dimensional distributions of $\breve{U}^n$ converge weakly to those of $\breve{U}$, and $\{\breve{U}^n : n \geq 1\}$ is tight. Denote $\breve{U}^n(t_1, t_2) := \breve{U}^n(t_1, t_2, \cdot)$ for $t_1, t_2 \in [0, \infty)$. Without abuse of notation, we let $\breve{U}^n(t) := \breve{U}^n(t, t, \cdot)$ for $t \geq 0$. In order to show the convergence of the finite-dimensional distributions of $\breve{U}^n$, it suffices to prove for any $l \in \mathbb{N}$ and $\boldsymbol{t}^k := (t^k_1, t^k_2)$, where $t^k_1, t^k_2 \in [0, \infty)$ and $k = 1, ..., l$,

$$(\breve{U}^n(t^1_1, t^1_2), ..., \breve{U}^n(t^l_1, t^l_2)) \Rightarrow (\breve{U}(t^1_1, t^1_2), ..., \breve{U}(t^l_1, t^l_2))$$

in $\mathbb{D}^l_2$ as $n \to \infty$. By the definition of $\breve{U}^n$, it is equivalent to prove

$$(\breve{U}^n(t^1_1 \wedge t^1_2), ..., \breve{U}^n(t^l_1 \wedge t^l_2)) \Rightarrow (\breve{U}(t^1_1 \wedge t^1_2), ..., \breve{U}(t^l_1 \wedge t^l_2))$$

in $\mathbb{D}^l_2$ as $n \to \infty$, which follows directly from Theorem 3.1 of [33].

Now, we focus on the tightness of $\breve{U}^n$. From Corollary 4.2 of [19], it is equivalent to show that there exists a sequence $\{\boldsymbol{\alpha}^l := (\alpha_1^l, \alpha_2^l) \in \mathbb{R}_+^2 : l \geq 1\}$ satisfying $\min_k \alpha_k^l \to \infty$ as $l \to \infty$ such that

(i) for each $\boldsymbol{\alpha}^l$ and every $\epsilon > 0$ there exists a compact set $M_{l,\epsilon} \subset \mathbb{D}_2$ such that

$$P(\breve{U}^n(t_1, t_2) \in M_{l,\epsilon}, \forall \boldsymbol{t} \in [0, \alpha_1^l] \times [0, \alpha_2^l]) > 1 - \epsilon, \quad n \geq 1;$$

(ii) for each $l \geq 1$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(\omega_\delta^{\boldsymbol{\alpha}^l}(\breve{U}^n) \geq \epsilon) = 0,$$

where

$$\omega_\delta^{\boldsymbol{\alpha}^l}(\breve{U}^n) := \inf_{\Delta_{\boldsymbol{\alpha}^l}(\delta)} \max_{B \in \Delta_{\boldsymbol{\alpha}^l}(\delta)} \omega_{\breve{U}^n}(B),$$

and $\omega_{\breve{U}^n}(B) := \sup_{\boldsymbol{s},\boldsymbol{t} \in B} d_{\mathbb{D}_2}(\breve{U}^n(s_1, s_2), \breve{U}^n(t_1, t_2))$, and $d_{\mathbb{D}_2}$ is the metric in space $\mathbb{D}_2$.

Recall from Theorem 3.1 of [33] that the processes $\ddot{U}^n := \{\ddot{U}(t, \boldsymbol{x}) := \breve{U}^n(t, t, \boldsymbol{x}) : t \geq 0 \text{ and } \boldsymbol{x} \in [0,1]^2\}$ are tight in $\mathbb{D}([0,\infty), \mathbb{D}_2)$. Let $\dot{U}^n(t) := \ddot{U}^n(t, \cdot)$ for $t \geq 0$. From Corollary 4.1 of [19] we have that there exists a sequence $\{\alpha_0^l \in \mathbb{R}_+ : l \geq 1\}$ satisfying $\alpha_0^l \to \infty$ as $l \to \infty$ such that

(i') for each α_0^l and every $\epsilon > 0$ there exists a compact set $M_{l,\epsilon} \subset \mathbb{D}_2$ such that

$$P\big(\ddot{U}^n(t) \in M_{l,\epsilon}, \forall t \in [0, \alpha_0^l]\big) > 1 - \epsilon, \quad n \geq 1;$$

(ii') for each $l \geq 1$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\big(\omega_\delta^{\alpha_0^l}(\ddot{U}^n) \geq \epsilon\big) = 0,$$

where

$$\omega_\delta^{\alpha_0^l}(\breve{U}^n) := \inf_{\Delta_{\alpha_0^l}(\delta)} \max_{B \in \Delta_{\alpha_0^l}(\delta)} \omega_{\ddot{U}^n}(B),$$

and $\omega_{\ddot{U}^n}(B) := \sup_{s,t \in B} d_{\mathbb{D}_2}(\ddot{U}^n(s), \ddot{U}^n(t))$.

We set $\boldsymbol{\alpha}^l = (\alpha_0^l, \alpha_0^l)$ for $l \geq 0$. By the definition of $\breve{U}^n$ in (6.1), we see that conditions (i') and (ii") imply (i) and (ii) hold, respectively. Therefore, the tightness of $\breve{U}^n$ holds, which completes the proof of Proposition 4.1. □

6.2. *Proof of Proposition 4.2.* Here we only focus on the process $\hat{W}$, and the other processes can be analyzed similarly. We first show $\hat{W}$ is well-defined.

We introduce some notations here. For a set $\mathcal{J}$, let $|\mathcal{J}|$ be the cardinality of $\mathcal{J}$. Let $\mathcal{J}_k^1$ and $\mathcal{J}_{N-k}^2$ be the partition of $\mathcal{A} := \{1, ..., N\}$, where N is a positive integer, $\mathcal{J}_k^1 \cap \mathcal{J}_{N-k}^2 = \varnothing$, $|\mathcal{J}_k^1| = k$ and $|\mathcal{J}_{N-k}^2| = N - k$. Let $\Phi : \mathbb{R}^N \to \mathbb{R}$. For $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^N$, define $\Phi^{\mathcal{J}_k^1, \mathcal{J}_{N-k}^2}(\boldsymbol{x}; \boldsymbol{y}) := \Phi(\boldsymbol{z})$, where $z_j = x_j$ for $j \in \mathcal{J}_k^1$ and $z_j = y_j$ for $j \in \mathcal{J}_{N-k}^2$. Then, we define

$$\Delta\Phi(\boldsymbol{x}; \boldsymbol{y}) := \sum_{k=0}^{N} (-1)^k \sum_{\mathcal{J}_k^1, \mathcal{J}_{N-k}^2 \text{ partitions of } \mathcal{A}} \Phi^{\mathcal{J}_k^1, \mathcal{J}_{N-k}^2}(\boldsymbol{x}; \boldsymbol{y}). \tag{6.2}$$

In the rest of the paper, we will use $\Delta\hat{K}^n$ and $\Delta\hat{K}$ as defined in (6.2) when $N = 4$. Notation ΔF is defined as (6.2) when $N = 2$.

By the definition of mean-square integrals, we have

$$\lim_{l \to \infty} E[(\hat{W}(t) - \hat{W}^{(l)}(t))^2] = 0, \quad t \geq 0,$$

where

$$\hat{W}^{(l)}(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}_t^{(l)}(s_1, s_2, \boldsymbol{x}) d\hat{K}(\lambda s_1, \lambda s_2, \boldsymbol{x}), \tag{6.3}$$

with

$$\mathbf{1}_t^{(l)}(s_1, s_2, \boldsymbol{x}) := \sum_{i=1}^{l} \sum_{j=1}^{l} [\mathbf{1}(s_{i-1}^l < s_1 \leq s_i^l, s_{j-1}^l < s_2 \leq s_j^l) \mathbf{1}(x_j \leq t - s_j^l, \forall j)],$$

and $0 = s_0^l < s_1^l < ... < s_l^l = t$ satisfying $\max_{1 \leq i \leq l} |s_i^l - s_{i-1}^l| \to 0$ as $l \to \infty$. We call $\{s_i^l : 0 \leq i \leq l\}$ is a partition of $[0, t]$. Define $\hat{W}^{(\ell)}(t)$ and its associated partition $\{s_i^\ell : 0 \leq i \leq \ell\}$ of $[0, t]$ similarly, $t \geq 0$. To show $\hat{W}$ is well-defined, it suffices to prove

$$\lim_{l, \ell \to \infty} E[(\hat{W}^{(l)}(t) - \hat{W}^{(\ell)}(t))^2] = 0, \quad t \geq 0. \tag{6.4}$$

Without loss of generality, we assume that the partition $\{s_i^\ell : 0 \leq i \leq \ell\}$ of $[0, t]$ is finer than the partition $\{s_i^l : 0 \leq i \leq l\}$. By (6.3), for $t \geq 0$,

$$\begin{aligned}
&\hat{W}^{(l)}(t) - \hat{W}^{(\ell)}(t) \\
&= \sum_{i=1}^{l} \sum_{j=1}^{l} \sum_{p : s_{i-1}^l < s_p^\ell \leq s_i^l} \sum_{q : s_{j-1}^l < s_q^\ell \leq s_j^l} \Delta\hat{K}\big((\lambda s_{p-1}^\ell, \lambda s_{q-1}^\ell, t - s_i^l, t - s_j^l); \\
&\qquad\qquad (\lambda s_p^\ell, \lambda s_q^\ell, t - s_p^\ell, t - s_q^\ell)\big).
\end{aligned}$$

By the definition of $\hat{K}$, we can easily obtain that for $0 \le s_1 \le t_1$, $0 \le s_2 \le t_2$ and $\mathbf{0} \le \boldsymbol{x} \le \boldsymbol{y}$,

$$E[(\Delta\hat{K}((s_1, s_2, \boldsymbol{x}); (t_1, t_2, \boldsymbol{y})))^2] = [(t_1 - s_1) \wedge (t_2 - s_2)]\Delta F(\boldsymbol{x}; \boldsymbol{y})(1 - \Delta F(\boldsymbol{x}; \boldsymbol{y})), \tag{6.5}$$

and for $0 \le s_1' \le t_1'$, $0 \le s_2' \le t_2'$, $\mathbf{0} \le \boldsymbol{x}' \le \boldsymbol{y}'$, $t_1 \le s_1'$ and $t_2 \le s_2'$,

$$E[\Delta\hat{K}((s_1, s_2, \boldsymbol{x}); (t_1, t_2, \boldsymbol{y}))\Delta\hat{K}((s_1', s_2', \boldsymbol{x}'); (t_1', t_2', \boldsymbol{y}'))] = 0. \tag{6.6}$$

By (6.5) and (6.6), we have, for $t \ge 0$,

$$\begin{aligned}
&E[(\hat{W}^{(l)}(t) - \hat{W}^{(\ell)}(t))^2] \\
&= \sum_{i=1}^{l}\sum_{j=1}^{l} \sum_{p: s_{i-1}^l < s_p^\ell \le s_i^l} \sum_{q: s_{j-1}^l < s_q^\ell \le s_j^l} \lambda[(s_p^\ell - s_{p-1}^\ell) \wedge (s_q^\ell - s_{q-1}^\ell)] \\
&\qquad \times \Delta F((t - s_i^l, t - s_j^l); (t - s_p^\ell, t - s_q^\ell)) \\
&\qquad\quad \times \left(1 - \Delta F((t - s_i^l, t - s_j^l); (t - s_p^\ell, t - s_q^\ell))\right) \\
&\le \sum_{i=1}^{l}\sum_{j=1}^{l} \sum_{p: s_{i-1}^l < s_p^\ell \le s_i^l} \sum_{q: s_{j-1}^l < s_q^\ell \le s_j^l} \lambda[(s_p^\ell - s_{p-1}^\ell) \wedge (s_q^\ell - s_{q-1}^\ell)] \\
&\qquad\qquad \times \Delta F((t - s_i^l, t - s_j^l); (t - s_p^\ell, t - s_q^\ell)) \\
&\le \sum_{i=1}^{l}\sum_{j=1}^{l} \lambda[(s_i^l - s_{i-1}^l) \wedge (s_j^l - s_{j-1}^l)]\Delta F((t - s_i^l, t - s_j^l); (t - s_{i-1}^l, t - s_{j-1}^l)) \\
&\le \max_{1 \le i \le l} \max_{1 \le j \le l} \lambda[(s_i^l - s_{i-1}^l) \wedge (s_j^l - s_{j-1}^l)].
\end{aligned}$$

Since $\max_{1 \le i \le l}(s_i^l - s_{i-1}^l) \to 0$ as $l \to \infty$, we have proved (6.4), which implies that the process $\hat{W}$ is well-defined.

Recall from (4.14) that $\hat{K}$ is Gaussian with mean 0. Then, for a fixed $t \ge 0$, $\hat{W}^{(l)}(t)$ is normally distributed with mean 0. By the definition of $\hat{W}$, $\hat{W}^{(l)}$ converges to $\hat{W}$ in probability as $l \to \infty$. Recall the fact that if a sequence of normally distributed random variables in probability converges to a random variable, the limit is also a normal random variable (see, e.g., Lemma 4.9.4 of [32]). Thus, $\hat{W}(t)$ is normally distributed, $t \ge 0$, which implies the process $\hat{W}$ is Gaussian.

Next, we will show (4.19) holds. By the definition of the process $\hat{W}$, we see

$$E[(\hat{W}(t) - \hat{W}(s))^2] = \lim_{l \to \infty} E[(\hat{W}^{(l)}(t) - \hat{W}^{(l)}(s))^2], \tag{6.7}$$

where we assume the same partition $\{s_i^l : 0 \le i \le l\}$ of $[0,t]$ is applied for $\hat{W}^{(l)}(t)$ and $\hat{W}^{(l)}(s)$ for $0 \le s \le t$. By (6.3), it is easy to see that

$$\begin{aligned}&\hat{W}^{(l)}(t) - \hat{W}^{(l)}(s)\\ &= \sum_{i=1}^{l}\sum_{j=1}^{l} \Delta\hat{K}((s_{i-1}^l, s_{j-1}^l, s - s_i^l, s - s_j^l); (s_i^l, s_j^l, t - s_i^l, t - s_j^l)),\end{aligned}$$

for $t \ge s \ge 0$, where we set $\hat{K}(s_1, s_2, x_1, x_2) = 0$ if $x_1 < 0$ or $x_2 < 0$. Thus, together with (6.5) and (6.6), we obtain

$$\begin{aligned}&E[(\hat{W}^{(l)}(t) - \hat{W}^{(l)}(s))^2]\\ &= \sum_{i=1}^{l}\sum_{j=1}^{l} \lambda[(s_i^l - s_{i-1}^l) \wedge (s_j^l - s_{j-1}^l)]\Delta F((s - s_i^l, s - s_j^l); (t - s_i^l, t - s_j^l))\\ &\qquad\qquad \times (1 - \Delta F((s - s_i^l, s - s_j^l); (t - s_i^l, t - s_j^l))), \quad t \ge s \ge 0.\end{aligned}$$

By Lebesgue's theorem, we have

$$\begin{aligned}&\lim_{l\to\infty} E[(\hat{W}^{(l)}(t) - \hat{W}^{(l)}(s))^2]\\ &= \lambda \int_0^t \int_0^t [\Delta F((s - s_1, s - s_2); (t - s_1, t - s_2))\\ &\qquad \times (1 - \Delta F((s - s_1, s - s_2); (t - s_1, t - s_2)))]\, d(s_1 \wedge s_2), \quad t \ge s \ge 0,\end{aligned}$$

which by (6.7) implies that (4.19) holds.

We now prove the process $\hat{W}$ is continuous. Note that (4.19) indicates $\hat{W}$ is continuous in probability. To show the process $\hat{W}$ has continuous sample paths *a.s.*, by Lemma 4.9.6 of [32], it is sufficent to prove that for any partition $\{s_i^l : 0 \le i \le l\}$ of $[0,t]$,

$$\lim_{L\to\infty} \limsup_{l\to\infty} P\Big(\sum_{i=1}^{l} (\hat{W}(s_i^l) - \hat{W}(s_{i-1}^l))^2 \ge L\Big) = 0. \tag{6.8}$$

By Markov inequality and (4.19), we note that

$$\begin{aligned}&P\Big(\sum_{i=1}^{l} (\hat{W}(s_i^l) - \hat{W}(s_{i-1}^l))^2 \ge L\Big)\\ &\le \frac{1}{L}\sum_{i=1}^{l} E[(\hat{W}(s_i^l) - \hat{W}(s_{i-1}^l))^2]\end{aligned}$$

$$
\begin{aligned}
&= \frac{\lambda}{L}\sum_{i=1}^{l}\int_0^t\int_0^t [\Delta F((s_{i-1}^l - s_1, s_{i-1}^l - s_2);(s_i^l - s_1, s_i^l - s_2)) \\
&\qquad \times (1 - \Delta F((s_{i-1}^l - s_1, s_{i-1}^l - s_2);(s_i^l - s_1, s_i^l - s_2)))]d(s_1 \wedge s_2) \\
&\le \frac{\lambda}{L}\sum_{i=1}^{l}\int_0^t\int_0^t [\Delta F((s_{i-1}^l - s_1, s_{i-1}^l - s_2);(s_i^l - s_1, s_i^l - s_2))]d(s_1 \wedge s_2) \\
&\le \frac{\lambda t}{L}.
\end{aligned}
$$

Therefore, we see (6.8) holds, which implies that $\hat{W}$ is a continuous process. By an analogous approach proving (4.19), we can also show the covariance functions among $\hat{W}_k$, $\hat{W}_k^c$, $k = 1, 2$, and $\hat{W}$. We omit the details here for brevity. □

6.3. *Preliminaries.* In this section, we will establish some preliminary results in order to prove Theorem 4.1. We first give representations for the processes $\boldsymbol{X}^n$, $\boldsymbol{Y}^n$ and S^n. Define the empirical processes driven by the residual service times $\{\tilde{\eta}_k^i : i \ge 1\}$, for $k = 1, 2$, by

$$
\hat{U}_k^{n,Y}(x) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor n(\bar{Y}_{k'}^n(0))\rfloor} (\mathbf{1}(\tilde{\eta}_k^{i,Y_k} \le x) - F_{k,e}(x)), \quad x \ge 0,
$$

and define the empirical process driven by the residual service vector $\{\boldsymbol{\eta}^{i,J} : i \ge 1\}$ as follows:

$$
\hat{U}^n(\boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor n(\bar{J}^n(0))\rfloor} (\mathbf{1}(\tilde{\boldsymbol{\eta}}^{i,J} \le \boldsymbol{x}) - F_{1,e}(x_1)F_{2,e}(x_2)), \quad \boldsymbol{x} \ge \mathbf{0}.
$$

Without abuse of notation, we write $\hat{U}^n(t) \equiv \hat{U}^n(t, t)$ for $t \ge 0$. Let $\hat{U}_1^n(t) := \hat{U}^n(t, \infty)$ and $\hat{U}_2^n(t) := \hat{U}^n(\infty, t)$, $t \ge 0$.

LEMMA 6.1. *The processes $\hat{\boldsymbol{X}}^n$, $\hat{\boldsymbol{Y}}^n$ and $\hat{S}^n$ defined in (4.6) have the following representations: for $t \ge 0$ and $k = 1, 2$,*

$$
\begin{aligned}
\hat{X}_k^n(t) = \hat{X}_k^{n,0}(t) &- N_k\sqrt{n}(1 - \rho_k^n)F_{k,e}(t)(t) \\
&- \hat{U}_k^{n,Y}(t) - \hat{V}_k^{n,0}(t) - \hat{U}_k^n - \hat{W}_k^n(t) \\
&+ \int_0^t F_k^c(t - s)d\hat{A}^n(s) + \int_0^t (\hat{X}_k^n(t - s))^+ dF_k(s),
\end{aligned} \tag{6.9}
$$

$$\begin{aligned}\hat{Y}_k^n(t) = {} & \hat{Y}_k^{n,0}(t) - \hat{U}_{k'}^{n,Y}(t) + N_k\sqrt{n}(1-\rho_k^n)F_{k,e}(t) + \hat{U}_k^n(t) - \hat{U}^n(t) \\ & + \hat{V}_k^{n,0}(t) - \hat{V}^{n,0}(t) + \hat{W}_k^{n,c}(t) - \hat{M}^{n,0}(t) - \hat{\Psi}^n(t) \\ & + \int_0^t F_k(t-s)d\hat{A}^n(s) - \int_0^t (\hat{X}_k^n(t-s))^+ dF_k(s),\end{aligned} \tag{6.10}$$

$$\begin{aligned}\hat{S}^n(t) = {} & \hat{S}^{n,0}(t) + \hat{U}_1^{n,Y}(t) + \hat{U}_2^{n,Y}(t) + \hat{U}^n(t) \\ & + \hat{V}^{n,0}(t) + \hat{M}^{n,0}(t) + \hat{W}^n(t) + \hat{\Psi}^n(t),\end{aligned} \tag{6.11}$$

where

$$\hat{X}_k^{n,0}(t) := \hat{X}_k^n(0)F_{k,e}^c(t) + (\hat{X}_k^n(0))^+(F_k^c(t) - F_{k,e}^c(t)), \tag{6.12}$$

$$\begin{aligned}\hat{S}^{n,0}(t) := {} & \hat{Y}_2^n(0)F_{1,e}(t) + \hat{Y}_1^n(0)F_{2,e}(t) + \hat{Z}_2^n(0)F_1(t)F_{2,e}(t) \\ & + \hat{Z}_1^n(0)F_{1,e}(t)F_2(t) + \hat{J}^n(0)F_{1,e}(t)F_{2,e}(t) + \hat{I}^n(0)F_m(t),\end{aligned} \tag{6.13}$$

$$\begin{aligned}\hat{Y}_k^{n,0}(t) := {} & \hat{Y}_k^n(0) + \hat{X}_k^n(0)F_{k,e}(t) \\ & + (\hat{X}_k^n(0))^+(F_k(t) - F_{k,e}(t)) - \hat{S}^{n,0}(t),\end{aligned} \tag{6.14}$$

$$\hat{W}_k^n(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} (\mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i \le t) - F_k(t - \tau_i^n - w_k^{n,i})), \tag{6.15}$$

$$\begin{aligned}\hat{W}^n(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \big(\mathbf{1}(\tau_i^n + w_j^{n,i} + \eta_j^i \le t, \forall j) \\ - F(t - \tau_i^n - w_1^{n,i}, t - \tau_i^n - w_2^{n,i})\big),\end{aligned} \tag{6.16}$$

$$\hat{W}_k^{n,c}(t) := \hat{W}_k^n(t) - \hat{W}^n(t), \tag{6.17}$$

$$\begin{aligned}\hat{\Psi}^n(t) := {} & \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} F(t - \tau_i^n - w_1^{n,i}, t - \tau_i^n - w_2^{n,i}) \\ & - \frac{\lambda^n}{\sqrt{n}} \int_0^t \int_0^t ((t-s_1) \wedge (t-s_2)) dF(s_1, s_2),\end{aligned} \tag{6.18}$$

$$\hat{M}^{n,0}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (F_{1,e}(t)F_2(t-\tilde{w}_2^{n,i,R}) - F_{1,e}(t)F_2(t)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_2^n(0)} (F_1(t-\tilde{w}_1^{n,i,R})F_{2,e}(t) - F_1(t)F_{2,e}(t)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{I^n(0)} (F(t-\tilde{w}_1^{n,i,I}, t-\tilde{w}_2^{n,i,I}) - F_m(t)), \tag{6.19}$$

$$\hat{V}_k^{n,0}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_k^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,Z} \le t) - F_{k,e}(t)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{Q_k^n(0)} (\mathbf{1}(\tilde{w}_k^{n,i} + \eta_k^{i,Q} \le t) - F_k(t-\tilde{w}_k^{n,i})), \tag{6.20}$$

$$\begin{aligned}\hat{V}^{n,0}(t) := {} & \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (\mathbf{1}(\tilde{\eta}_1^{i,Z} \le t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \le t) - F_{1,e}(t)F_2(t-\tilde{w}_2^{n,i,R})) \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_2^n(0)} (\mathbf{1}(\tilde{w}_1^{n,i,R} + \eta_1^{i,R} \le t, \tilde{\eta}_2^{i,Z} \le t) - F_1(t-\tilde{w}_1^{n,i,R})F_{2,e}(t)) \\ & + \frac{1}{\sqrt{n}} \sum_{i=1}^{I^n(0)} (\mathbf{1}(\tilde{w}_j^{n,i,I} + \eta_j^{i,I} \le t, \forall j) - F(t-\tilde{w}_1^{n,i,I}, t-\tilde{w}_2^{n,i,I})).\end{aligned} \tag{6.21}$$

PROOF. From the system dynamic equation of $X_k^n(t)$ in (4.3) and the decomposition of $X_k^n(0)$ in (4.2), we obtain, for $k = 1, 2$, and $t \ge 0$,

$$\begin{aligned}& X_k^n(t) - N_k^n \\ & = \sum_{i=1}^{Y_{k'}^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,Y_k} > t) - F_{k,e}^c(t)) + \sum_{i=1}^{Z_k^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,Z} > t) - F_{k,e}^c(t)) \\ & \quad + \sum_{i=1}^{J^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,J} > t) - F_{k,e}^c(t)) + \sum_{i=1}^{Q_k^n(0)} (\mathbf{1}(\tilde{w}_k^{n,i} + \eta_k^{i,Q} > t) - F_k^c(t-\tilde{w}_k^{n,i})) \\ & \quad + \sum_{i=1}^{A^n(t)} (\mathbf{1}(\tau_i^n + w_k^{n,i} + \eta_k^i > t) - F_k^c(t-\tau_i^n - w_k^{n,i})) + Y_{k'}^n(0)F_{k,e}^c(t)\end{aligned}$$

$$
\begin{aligned}
&+ Z_k^n(0)F_{k,e}^c(t) + J^n(0)F_{k,e}^c(t) + \sum_{i=1}^{Q_k^n(0)} F_k^c(t-\tilde{w}_k^{n,i}) \\
&+ \sum_{i=1}^{A^n(t)} F_k^c(t-\tau_i^n - w_k^{n,i}) - N_k^n \\
&= -\sqrt{n}\hat{U}_k^{n,Y}(t) - \sqrt{n}\hat{U}_k^n(t) - \sqrt{n}\hat{W}_k^n(t) + B_k^n(0)F_{k,e}^c(t) + \sum_{i=1}^{Q_k^n(0)} F_k^c(t-\tilde{w}_k^{n,i}) \\
&+ \sum_{i=1}^{Z_k^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,Z} > t) - F_{k,e}^c(t)) + \sum_{i=1}^{Q_k^n(0)} (\mathbf{1}(\tilde{w}_k^{n,i} + \eta_k^{i,Q} > t) - F_k^c(t-\tilde{w}_k^{n,i})) \\
&+ \sum_{i=1}^{A^n(t)} (F_k^c(t-\tau_i^n - w_k^{n,i}) - F_k^c(t-\tau_i^n)) + \sum_{i=1}^{A^n(t)} F_k^c(t-\tau_i^n) - N_k^n.
\end{aligned}
$$

We then have

$$
\begin{aligned}
X_k^n(t) - N_k^n = &-\sqrt{n}(\hat{U}_k^{n,Y}(t) + \hat{U}_k^n(t) + \hat{W}_k^n(t) + \hat{V}_k^{n,0}(t)) \\
&+ B_k^n(0)F_{k,e}^c(t) + \sum_{i=1}^{Q_k^n(0)} F_k^c(t-\tilde{w}_k^{n,i}) \\
&+ \sum_{i=1}^{A^n(t)} (F_k^c(t-\tau_i^n - w_k^{n,i}) - F_k^c(t-\tau_i^n)) \\
&+ \int_0^t F_k^c(t-s)dA^n(s) - N_k^n,
\end{aligned} \tag{6.22}
$$

for $t \geq 0$. Note that, by Propostion 2.1 of [51], we have, for $t \geq 0$ and $k = 1, 2$,

$$
\begin{aligned}
&\sum_{i=1}^{A^n(t)} (F_k^c(t-\tau_i^n - w_k^{n,i}) - F_k^c(t-\tau_i^n)) \\
&\quad = \int_0^t (X_k^n(t-s) - N_k^n)^+ dF_k(s) - \sum_{i=1}^{Q_k^n(0)} (F_k^c(t-\tilde{w}_k^{n,i}) - F_k^c(t)).
\end{aligned}
$$

Thus, following from (6.22), we obtain, for $t \geq 0$,

$$
\begin{aligned}
X_k^n(t) - N_k^n = &-\sqrt{n}(\hat{U}_k^{n,Y}(t) + \hat{U}_k^n(t) + \hat{W}_k^n(t) + \hat{V}_k^{n,0}(t)) \\
&+ Q_k^n(0)F_k^c(t) + B_k^n(0)F_{k,e}^c(t) - N_k^n
\end{aligned}
$$

$$+\int_0^t (X_k^n(t-s)-N_k^n)^+ dF_k(s)+\int_0^t F_k^c(t-s)dA^n(s). \tag{6.23}$$

Notice that, for $t\geq 0$ and $k=1,2$,

$$\begin{aligned}
&Q_k^n(0)F_k^c(t)+B_k^n(0)F_{k,e}^c(t)-N_k^n+\int_0^t F_k^c(t-s)dA^n(s)\\
&=Q_k^n(0)F_k^c(t)+(X_k^n(0)-Q_k^n(0))F_{k,e}^c(t)-N_k^n\\
&\qquad+\int_0^t F_k^c(t-s)d(A^n(s)-\lambda^n s)+\lambda^n\int_0^t F_k^c(t-s)ds\\
&=Q_k^n(0)(F_k^c(t)-F_{k,e}^c(t))+(X_k^n(0)-N_k^n)F_{k,e}^c(t)\\
&\qquad-N_k^n(1-\rho_k^n)F_{k,e}(t)+\sqrt{n}\int_0^t F_k^c(t-s)d\hat{A}^n(s).
\end{aligned} \tag{6.24}$$

Plugging (6.24) into (6.23), and dividing $\sqrt{n}$ on both sides of (6.23), we then obtain (6.9) holds.

Next, to derive the representation of $\hat{S}^n$, we center each term in (4.4) by its mean conditional on arrival times, residual waiting times and waiting times, and by some algebraic manipulations, we obtain, for $t\geq 0$,

$$\begin{aligned}
&S^n(t)-\tilde{S}^n(t)\\
&=(Y_2^n(0)-n\bar{Y}_2(0))F_{1,e}(t)+(Y_1^n(0)-n\bar{Y}_1(0))F_{2,e}(t)+Z_2^n(0)F_1(t)F_{2,e}(t)\\
&\quad+Z_1^n(0)F_{1,e}(t)F_2(t)+(J_1^n(0)-n\bar{J}(0))\wedge(J_2^n(0)-n\bar{J}(0))F_{1,e}(t)F_{2,e}(t)\\
&\quad+I^n(0)F_m(t)+\sqrt{n}\Big(\hat{U}_1^{n,Y}(t)+\hat{U}_2^{n,Y}(t)+\hat{U}^n(t)+\hat{V}^{n,0}(t)\\
&\qquad\qquad+\hat{W}^n(t)+\hat{M}^{n,0}(t)+\hat{\Psi}^n(t)\Big).
\end{aligned}$$

Dividing $\sqrt{n}$ on both sides of the previous equation, we then have (6.11).

To show the representation of $\hat{Y}_k^n$, $k=1,2$, by (4.5) and the definition of $\tilde{Y}_k^n$ in (4.9), we have, for $t\geq 0$,

$$\begin{aligned}
Y_k^n(t)-\tilde{Y}_k^n(t)&=(Y_k^n(0)-n\bar{Y}_k(0))+(X_k^n(0)-N_k^n)\\
&\quad+(A^n(t)-\lambda^n t)-(X_k^n(t)-N_k^n)-(S(t)-\tilde{S}^n(t)).
\end{aligned} \tag{6.25}$$

Dividing $\sqrt{n}$ on both sides of (6.25), and plugging (6.9) and (6.11), we obtain (6.10). □

Let $\bar{E}_k^n:=n^{-1}E_k^n$, $k=1,2$. The weak convergence of $\bar{E}_k^n$, $k=1,2$, is established in Lemma 6.2.

LEMMA 6.2. *Under Assumptions 1 and 4-8,*

$$(\bar{E}_1^n, \bar{E}_2^n) \Rightarrow (\bar{a}, \bar{a}) \quad \text{in} \quad \mathbb{D}^2 \quad \text{as} \quad n \to \infty,$$

where $\bar{a}(t) = \lambda t$, $t \geq 0$, *is the fluid limit of the arrival process.*

PROOF. Note that, for $k = 1, 2$,

$$A^n(t) - (X_k^n(t) - N_k^n)^+ \leq E_k^n(t) \leq A^n(t), \quad t \geq 0, \quad a.s.$$

Thus, for each $T > 0$ and $\epsilon > 0$,

(6.26)
$$\begin{aligned} &P\Big(\sup_{0\leq t\leq T} |\bar{E}_k^n(t) - \bar{a}(t)| > \epsilon \Big) \\ &\leq P\Big(\sup_{0\leq t\leq T} |\bar{A}^n(t) - \bar{a}(t)| > \frac{\epsilon}{2} \Big) + P\Big(\sup_{0\leq t\leq T} |\bar{E}_k^n(t) - \bar{A}^n(t)| > \frac{\epsilon}{2} \Big) \\ &\leq P\Big(\sup_{0\leq t\leq T} |\bar{A}^n(t) - \bar{a}(t)| > \frac{\epsilon}{2} \Big) + P\Big(\sup_{0\leq t\leq T} |(\bar{X}_k^n(t) - \bar{N}_k^n)^+| > \frac{\epsilon}{2} \Big). \end{aligned}$$

From Assumptions 4 and 5, we first see that the first term on the RHS of (6.26) goes to 0 as $n \to \infty$. Also, by Assumption 6 and Corollary 5.1 of [51], we have $\bar{X}_k^n \Rightarrow N_k$ in $\mathbb{D}$ as $n \to \infty$, which implies that $(\bar{X}_k^n(t) - \bar{N}_k^n)^+ \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$. Thus, the second term on the RHS of (6.26) converges to 0 as $n \to \infty$. Hence, we obtain that for $k = 1, 2$, $\bar{E}_k^n \Rightarrow \bar{a}$ in $\mathbb{D}$ as $n \to \infty$. Since $\bar{a}$ is a deterministic function in $\mathbb{R}_+$, by Theorem 11.4.5 of [58] we see that Lemma 6.2 holds. □

Lemma 6.2 directly implies the stochastic boundedness of the processes $\bar{E}_k^n$, $k = 1, 2$, which is stated below.

LEMMA 6.3. *For each* $k = 1, 2$, *and* $T \geq 0$, *there exists a* $\kappa \geq 0$ *such that*

$$P(\bar{E}_k^n(T) \geq \kappa) \to 0 \quad \text{as} \quad n \to \infty.$$

6.4. *Convergence of* $\hat{W}^n$, $\hat{W}_k^n$ *and* $\hat{W}_k^{n,c}$. It follows from the definitions of $\hat{W}^n$ in (6.16), $\hat{W}_k^n$ in (6.15), $\hat{W}_k^{n,c}$ in (6.17) and E_k^n, $k = 1, 2$, that

$$\hat{W}^n(t) = \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \leq t, \forall j) d\hat{K}^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}), \tag{6.27}$$

$$\hat{W}_k^n(t) = \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_k + x_k \leq t) d\hat{K}^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}), \tag{6.28}$$

and

$$\hat{W}_k^{n,c}(t) = \hat{W}_k^n(t) - \hat{W}^n(t)$$
$$= \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_k + t_k \le t, s_{k'} + t_{k'} > t) d\hat{K}^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}), \tag{6.29}$$

for $t \ge 0$.

The integrals above are well-defined as Stieltjes integrals for functions of bounded variation as integrators. We will first prove the tightness of these processes. Here we focus on showing the tightness of $\hat{W}^n$, as the tightness of $\hat{W}_k^n$ and $\hat{W}_k^{n,c}$, $k = 1, 2$, follows from a similar argument.

Note that

$$\hat{K}^n(t_1, t_2, \boldsymbol{x}) = \hat{K}_1^n(t_1, t_2, \boldsymbol{x}) + \hat{K}_2^n(t_1, t_2, \boldsymbol{x}), \quad t_1, t_2 \ge 0, \ \boldsymbol{x} \in \mathbb{R}_+^2,$$

where for $t_1, t_2 \ge 0, \ \boldsymbol{x} \in \mathbb{R}_+^2$ and $i = 1, 2$,

$$\hat{K}_i^n(t_1, t_2, \boldsymbol{x}) := \sqrt{n} \bar{K}_i^n(t_1, t_2, \boldsymbol{x}),$$

and $\bar{K}_i^n(t_1, t_2, \boldsymbol{x})$ are defined in (5.7) and (5.8). We then decompose $\hat{W}^n$ into two processes, $\hat{G}^n := \{\hat{G}^n(t) : t \ge 0\}$ and $\hat{H}^n := \{\hat{H}^n(t) : t \ge 0\}$ as follows:

$$\hat{W}^n(t) = \hat{H}^n(t) + \hat{G}^n(t), \quad t \ge 0,$$

where

$$\hat{H}^n(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \le t, \forall j) d\hat{K}_1^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}), \tag{6.30}$$

$$\hat{G}^n(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \le t, \forall j) d\hat{K}_2^n(\bar{E}_1^n(s_1), \bar{E}_2^n(s_2), \boldsymbol{x}). \tag{6.31}$$

Set $\hat{H}^n := \{\hat{H}^n(t) : t \ge 0\}$ and $\hat{G}^n := \{\hat{G}^n(t) : t \ge 0\}$. We prove the tightness property for $\hat{H}^n$ and $\hat{G}^n$ in the next lemma.

LEMMA 6.4. *Under Assumptions 1 and 4-8, the sequences* $\{\hat{H}^n : n \ge 1\}$ *and* $\{\hat{G}^n : n \ge 1\}$ *are tight.*

Before proving the lemma, we present some preliminary results. We use the notation $\hat{\tau}_j^{n,i}$ to be the time at which task j of job i enters service after time 0−, i.e., $\hat{\tau}_j^{n,i} = \inf\{t \ge 0 : E_j^n(t) \ge i\}$, $j = 1, 2$. Recall the

definition of the processes $H^{n,i}(t)$ in (5.12), $t \geq 0$. We define the filtration $\mathcal{H}^n := \{\mathcal{H}^n_t : t \geq 0\}$ by

$$\begin{aligned}\mathcal{H}^n_t &:= \sigma(X^n_j(0), Y^n_j(0), \forall j) \\ &\vee \sigma\left(\mathbf{1}(\eta^i_j \leq s - \hat{\tau}^{n,i}_j, \forall j), s \leq t, i = 1, ..., E^n_1(t) \wedge E^n_2(t)\right) \\ &\vee \sigma(E^n_j(s), s \leq t, \forall j) \vee \sigma(\xi^n_i, i \geq 1) \vee \mathcal{N},\end{aligned} \tag{6.32}$$

where $\mathcal{N}$ includes all the null sets. Note that here we include the initial quantities in the filtration $\mathcal{H}^n := \{\mathcal{H}^n_t : t \geq 0\}$ in (5.15). It is easy to verify that $\mathcal{H}^n$ is a filtration and satisfies the usual conditions [23].

Define

$$\hat{H}^n_\kappa(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n_1(t)\wedge E^n_2(t)\wedge\kappa} H^{n,i}(t) \tag{6.33}$$

for $\kappa \in \mathbb{N}$ and $t \geq 0$. Set $\hat{H}^n_\kappa := \{\hat{H}^n_\kappa(t) : t \geq 0\}$. We first state the martingale property of $\hat{H}^n_\kappa$ in Lemma 6.5. The proof is identical to that of Lemma 5.6, so we omit the details here for brevity.

LEMMA 6.5. *Under Assumptions 1 and 4-8, for each $\kappa \geq 1$, the process $\hat{H}^n_\kappa$ is an $\mathcal{H}^n$-square-integrable martingale with the predictable quadratic variation process*

$$\langle \hat{H}^n_\kappa \rangle(t) = \frac{1}{n} \sum_{i=1}^{E^n_1(t)\wedge E^n_2(t)\wedge\kappa} \int_0^{\eta^i_1 \wedge (t-\hat{\tau}^{n,i}_1)^+} \int_0^{\eta^i_2 \wedge (t-\hat{\tau}^{n,i}_2)^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}), \quad t \geq 0.$$

Define a multiparameter process $\hat{T}^n := \{\hat{T}^n(t_1, t_2, \boldsymbol{x}) : t_1 \geq 0, t_2 \geq 0, \boldsymbol{x} \in \mathbb{R}^2_+\}$ by

$$\hat{T}^n(t_1, t_2, \boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n_1(t_1)\wedge E^n_2(t_2)} (\mathbf{1}(\boldsymbol{\eta}^i \geq \boldsymbol{x}) - F^c(\boldsymbol{x})),$$

for $t_1, t_2 \geq 0$, $\boldsymbol{x} \in \mathbb{R}^2_+$. We then obtain the convergence of the processes $\hat{T}^n$ following from a similar argument as Proposition 4.1.

LEMMA 6.6. *Under Assumptions 1 and 4-8,*

$$\hat{T}^n \Rightarrow \hat{T} \quad \text{in} \quad \mathbb{D}([0,\infty)^2, \mathbb{D}([0,\infty)^2, \mathbb{R})) \quad \text{as} \quad n \to \infty,$$

where $\hat{T}(t_1, t_2, \boldsymbol{x})$ *is a continuous Gaussian random field with mean function* $E[\hat{T}(t_1, t_2, \boldsymbol{x})] = 0$ *and covariance function*

$$Cov(\hat{T}(t_1, t_2, \boldsymbol{x}), \hat{T}(s_1, s_2, \boldsymbol{y})) = [(t_1 \wedge t_2) \wedge (s_1 \wedge s_2)](F^c(\boldsymbol{x} \vee \boldsymbol{y}) - F^c(\boldsymbol{x})F^c(\boldsymbol{y})),$$

for $t_i, s_i \geq 0$, $i = 1, 2$, *and* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2_+$.

We can then rewrite $\hat{G}^n$ in (6.31) as

$$\hat{G}^n(t) = \int_0^t \int_0^t \frac{\hat{T}^n(t - x_1, t - x_2, \boldsymbol{x})}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}), \quad t \geq 0.$$

We are now ready to prove Lemma 6.4.

Proof of Lemma 6.4. We first prove the tightness of $\{\hat{H}^n\}$. We follow the argument of Lemma 3.7 of [29] by using Aldous' sufficient condition (see, e.g., [7]) to verify the tightness of $\{\hat{H}^n : n \geq 1\}$. This requires us to check for $L > 0$ and $\epsilon > 0$,

$$\lim_{\tilde{\kappa} \to \infty} \limsup_{n \to \infty} P\left(\sup_{0 \leq t \leq L} |\hat{H}^n(t)| > \tilde{\kappa} \right) = 0, \tag{6.34}$$

and

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{\tau \in \mathcal{C}_L^n} P\left(\sup_{0 \leq t \leq \delta} |\hat{H}^n(\tau + t) - \hat{H}^n(t)| > \epsilon \right) = 0, \tag{6.35}$$

where $\mathcal{C}_L^n$ is the set of all $\mathcal{H}^n$-stopping times bounded by L, where the filtration $\mathcal{H}^n$ is defined in (6.32). Since the proofs of (6.34) and (6.35) are analogous, we only verify (6.35) here. Fix $T > 0$. For each $\epsilon > 0$ and $\kappa \in \mathbb{N}$, by Lemma 6.5, we have

$$\begin{aligned} &P\left(\sup_{0 \leq t \leq T} |\hat{H}^n(\tau + t) - \hat{H}^n(\tau)| > \epsilon \right) \\ &\leq P\left(\bar{E}_1^n(T) \wedge \bar{E}_2^n(T) > \kappa \right) + P\left(\sup_{0 \leq t \leq T} \left| \hat{H}_{\kappa n}^n(\tau + t) - \hat{H}_{\kappa n}^n(\tau) \right| > \epsilon \right). \end{aligned}$$

Lemma 6.3 implies that for κ sufficiently large, the first term on the RHS of the inequality above goes to 0 as $n \to \infty$. Next, we only need to show the second term converges to 0 as $n \to \infty$. By the Lenglart-Rebolledo inequality [32], it follows that for any $\gamma > 0$,

$$P\left(\sup_{0 \leq t \leq T} \left| \hat{H}_{\kappa n}^n(\tau + t) - \hat{H}_{\kappa n}^n(\tau) \right| > \epsilon \right)$$

$$\leq \frac{\gamma}{\epsilon^2} + P\left(\langle \hat{H}^n_{\kappa n}\rangle(\tau+T) - \langle \hat{H}^n_{\kappa n}\rangle(\tau) > \gamma\right).$$

Note from Lemma 6.5 that

$$\langle \hat{H}^n_{\kappa n}\rangle(\tau+T) - \langle \hat{H}^n_{\kappa n}\rangle(\tau) \leq \frac{1}{n} \sup_{s\leq L, |t-s|\leq T} \sum_{i=E^n_1(s)\wedge E^n_2(s)+1}^{E^n_1(t)\wedge E^n_2(t)} \int_0^{\eta^i_1}\int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}).$$

It follows (5.26) and the FLLN that

$$\frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} \int_0^{\eta^i_1}\int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \Rightarrow t \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n\to\infty.$$

This convergence being uniform in t on bounded intervals together with Lemma 6.2 implies that

$$\lim_{T\to 0}\limsup_{n\to\infty} P\left(\frac{1}{n} \sup_{s\leq L, |t-s|\leq T} \sum_{i=E^n_1(s)\wedge E^n_2(s)+1}^{E^n_1(t)\wedge E^n_2(t)} \int_0^{\eta^i_1}\int_0^{\eta^i_2} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) > \gamma\right) = 0,$$

from which we obtain

$$P\left(\langle \hat{H}^n_{\kappa n}\rangle(\tau+T) - \langle \hat{H}^n_{\kappa n}\rangle(\tau) > \gamma\right) \to 0 \quad \text{as} \quad n\to\infty,$$

which completes the proof of the tightness of $\{\hat{H}^n\}$.

Next, we will prove the tightness of $\{\hat{G}^n\}$. For $\epsilon > 0$ and $t \geq 0$, we decompose $\hat{G}^n$ as

$$\hat{G}^n(t) = \hat{G}^{n,\epsilon}_1(t) + \hat{G}^{n,\epsilon}_2(t),$$

where

$$\hat{G}^{n,\epsilon}_1(t) := \int_0^t\int_0^t \frac{\hat{T}^n(t-x_1, t-x_2, \boldsymbol{x})\mathbf{1}(F^c(\boldsymbol{x})\geq\epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}),$$
$$\hat{G}^{n,\epsilon}_2(t) := \int_0^t\int_0^t \frac{\hat{T}^n(t-x_1, t-x_2, \boldsymbol{x})\mathbf{1}(F^c(\boldsymbol{x})<\epsilon)}{F^c(\boldsymbol{x})} dF(\boldsymbol{x}).$$

To show the tightness of $\{\hat{G}^n\}$, by Lemma 3.32 of Chapter VI in [29], it suffices to prove the following:

(i) $\{\hat{G}^{n,\epsilon}_1 : n \geq 1\}$ is tight;
(ii) For each $\delta > 0$ and $T > 0$,

$$\lim_{\epsilon\to 0}\lim_{n\to\infty} P\left(\sup_{0\leq t\leq T} |\hat{G}^{n,\epsilon}_2(t)| > \delta\right) = 0.$$

The rest of the proof proceeds analogously as in Lemma 6.6 of [33] and Lemma 3.4 of [29] by applying Lemma 6.6. We omit the details here for brevity. □

To proceed with the proof for the convergence of $\hat{W}^n$, $\hat{W}_k^n$ and $\hat{W}_k^{n,c}$, $k = 1, 2$, we need to show the convergence of the finite-dimensional distributions of $\hat{W}^n$, $\hat{W}_k^n$ and $\hat{W}_k^{n,c}$, $k = 1, 2$. Recall Lemma 5.2 of [29]. For $x_1, x_2 \geq 0$ and $\boldsymbol{y} := (y_1, y_2) \in \mathbb{R}_+^2$, let $\chi_i(x_1, x_2, \boldsymbol{y})$, $i \in \mathbb{N}$, be real-valued bounded Borel functions such that $E[\chi_i(x_1, x_2, \boldsymbol{\eta}^i)] = 0$. Define the processes $\zeta_\kappa^n := \{\zeta_\kappa^n(t) : t \geq 0\}$ and $\langle \zeta_\kappa^n \rangle := \{\langle \zeta_\kappa^n \rangle(t) : t \geq 0\}$, $\kappa \in \mathbb{N}$, by

$$\begin{aligned} \zeta_\kappa^n(t) &:= \sum_{i=1}^{E_1^n(t) \wedge E_2^n(t) \wedge \kappa} \chi_i(\hat{\tau}_1^{n,i}, \hat{\tau}_2^{n,i}, \boldsymbol{\eta}^i) \\ \langle \zeta_\kappa^n \rangle(t) &:= \sum_{i=1}^{E_1^n(t) \wedge E_2^n(t) \wedge \kappa} \bar{\chi}_i(\hat{\tau}_1^{n,i}, \hat{\tau}_2^{n,i}), \end{aligned} \tag{6.36}$$

where $\bar{\chi}_i(x_1, x_2) := E[(\chi_i(x_1, x_2, \boldsymbol{\eta}^i))^2]$. We also set the σ-fields $\hat{\mathcal{F}}_t^n := \sigma(\hat{\tau}_1^{n,i}, \hat{\tau}_2^{n,i}, \boldsymbol{\eta}^i, 1 \leq i \leq \lfloor t \rfloor) \vee \mathcal{N}$ and $\mathcal{F}_t^n := \sigma((\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}) \wedge (\hat{\tau}_1^{n, E_1^n(t) \wedge E_2^n(t)+1} \vee \hat{\tau}_2^{n, E_1^n(t) \wedge E_2^n(t)+1}), \boldsymbol{\eta}^{i \wedge (E_1^n(t) \wedge E_2^n(t))}, i \geq 1) \vee \mathcal{N}$, and define the filtrations $\hat{\mathcal{F}}^n := \{\hat{\mathcal{F}}_t^n : t \geq 0\}$ and $\mathcal{F}^n := \{\mathcal{F}_t^n : t \geq 0\}$, where $\mathcal{N}$ includes all the null sets. We can then show the following results.

LEMMA 6.7. *(i) $\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}$, $i = 1, 2, \ldots$, are $\mathcal{F}^n$-stopping times, and the following inclusions hold: $\mathcal{F}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}} \supset \hat{\mathcal{F}}_{i+1}^n$, $\mathcal{G}_i^n \subset \hat{\mathcal{F}}_i^n$, where $\mathcal{G}_i^n := \sigma(\mathcal{B} \cap \{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > t\}, t \geq 0, \mathcal{B} \in \mathcal{F}_t^n)$;*

(ii) The process $E_1^n \wedge E_2^n := \{E_1^n(t) \wedge E_2^n(t) : t \geq 0\}$ is $\mathcal{F}^n$-predicatable;

(iii) The processes ζ_κ^n, $\kappa = 1, 2, \ldots$, are $\mathcal{F}^n$-square-integrable martingales with the processes $\langle \zeta_\kappa^n \rangle$ as predictable quadratic-variation processes.

PROOF. The proof follows from a similar argument as the proof of Lemma 5.2 of [29]. □

Now, we are ready to prove the convergence of $\hat{W}^n$, joint with $\hat{W}_k$ and $\hat{W}_k^c$, $k = 1, 2$.

LEMMA 6.8. *Under Assumptions 1 and 4-8,*

$$(\hat{W}_1^n, \hat{W}_2^n, \hat{W}_1^{n,c}, \hat{W}_2^{n,c}, \hat{W}^n) \Rightarrow (\hat{W}_1, \hat{W}_2, \hat{W}_1^c, \hat{W}_2^c, \hat{W}) \quad \text{in} \quad \mathbb{D}^5 \quad \text{as} \quad n \to \infty,$$

where $\hat{W}_k$, $\hat{W}_k^c$ and $\hat{W}$ are defined in (4.16), (4.18) and (4.17), $k = 1, 2$, respectively.

PROOF. Lemma 6.4 imply the tightness of $\hat{W}^n$. And a similar argument can also be used to show the tightness of $\hat{W}_k^n$, $\hat{W}_k^{n,c}$, $k = 1, 2$. Define the processes $\tilde{W}^n := \{\tilde{W}^n(t) : t \geq 0\}$, $\tilde{W}_k^n := \{\tilde{W}_k^n(t) : t \geq 0\}$ and $\tilde{W}_k^{n,c} := \{\tilde{W}_k^{n,c}(t) : t \geq 0\}$, $k = 1, 2$, by

$$\tilde{W}^n(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \leq t, \forall j) d\hat{K}^n(\lambda s_1, \lambda s_2, \boldsymbol{x}),$$

$$\tilde{W}_k^n(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_k + x_k \leq t) d\hat{K}^n(\lambda s_1, \lambda s_2, \boldsymbol{x}), \quad k = 1, 2,$$

and

$$\tilde{W}_k^{n,c}(t) := \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_k + x_k \leq t, s_{k'} + x_{k'} > t) d\hat{K}^n(\lambda s_1, \lambda s_2, \boldsymbol{x}),$$

for $t \geq 0$. Tightness of $\tilde{W}^n$, $\tilde{W}_k^n$ and $\tilde{W}_k^{n,c}$, $k = 1, 2$, can be proved similarly.

It remains to establish their joint convergence in finite-dimensional distributions. For that, let

(6.37)

$$\begin{aligned}
&\hat{W}^{n,(l)}(t) \\
&\quad := \sum_{i=1}^{l} \sum_{j=1}^{l} \Delta\hat{K}^n((\bar{E}_1^n(s_{i-1}^l), \bar{E}_2^n(s_{j-1}^l), \mathbf{0}); (\bar{E}_1^n(s_i^l), \bar{E}_2^n(s_j^l), t - s_i^l, t - s_j^l)), \\
&\hat{W}_k^{n,(l)}(t) \\
&\quad := \sum_{i=1}^{l} \sum_{j=1}^{l} \Delta\hat{K}_{(k)}^n((\bar{E}_1^n(s_{i-1}^l), \bar{E}^n(s_{j-1}^l), \mathbf{0}); (\bar{E}_1^n(s_i^l), \bar{E}_2^n(s_j^l), t - s_i^l, t - s_j^l)),
\end{aligned}$$

and

$$\hat{W}_k^{n,c,(l)}(t) := \hat{W}_k^{n,(l)}(t) - \hat{W}^{n,(l)}(t),$$

for $k = 1, 2$, where $0 = s_0^l < s_1^l < \cdots < s_l^l = t$ and $\max_{1 \leq i \leq l} |s_i^l - s_{i-1}^l| \to 0$ as $l \to \infty$, and $\hat{K}_{(1)}^n(t_1, t_2, \boldsymbol{x}) := \hat{K}^n(t_1, t_2, x_1, \infty)$ and $\hat{K}_{(2)}^n(t_1, t_2, \boldsymbol{x}) := \hat{K}^n(t_1, t_2, \infty, x_2)$ for $t_1, t_2 \in \mathbb{R}_+$ and $\boldsymbol{x} \in \mathbb{R}_+^2$.

We also define in analogy, for $t \geq 0$,

$$\hat{W}^{(l)}(t) := \sum_{i=1}^{l} \sum_{j=1}^{l} \Delta\hat{K}((\lambda s_{i-1}^l, \lambda s_{j-1}^l, \mathbf{0}); (\lambda s_i^l, \lambda s_j^l, t - s_i^l, t - s_j^l)),$$

$$\hat{W}_k^{(l)}(t) := \sum_{i=1}^{l}\sum_{j=1}^{l} \Delta\hat{K}_{(k)}((\lambda s_{i-1}^l, \lambda s_{j-1}^l, \mathbf{0}); (\lambda s_i^l, \lambda s_j^l, t-s_i^l, t-s_j^l)), \quad k=1,2,$$

$$\hat{W}_k^{c,(l)}(t) := \hat{W}_k^{(l)}(t) - \hat{W}^{(l)}(t), \quad k=1,2,$$

$$\tilde{W}^{n,(l)}(t) := \sum_{i=1}^{l}\sum_{j=1}^{l} \Delta\hat{K}^n((\lambda s_{i-1}^l, \lambda s_{j-1}^l, \mathbf{0}); (\lambda s_i^l, \lambda s_j^l, t-s_i^l, t-s_j^l)),$$

$$\tilde{W}_k^{n,(l)}(t) := \sum_{i=1}^{l}\sum_{j=1}^{l} \Delta\hat{K}_{(k)}^n((\lambda s_{i-1}^l, \lambda s_{j-1}^l, \mathbf{0}); (\lambda s_i^l, \lambda s_j^l, t-s_i^l, t-s_j^l)), \ k=1,2,$$

and

$$\tilde{W}_k^{n,c,(l)}(t) := \tilde{W}_k^{n,(l)}(t) - \tilde{W}^{n,(l)}(t), \quad k=1,2, \tag{6.38}$$

where $\hat{K}_{(1)}(t_1,t_2,\boldsymbol{x}) := \hat{K}(t_1,t_2,x_1,\infty)$ and $\hat{K}_{(2)}(t_1,t_2,\boldsymbol{x}) := \hat{K}(t_1,t_2,\infty,x_2)$ for $t_1,t_2 \in \mathbb{R}_+$ and $\boldsymbol{x} \in \mathbb{R}_+^2$. Set $\tilde{W}_k^{n,(l)} := \{\tilde{W}_k^{n,(l)}(t) : t \geq 0\}$, $\tilde{W}_k^{n,c,(l)} := \{\tilde{W}_k^{n,c,(l)}(t) : t \geq 0\}$, $\hat{W}_k^{(l)} := \{\hat{W}_k^{(l)}(t) : t \geq 0\}$, $\hat{W}_k^{c,(l)} := \{\hat{W}_k^{c,(l)}(t) : t \geq 0\}$, $k=1,2$, $\tilde{W}^{n,(l)} := \{\tilde{W}^{n,(l)}(t) : t \geq 0\}$ and $\hat{W}^{(l)} := \{\hat{W}^{(l)}(t) : t \geq 0\}$. Since $\hat{W}^{(l)}$ converges to $\hat{W}$ as $l \to \infty$ in probability by definition, in order to show the joint convergence in finite dimensional distributions it is sufficient to show the following conditions:

(a) $(\tilde{W}_1^{n,(l)}, \tilde{W}_2^{n,(l)}, \tilde{W}_1^{n,c,(l)}, \tilde{W}_2^{n,c,(l)}, \tilde{W}^{n,(l)}) \overset{d_f}{\Rightarrow} (\hat{W}_1^{(l)}, \hat{W}_2^{(l)}, \hat{W}_1^{c,(l)}, \hat{W}_2^{c,(l)}, \hat{W}^{(l)})$ as $n \to \infty$;

(b) For $\gamma > 0$ and $t > 0$,

$$\begin{aligned}
&\lim_{l\to\infty}\limsup_{n\to\infty} P(|\hat{W}^{n,(l)}(t) - \hat{W}^n(t)| > \gamma) = 0,\\
&\lim_{l\to\infty}\limsup_{n\to\infty} P(|\tilde{W}^{n,(l)}(t) - \tilde{W}^n(t)| > \gamma) = 0,\\
&\lim_{l\to\infty}\limsup_{n\to\infty} P(|\hat{W}_k^{n,(l)}(t) - \hat{W}_k^n(t)| > \gamma) = 0, \quad k=1,2,\\
&\lim_{l\to\infty}\limsup_{n\to\infty} P(|\tilde{W}_k^{n,(l)}(t) - \tilde{W}_k^n(t)| > \gamma) = 0, \quad k=1,2,\\
&\lim_{l\to\infty}\limsup_{n\to\infty} P(|\tilde{W}_k^{n,c,(l)}(t) - \tilde{W}_k^{n,c}(t)| > \gamma) = 0, \quad k=1,2;
\end{aligned} \tag{6.39}$$

(c) For $T > 0$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P\Big(\sup_{0\leq t\leq T} |\tilde{W}^{n,(l)}(t) - \hat{W}^{n,(l)}(t)| > \epsilon\Big) = 0,$$

$$\lim_{n\to\infty} P\Big(\sup_{0\le t\le T}|\tilde{W}_k^{n,(l)}(t)-\hat{W}_k^{n,(l)}(t)|>\epsilon\Big)=0,\quad k=1,2,$$
$$\lim_{n\to\infty} P\Big(\sup_{0\le t\le T}|\tilde{W}_k^{n,c,(l)}(t)-\hat{W}_k^{n,c,(l)}(t)|>\epsilon\Big)=0,\quad k=1,2.$$

First, we focus on the proof of (a). For any $t_{1,i}$, $t_{2,j}$ $t_p\ge 0$, $c_{1,i}$, $c_{2,j}$ and $c_p\in\mathbb{R}$ and positive integers I_1, I_2 and I_3, for $i=1,...,I_1$, $j=1,...,I_2$ and $p=1,...,I_3$, since the distribution function F is continuous by Assumption 1, by the weak convergence of $\hat{K}^n$ to $\hat{K}$ as $n\to\infty$ and the continuity of $\hat{K}$, we immediately see that, as $n\to\infty$,

$$\sum_{k=1}^{2}\sum_{i=1}^{I_k}c_{k,i}\tilde{W}_k^{n,(l)}(t_{k,i})+\sum_{p=1}^{I_3}\tilde{W}^{n,(l)}(t_p)\Rightarrow\sum_{k=1}^{2}\sum_{i=1}^{I_k}c_{k,i}\hat{W}_k^{(l)}(t_{k,i})+\sum_{p=1}^{I_3}c_p\hat{W}^{(l)}(t_p).$$

By the Cramér-Wold theorem (see, e.g., Theorem 3.95 of [13]), we see

$$(\tilde{W}_1^{n,(l)},\tilde{W}_2^{n,(l)},\tilde{W}^{n,(l)})\overset{d_f}{\Rightarrow}(\hat{W}_1^{(l)},\hat{W}_2^{(l)},\hat{W}^{(l)})\quad\text{as}\quad n\to\infty.$$

By (6.38) and (6.38), together with the continuous mapping theorem [7], we conclude that (a) holds.

We will next prove (b). For brevity, we here only provide the proof for (6.39), as other proofs follow similarly. For the points $0=s_0^l<s_1^l<...<s_l^l=t$ satisfying $\max_{1\le i\le l}|s_i^l-s_{i-1}^l|\to 0$ as $l\to\infty$, let

$$\begin{aligned}\chi_i(x_1,x_2,\boldsymbol{y})=&\sum_{p=1}^{l}\sum_{j=1}^{l}\mathbf{1}(s_{p-1}^l<x_1\le s_p^l)\mathbf{1}(s_{j-1}^l<x_2\le s_j^l)\\&\times(\mathbf{1}(t-s_p^l<y_1\le t-x_1,t-s_j^l<y_2\le t-x_2)\\&\qquad-\Delta F((t-s_p^l,t-s_j^l);(t-x_1,t-x_2))).\end{aligned}$$

Then it is easy to verify that

$$\begin{aligned}\bar{\chi}_i(x_1,x_2)&=E[(\chi_i(x_1,x_2,\boldsymbol{\eta}^i))^2]\\&=\sum_{p=1}^{l}\sum_{j=1}^{l}\Big(\mathbf{1}(s_{p-1}^l<x_1\le s_p^l)\mathbf{1}(s_{j-1}^l<x_2\le s_j^l)\\&\qquad\times\Delta F((t-s_p^l,t-s_j^l);(t-x_1,t-x_2))\\&\qquad\times(1-\Delta F((t-s_p^l,t-s_j^l);(t-x_1,t-x_2)))\Big).\end{aligned}$$

Recall from (6.36), by (6.27) and (6.37), for $\kappa\in\mathbb{N}$,

$$\frac{1}{\sqrt{n}}\zeta_\kappa^n(t)=\hat{W}^n(t)-\hat{W}^{n,(l)}(t)\quad\text{on}\quad\{E_1^n(t)\wedge E_2^n(t)\le\kappa\}.$$

Hence, we have

$$\begin{aligned}\frac{1}{n}\langle\zeta_\kappa^n\rangle(t) &\le \frac{1}{n}\sum_{i=1}^{E_1^n(t)\wedge E_2^n(t)}\sum_{p=1}^{l}\sum_{j=1}^{l}\mathbf{1}(s_{p-1}^l<\hat\tau_1^{n,i}\le s_p^l)\mathbf{1}(s_{j-1}^l<\hat\tau_2^{n,i}\le s_j^l)\\ &\qquad\times\Delta F((t-s_p^l,t-s_j^l);(t-s_{p-1}^l,t-s_{j-1}^l))\\ &= \frac{1}{n}\sum_{p=1}^{l}\sum_{j=1}^{l}[(E_1^n(s_p^l)-E_1^n(s_{p-1}^l))\wedge(E_2^n(s_j^l)-E_2^n(s_{j-1}^l))]\\ &\qquad\times\Delta F((t-s_p^l,t-s_j^l);(t-s_{p-1}^l,t-s_{j-1}^l))\\ &\le \sup_{1\le p\le l}\sup_{1\le j\le l}[(\bar E_1^n(s_p^l)-\bar E_1^n(s_{p-1}^l))\wedge(\bar E_2^n(s_j^l)-\bar E_2^n(s_{j-1}^l))].\end{aligned}$$

Then, by Lemma 6.7 with the Lenglart-Rebolledo inequality (see, e.g., [32]), we have that for any $\kappa\in\mathbb{N}$, $\gamma>0$ and $\epsilon>0$,

$$\begin{aligned}&P\big(|\hat W^{n,(l)}(t)-\hat W^n(t)|>\gamma\big)\\ &\le P(E_1^n(t)\wedge E_2^n(t)>n\kappa)+P(n^{-1/2}|\zeta_\kappa^n(t)|>\gamma)\\ &\le P(\bar E_1^n(t)\wedge\bar E_2^n(t)>\kappa)+\frac{\epsilon}{\gamma^2}\\ &\quad+P\bigg(\sup_{1\le p\le l}\sup_{1\le j\le l}[(\bar E_1^n(s_p^l)-\bar E_1^n(s_{p-1}^l))\wedge(\bar E_2^n(s_j^l)-\bar E_2^n(s_{j-1}^l))]>\epsilon\bigg).\end{aligned}$$

By Lemma 6.2 and the fact that $\max_{1\le i\le l}|s_i^l-s_{i-1}^l|\to 0$ as $l\to\infty$, we have both terms on the RHS of the above inequality vanish, i.e.,

$$\begin{aligned}&\lim_{\kappa\to\infty}\limsup_{n\to\infty}P(\bar E_1^n(t)\wedge\bar E_2^n(t)>\kappa)=0,\\ &\lim_{l\to\infty}\limsup_{n\to\infty}P\bigg(\sup_{1\le p\le l}\sup_{1\le j\le l}[(\bar E_1^n(s_p^l)-\bar E_1^n(s_{p-1}^l))\wedge(\bar E_2^n(s_j^l)-\bar E_2^n(s_{j-1}^l))]>\epsilon\bigg)\\ &\qquad=0,\end{aligned}$$

which completes of the proof of (b). It is quite straightforward to see that (c) also holds, since $\hat K$ is continuous.

In summary, we have shown that

$$\begin{aligned}&\big(\hat W_1^n,\hat W_2^n,\ \tilde W_1^n,\tilde W_2^n,\hat W_1^{n,c},\hat W_2^{n,c},\tilde W_1^{n,c},\tilde W_2^{n,c},\hat W^n,\tilde W^n\big)\\ &\Rightarrow\big(\hat W_1,\hat W_2,\hat W_1,\hat W_2,\hat W_1^c,\hat W_2^c,\hat W_1^c,\hat W_2^c,\hat W,\hat W\big)\end{aligned}\tag{6.40}$$

in $\mathbb{D}^{10}$ as $n\to\infty$. Thus Lemma 6.8 has been proved. □

6.5. *Convergence of the Initial Quantities.* In this section, we prove the weak convergence of the initial quantities. We first define $\hat{\mathcal{E}}_k^{n,e} := \{\hat{\mathcal{E}}^{n,e}(t) : t \geq 0\}$, $k = 1, 2$, by

$$\hat{\mathcal{E}}_k^{n,e}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{Q_k^n(0)} \mathbf{1}(\tilde{w}_k^{n,i} > t), \quad t > 0, \quad \text{and} \quad \hat{\mathcal{E}}_k^{n,e}(0) := 0. \tag{6.41}$$

Let $\hat{M}^{n,0} := \{\hat{M}^{n,0}(t) : t \geq 0\}$, where $\hat{M}^{n,0}(t)$ for $t \geq 0$ is defined in (6.19).

LEMMA 6.9. *Under Assumptions 1 and 4-8,*

$$(\hat{\mathcal{E}}_1^{n,e}, \hat{\mathcal{E}}_2^{n,e}, \hat{M}^{n,0}) \Rightarrow (0, 0, 0) \quad \text{in} \quad \mathbb{D}^3 \quad \text{as} \quad n \to \infty.$$

PROOF. By the definition of $\hat{\mathcal{E}}_k^{n,e}$, $k = 1, 2$, it is sufficient to show that for $\epsilon > 0$ and $0 < T_1 < T_2$,

$$\lim_{n\to\infty} P\left(\sup_{T_1 \leq t \leq T_2} \left| \hat{\mathcal{E}}_k^{n,e}(t) \right| > \epsilon \right) = 0.$$

Recall that in the sequence $\{\tilde{w}_k^{n,i} : i = 1, ..., Q_k^n(0)\}$, $\tilde{w}_k^{n,1}$ represents the residual waiting time of the task in the front of the queue and $\tilde{w}_k^{n,Q_k^n(0)}$ represents that of the task in the end of the queue at station k at time $0-$, $k = 1, 2$. Under the non-idling FCFS discipline, $k = 1, 2$,

$$\tilde{w}_k^{n,1} \leq \tilde{w}_k^{n,2} \leq ... \leq \tilde{w}_k^{n,Q_k^n(0)}, \quad a.s. \tag{6.42}$$

We thus obtain, for $k = 1, 2$,

$$\begin{aligned} P\left(\sup_{T_1 \leq t \leq T_2} \left| \hat{\mathcal{E}}_k^{n,e}(t) \right| > \epsilon \right) &\leq P\left(\sup_{T_1 \leq t \leq T_2} \left| \hat{Q}_k^n(0) \mathbf{1}(\tilde{w}_k^{n,Q_k^n(0)} > t) \right| > \epsilon \right) \\ &\leq P\left(\left| \hat{Q}_k^n(0) \mathbf{1}(\tilde{w}_k^{n,Q_k^n(0)} > T_1) \right| > \epsilon \right) \\ &= P\left(\left| \hat{Q}_k^n(0) \mathbf{1}(\tilde{w}_k^{n,Q_k^n(0)} > T_1) \right| > \epsilon, \tilde{w}_k^{n,Q_k^n(0)} > T_1 \right) \\ &\leq P\big(\tilde{w}_k^{n,Q_k^n(0)} > T_1\big). \end{aligned} \tag{6.43}$$

Assumption 8 implies that the RHS of (6.43) converges to 0 as $n \to \infty$, which completes the proof of the convergence of $(\hat{\mathcal{E}}_1^{n,e}, \hat{\mathcal{E}}_2^{n,e})$.

Define $\hat{M}_k^{n,0} := \{\hat{M}_k^{n,0}(t) : t \geq 0\}$ by

$$\hat{M}_k^{n,0}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_k^n(0)} (F_{k,e}(t) F_{k'}(t - \tilde{w}_{k'}^{n,i,R}) - F_{k,e}(t) F_{k'}(t)), \quad k = 1, 2,$$

and

$$\hat{M}_3^{n,0}(t) := \frac{1}{\sqrt{n}} \sum_{i=1}^{I^n(0)} (F(t - \tilde{w}_1^{n,i,I}, t - \tilde{w}_2^{n,i,I}) - F_m(t)),$$

for $t \geq 0$.

To show $\hat{M}^{n,0} \Rightarrow 0$, by the definition of $\hat{M}^{n,0}$ in (6.19), Theorem 11.4.5 of [58] and the continuous mapping theorem [7], it suffices to show that for $k = 1, 2, 3$,

$$\hat{M}_k^{n,0} \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty.$$

Here we only provide the proof for the weak convergence of $\hat{M}_3^{n,0}$ for brevity, as the proofs for $\hat{M}_1^{n,0}$ and $\hat{M}_2^{n,0}$ are similar.

Denote $\tilde{w}_m^{n,l,I} := \max_{k=1,2} \tilde{w}_k^{n,l,I}$ for $l = 1, \cdots, I^n(0)$. By (6.42), we first obtain, for each $t \geq 0$,

$$\begin{aligned} |\hat{M}_3^{n,0}(t)| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{I^n(0)} (F_m(t) - F_m(t - \tilde{w}_m^{n,l,I})) \\ &\leq \hat{I}^n(0)(F_m(t) - F_m(t - \tilde{w}_m^{n,I^n(0),I})). \end{aligned} \tag{6.44}$$

Note the fact that, under Assumption 8, the sequence of $\{\tilde{w}_m^{n,l,I} : l = 1, ..., I^n(0)\}$ converges to 0 *a.s.* as $n \to \infty$, and F_m is uniformly continuous by Assumption 1 and Theorem 1 of [36]. Together with (6.44), it is easy to see, when n is sufficiently large, for each $t \geq 0$, $|\hat{M}_3^{n,0}(t)| \leq \hat{I}^n(0)\gamma^n$, *a.s.*, where $\{\gamma^n : n \geq 1\}$ is a sequence of random variables converging to 0 *a.s.* as $n \to \infty$. Moreover, we have that $\sup_{0 \leq t \leq T} |\hat{M}_3^{n,0}(t)| \leq \hat{I}^n(0)\gamma^n$ for each $T > 0$ *a.s.* Since $\hat{I}^n(0) \Rightarrow \hat{I}(0)$ in $\mathbb{R}$ as $n \to \infty$, we have that $\hat{I}^n(0)\gamma^n \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$. Therefore, we obtain that $\hat{M}_3^{n,0} \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$. This completes the proof of the Lemma. □

LEMMA 6.10. *Under Assumptions 1 and 4-8,*

$$(\hat{V}_1^{n,0}, \hat{V}_2^{n,0}, \hat{V}^{n,0}) \Rightarrow (0, 0, 0) \quad \textit{in} \quad \mathbb{D}^3 \quad \textit{as} \quad n \to \infty.$$

PROOF. We start with the proof of the convergence of $\hat{V}_k^{n,0}$ for $k = 1, 2$. By Theorem 11.4.5 in [58] and the continuous mapping theorem in [7], it suffices to show that for $k = 1, 2$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{Z_k^n(0)} (\mathbf{1}(\tilde{\eta}_k^{i,Z} \leq t) - F_{k,e}(t)) \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{6.45}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{Q_k^n(0)}(\mathbf{1}(\tilde{w}_k^{n,i}+\eta_k^{i,Q}\le t)-F_k(t-\tilde{w}_k^{n,i}))\Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n\to\infty. \tag{6.46}$$

We first prove (6.45) holds. Let

$$\check{U}_k^n(t):=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}(\tilde{\eta}_k^{i,Z}\le t)-F_{k,e}(t)),\quad t\ge 0,$$

and $\check{U}_k^n:=\{\check{U}_k^n(t):t\ge 0\}$. By Lemma 3.1 of [29] and Assumption 6, we obtain

$$(\check{U}_k^n(t),\bar{Z}_k^n(0))\Rightarrow(\hat{B}_{0,k}(F_{k,e}(t)),0)\quad \text{in}\quad \mathbb{D}\times\mathbb{R}\quad \text{as}\quad n\to\infty.$$

Thus, the random time change theorem [7] implies (6.45) holds.

Before proving (6.46), we let $E_k^{n,Q}(t)$ be the cumulative number of initial jobs whose task k is in queue waiting for service at time $0-$, and has entered service by time $t\ge 0$, $k=1,2$. Set $E_k^{n,Q}:=\{E_k^{n,Q}(t):t\ge 0\}$ and $\bar{E}_k^{n,Q}:=E_k^{n,Q}/n$. Note that, for $k=1,2$, $0\le\bar{E}_k^{n,Q}(t)\le\bar{Q}_k^n(0)$ for $t\ge 0$, *a.s.* By Assumption 6 we easily see that $\bar{Q}_k^n(0)\Rightarrow 0$ in $\mathbb{R}$ as $n\to\infty$, which implies that for $k=1,2$,

$$\bar{E}_k^{n,Q}\Rightarrow 0\quad \text{in}\quad \mathbb{D}\quad \text{as}\quad n\to\infty.$$

For $k=1,2$, let

$$\tilde{V}_k^n(s,x):=\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor ns\rfloor}(\mathbf{1}(\eta_k^{i,Q}\le x)-F_k(x)),\quad s,x\ge 0.$$

We can rewrite the second term on the RHS of $\hat{V}_k^{n,0}$ in (6.20) by

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{Q_k^n(0)}(\mathbf{1}(\tilde{w}_k^{n,i}+\eta_k^{i,Q}\le t)-F_k(t-\tilde{w}_k^{n,i}))\\&=\int_0^t\int_0^t\mathbf{1}(s+x\le t)d\tilde{V}_k^n(\bar{E}_k^{n,Q}(s),x),\quad t\ge 0.\end{aligned}$$

From Proposition 5.1 in [51], we see that

$$\int_0^t\int_0^t\mathbf{1}(s+x\le t)d\tilde{V}_k^n(s,x)\Rightarrow\int_0^t\int_0^t\mathbf{1}(s+x\le t)d\tilde{V}_k(s,x)$$

in $\mathbb{D}$ as $n \to \infty$, where $\tilde{V}_k = \{\tilde{V}_k(t,x) : t \geq 0, x \geq 0\}$ is a standard Kiefer process, and the integral limit above is defined in the mean-square limit sense, similar to those in §6.2. Furthermore, analogous to Lemma 6.8, we obtain

$$\int_0^t \int_0^t \mathbf{1}(s+x \leq t) d\tilde{V}_k^n(\bar{E}_k^{n,Q}(s), x) \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty,$$

which implies (6.46) holds.

In the rest of the proof, we will show the convergence $\hat{V}^{n,0} \Rightarrow 0$. It suffices to show each term in (6.21) weakly converges to 0. Since the first two terms in (6.21) are symmetric, without loss of generality, we here only show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t - \tilde{w}_2^{n,i,R})) \Rightarrow 0$$

in $\mathbb{D}$ as $n \to \infty$. We prove the convergence by showing the upper and lower bounds in (6.47) and (6.48), respectively, converge to zero. Note that

$$\begin{aligned} (6.47) \quad & \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \left(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t - \tilde{w}_2^{n,i,R})\right) \\ & \geq \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \Big(\mathbf{1}(\tilde{w}_2^{n,i,R} + \tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) \\ & \qquad\qquad - F_{1,e}(t - \tilde{w}_2^{n,i,R}) F_2(t - \tilde{w}_2^{n,i,R})\Big) \\ & \quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (F_{1,e}(t - \tilde{w}_2^{n,i,R}) - F_{1,e}(t)), \quad t \geq 0, \quad a.s. \end{aligned}$$

We first show the RHS of (6.47) weakly converges to 0 as $n \to \infty$. By Theorem 11.4.5 in [58], it suffices to show the two terms on the RHS of (6.47) weakly converge to 0 as $n \to \infty$ separately. We first consider the first term on the RHS of (6.47).

Denote

$$\tilde{K}^n(t, \boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq x_1, \eta_2^{i,R} \leq x_2) - F_{1,e}(x_1) F_2(x_2)),$$

for $t \geq 0$ and $\boldsymbol{x} \in \mathbb{R}_+^2$, and $\tilde{K}^n := \{\tilde{K}^n(t, \boldsymbol{x}) : t \geq 0, \boldsymbol{x} \in \mathbb{R}_+^2\}$. Let $E_k^{n,Z}(t)$ be the cumulative number of initial jobs whose task k is in queue waiting for

service at time $0-$, and has entered service by time $t \geq 0$, but whose task k' is in service at time $0-$. Set $E_k^{n,Z} := \{E_k^{n,Z}(t) : t \geq 0\}$ and $\bar{E}_k^{n,Z} := E_k^{n,Z}/n$. Note that, for $k = 1, 2$, $0 \leq \bar{E}_k^{n,Z}(t) \leq \bar{Z}_{k'}^n(0)$ for $t \geq 0$, *a.s.* By Assumption 6 we easily see that $\bar{Z}_k^n(0) \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$, which implies that for $k = 1, 2$,

$$\bar{E}_k^{n,Z} \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty.$$

We rewrite the first term on the RHS of (6.47) by

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \Big(\mathbf{1}(\tilde{w}_2^{n,i,R} + \tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) \\
&\qquad\qquad - F_{1,e}(t - \tilde{w}_2^{n,i,R}) F_2(t - \tilde{w}_2^{n,i,R}) \Big) \\
&= \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s + x_j \leq t, \forall j) d\tilde{K}^n(\bar{E}_2^{n,Z}(s), \boldsymbol{x}).
\end{aligned}$$

By Theorem 3.3 in [33], we obtain that

$$\int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s + x_j \leq t, \forall j) d\tilde{K}^n(s, \boldsymbol{x}) \Rightarrow \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s + x_j \leq t, \forall j) d\tilde{K}(s, \boldsymbol{x})$$

in $\mathbb{D}$ as $n \to \infty$, where $\tilde{K} := \{\tilde{K}(t, \boldsymbol{x}) : t \geq 0, \boldsymbol{x} \in \mathbb{R}_+^2\}$ is a generalized Kiefer process with mean 0, and the covariance structure, for $s, t \in \mathbb{R}_+$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^2$,

$$\begin{aligned}
&Cov(\tilde{K}(s, \boldsymbol{x}), \tilde{K}(t, \boldsymbol{y})) \\
&= (s \wedge t) \left(F_{1,e}(x_1 \wedge y_1) F_2(x_2 \wedge y_2) - F_{1,e}(x_1) F_2(x_2) F_{1,e}(y_1) F_2(y_2) \right),
\end{aligned}$$

and the integral limit above is defined in the mean-square limit sense, similar to those in §6.2. Furthermore, analogous to Lemma 6.8, we obtain

$$\int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s + x_j \leq t, \forall j) d\tilde{K}^n(\bar{E}_2^{n,Z}(s), \boldsymbol{x}) \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty.$$

Thus, the first term on the RHS of (6.47) weakly converges to 0 as $n \to \infty$. The weak convergence of the second term on the RHS of (6.47) is similar to the proof of the convergence $\hat{M}^{n,0} \Rightarrow 0$ in Lemma 6.9, and we omit the details here for brevity.

On the other hand, we also have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \Big(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t - \tilde{w}_2^{n,i,R}) \Big)$$

$$\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \Big(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t) \Big)$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (F_2(t) - F_2(t - \tilde{w}_2^{n,i,R})), \quad t \geq 0, \quad a.s. \tag{6.48}$$

Following an analogous argument to show the terms on the RHS of (6.47) weakly converge to 0, we can also show the RHS of (6.48) weakly converges to 0.

Now, we are ready to prove the convergence $\hat{V}^{n,0} \Rightarrow 0$. It suffices to show that for $T > 0$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P\Bigg(\sup_{0\leq t\leq T} \Bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \big(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t - \tilde{w}_2^{n,i,R}) \big) \Bigg| > \epsilon \Bigg) = 0.$$

Note that

$$P\Bigg(\sup_{0\leq t\leq T} \Bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \big(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t - \tilde{w}_2^{n,i,R}) \big) \Bigg| > \epsilon \Bigg)$$

$$\leq P\Bigg(\sup_{0\leq t\leq T} \Bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \Big(\mathbf{1}(\tilde{\eta}_1^{i,Z} \leq t, \eta_2^{i,R} \leq t) - F_{1,e}(t) F_2(t) \Big) + \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (F_2(t) - F_2(t - \tilde{w}_2^{n,i,R})) \Bigg| > \epsilon \Bigg)$$

$$+ P\Bigg(\sup_{0\leq t\leq T} \Bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} \big(\mathbf{1}(\tilde{w}_2^{n,i,R} + \tilde{\eta}_1^{i,Z} \leq t, \tilde{w}_2^{n,i,R} + \eta_2^{i,R} \leq t) - F_{1,e}(t - \tilde{w}_2^{n,i,R}) F_2(t - \tilde{w}_2^{n,i,R}) \big) + \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_1^n(0)} (F_{1,e}(t - \tilde{w}_2^{n,i,R}) - F_{1,e}(t)) \Bigg| > \epsilon \Bigg).$$

By the fact that the terms on the RHS of (6.48) and (6.47) weakly converge to 0, we obtain that the RHS of the above inequality goes to 0 as $n \to \infty$.

Next, we will focus on proving that the third term in (6.21) weakly converges to 0 as $n \to \infty$. Let $E_k^{n,I}(t)$ be the cumulative number of initial jobs, whose task k has entered service by time $t \geq 0$, but whose both tasks are

in queue waiting for service at time 0−. Set $E_k^{n,I} := \{E_k^{n,I}(t) : t \geq 0\}$ and $\bar{E}_k^{n,I} := E_k^{n,I}/n$, $k = 1, 2$. Note that, for $k = 1, 2$, $0 \leq \bar{E}_k^{n,I}(t) \leq \bar{I}^n(0)$ for $t \geq 0$, *a.s.* Since by Assumption 6 $\bar{I}^n(0) \Rightarrow 0$ in $\mathbb{R}$ as $n \to \infty$, we have, for $k = 1, 2$,

$$\bar{E}_k^{n,I} \Rightarrow 0 \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty. \tag{6.49}$$

Let

$$\tilde{V}^n(t_1, t_2, \boldsymbol{x}) := \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt_1 \rfloor \wedge \lfloor nt_2 \rfloor} (\mathbf{1}(\eta_j^{i,I} \leq \boldsymbol{x}, \forall j) - F(\boldsymbol{x})),$$

for $t_1, t_2 \geq 0$, $\boldsymbol{x} \in \mathbb{R}_+^2$. We can rewrite the third term on the RHS of $\hat{V}^{n,0}$ in (6.21) by

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{I^n(0)} (\mathbf{1}(\tilde{w}_j^{n,i,I} + \eta_j^{i,I} \leq t, \forall j) - F(t - \tilde{w}_1^{n,i,I}, t - \tilde{w}_2^{n,i,I})) \\
&\qquad = \int_0^t \int_0^t \int_{\mathbb{R}_+^2} \mathbf{1}(s_j + x_j \leq t, \forall j) d\tilde{V}^n(\bar{E}_1^{n,I}(s_1), \bar{E}_2^{n,I}(s_2), \boldsymbol{x}), \quad t \geq 0.
\end{aligned}$$

Analogous to Lemma 6.8, together with (6.49), we can see that the term on the RHS of the previous equation weakly converges to 0 as $n \to \infty$, which further implies the weak convergence of the third term on the RHS of $\hat{V}^{n,0}$ in (6.21) to 0. Thus, we have completed the proof of Lemma 6.10. □

LEMMA 6.11. *Under Assumptions 1 and 4-8, we have*

$$\begin{aligned}
&\big(\hat{U}_1^{n,Y}(t), \hat{U}_2^{n,Y}(t), \hat{V}_1^{n,0}(t), \hat{V}_2^{n,0}(t), \hat{V}^{n,0}(t), \hat{M}^{n,0}(t), \hat{U}^n(\boldsymbol{t})\big) \\
&\Rightarrow \big(\bar{Y}_2(0)^{1/2} \hat{B}_{0,1}(F_{1,e}(t)), \bar{Y}_1(0)^{1/2} \hat{B}_{0,2}(F_{2,e}(t)), 0, 0, 0, 0, \bar{J}(0)^{1/2} \hat{U}(\boldsymbol{t})\big)
\end{aligned}$$

in $\mathbb{D}^6 \times \mathbb{D}([0, \infty)^2, \mathbb{R})$ *as* $n \to \infty$, *where the processes* $\hat{B}_{0,k} := \{\hat{B}_{0,k}(t) : t \geq 0\}$, $k = 1, 2$, *and the process* $\hat{U} := \{\hat{U}(\boldsymbol{t}) : \boldsymbol{t} \in \mathbb{R}_+^2\}$ *are defined in Theorem 4.1.*

PROOF. For $k = 1, 2$, let

$$\begin{aligned}
\tilde{U}_k^{n,Y}(t) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor n\bar{Y}_{k'}(0) \rfloor} (\mathbf{1}(\tilde{\eta}_k^{i,Y_k} \leq t) - F_{k,e}(t)), \\
\tilde{U}^n(\boldsymbol{t}) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor n\bar{J}(0) \rfloor} (\mathbf{1}(\tilde{\boldsymbol{\eta}}^{i,J} \leq \boldsymbol{t}) - F_{1,e}(t_1) F_{2,e}(t_2)).
\end{aligned} \tag{6.50}$$

Set $\tilde{U}_k^{n,Y} := \{\tilde{U}_k^{n,Y}(t) : t \geq 0\}$, $k = 1, 2$, and $\tilde{U}^n := \{\tilde{U}^n(\boldsymbol{t}) : \boldsymbol{t} \in \mathbb{R}_+^2\}$.

By Lemma 3.1 of [29] and Assumption 6, with the random time change theorem [7], since $\tilde{U}_k^{n,Y}$, $k = 1, 2$, and $\tilde{U}^n$ are independent by definition, we first obtain

$$\begin{aligned}\hat{U}_k^{n,Y}(t) &\Rightarrow \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t)) \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty,\\ \hat{U}^n &\Rightarrow \bar{J}(0)^{1/2}\hat{U} \quad \text{in} \quad \mathbb{D}([0,\infty)^2, \mathbb{R}) \quad \text{as} \quad n \to \infty,\end{aligned}$$

and

$$\begin{aligned}&\big(\tilde{U}_1^{n,Y}(t), \tilde{U}_2^{n,Y}(t), \tilde{U}^n(\boldsymbol{t})\big)\\ &\Rightarrow \big(\bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)), \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), \bar{J}(0)^{1/2}\hat{U}(\boldsymbol{t})\big)\end{aligned}$$

in $\mathbb{D}^2 \times \mathbb{D}([0,\infty)^2, \mathbb{R})$ as $n \to \infty$. In order to prove the convergence of $\tilde{U}_k^{n,Y}$, $\tilde{U}^n$, joint with $\hat{U}_k^{n,Y}$, $\hat{U}^n$, $k = 1, 2$, it suffices to check, for each $T > 0$ and $\gamma > 0$,

$$\begin{aligned}\lim_{n\to\infty} P\bigg(\sup_{0\leq t\leq T} |\tilde{U}_k^{n,Y}(t) - \hat{U}_k^{n,Y}(t)| > \gamma\bigg) &= 0, \quad k = 1, 2,\\ \lim_{n\to\infty} P\bigg(\sup_{0\leq t_1,t_2\leq T} |\tilde{U}^n(\boldsymbol{t}) - \hat{U}^n(\boldsymbol{t})| > \gamma\bigg) &= 0.\end{aligned}$$

The above equations are evident as $\hat{B}_{0,k}(F_{k,e}(\cdot))$ and $\hat{U}$ are all continuous, $k = 1, 2$. Thus, we have

$$\begin{aligned}&\big(\tilde{U}_1^{n,Y}(t), \tilde{U}_2^{n,Y}(t), \hat{U}_1^{n,Y}(t), \hat{U}_2^{n,Y}(t), \tilde{U}^n(\boldsymbol{t}), \hat{U}^n(\boldsymbol{t})\big)\\ &\Rightarrow \big(\bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)), \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), \bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)),\\ &\qquad \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), \bar{J}(0)^{1/2}\hat{U}(\boldsymbol{t}), \bar{J}(0)^{1/2}\hat{U}(\boldsymbol{t})\big)\end{aligned}$$

in $\mathbb{D}^4 \times \mathbb{D}^2([0,\infty)^2, \mathbb{R})$ as $n \to \infty$. Further, by Lemma 6.10, together with Theorem 11.4.5 of [58], we obtain

$$\begin{aligned}&\big(\tilde{U}_1^{n,Y}(t), \tilde{U}_2^{n,Y}(t), \hat{U}_1^{n,Y}(t), \hat{U}_2^{n,Y}(t), \hat{V}_1^{n,0}(t), \hat{V}_2^{n,0}(t), \hat{V}^{n,0}(t),\\ &\qquad \hat{M}^{n,0}(t), \tilde{U}^n(\boldsymbol{t}), \hat{U}^n(\boldsymbol{t})\big)\\ &\Rightarrow \big(\bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)), \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), \bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)),\\ &\qquad \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), 0, 0, 0, 0, \bar{J}(0)^{1/2}\hat{U}(\boldsymbol{t}), \bar{J}(0)^{1/2}\hat{U}(\boldsymbol{t})\big)\end{aligned}$$

in $\mathbb{D}^8 \times \mathbb{D}^2([0,\infty)^2, \mathbb{R})$ as $n \to \infty$, which completes the proof of Lemma 6.11. □

We can now conclude the weak convergence of the processes associated with the initial quantities $\hat{\boldsymbol{X}}^{n,0} := (\hat{X}_1^{n,0}, \hat{X}_2^{n,0})$, $\hat{\boldsymbol{Y}}^{n,0} := (\hat{Y}_1^{n,0}, \hat{Y}_2^{n,0})$ and $\hat{S}^{n,0}$ in (4.27)–(4.29).

LEMMA 6.12. *Under Assumptions 1 and 4-8,*

$$\left(\hat{\boldsymbol{X}}^{n,0}, \hat{\boldsymbol{Y}}^{n,0}, \hat{S}^{n,0}\right) \Rightarrow \left(\hat{\boldsymbol{X}}^{0}, \hat{\boldsymbol{Y}}^{0}, \hat{S}^{0}\right) \quad \text{in} \quad \mathbb{D}^5 \quad \text{as} \quad n \to \infty,$$

where $\hat{\boldsymbol{X}}^0 := (\hat{X}_1^0, \hat{X}_2^0)$, $\hat{\boldsymbol{Y}}^0 := (\hat{Y}_1^0, \hat{Y}_2^0)$ *and* $\hat{S}^0$ *are defined in Theorem 4.1.*

6.6. *Completing the Proof of Theorem 4.1.* In this section, we complete the proof of Theorem 4.1. We first provide the following lemmas for the proof.

LEMMA 6.13. *Under Assumptions 1 and 4-8,*

$$\begin{aligned}
&\Big(\hat{A}^n(t), \hat{\boldsymbol{X}}^{n,0}(t), \hat{\boldsymbol{Y}}^{n,0}(t), \hat{S}^{n,0}(t), \hat{V}^{n,0}(t), \hat{M}^{n,0}(t), \hat{U}^n(t), \hat{W}^n(t),\\
&\qquad N_k\sqrt{n}(1-\rho_k^n)F_{k,e}(t), \hat{V}_k^{n,0}(t), \hat{U}_k^{n,Y}(t), \hat{U}_k^n(t), \hat{W}_k^n(t), \hat{W}_k^{n,c}(t),\ k=1,2\Big)\\
&\Rightarrow \Big(\hat{A}(t), \hat{\boldsymbol{X}}^0(t), \hat{\boldsymbol{Y}}^0(t), \hat{S}^0(t), 0, 0, \bar{J}(0)^{1/2}\hat{U}(t), \hat{W}(t), N_k\beta_k F_{k,e}(t), 0,\\
&\qquad \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t)), \bar{J}(0)^{1/2}\hat{U}_k(t), \hat{W}_k(t), \hat{W}_k^c(t),\ k=1,2\Big)
\end{aligned}$$

in $\mathbb{D}^{22}$ *as* $n \to \infty$, *where all the limiting processes are defined in Theorem 4.1.*

PROOF. Let

$$\tilde{U}_1^n(t) := \tilde{U}^n(t,\infty), \quad \tilde{U}_2^n(t) := \tilde{U}^n(\infty, t), \quad t \geq 0,$$

where $\tilde{U}^n$ is defined in (6.50). Set $\tilde{U}_1^n := \{\tilde{U}_1^n(t) : t \in \mathbb{R}_+\}$ and $\tilde{U}_2^n := \{\tilde{U}_2^n(t) : t \in \mathbb{R}_+\}$. Without abuse of notation, we let $\tilde{U}^n(t) = \tilde{U}^n(t,t)$, $t \geq 0$. Recall $\tilde{U}_k^{n,Y}$ is defined in (6.50), $k = 1, 2$.

First, by Lemmas 6.11, 6.12 and 6.8, we obtain the convergence

$$\begin{aligned}
&\Big(\hat{A}^n(t), \hat{\boldsymbol{X}}^{n,0}(t), \hat{\boldsymbol{Y}}^{n,0}(t), \hat{S}^{n,0}(t), \tilde{U}^n(t), \tilde{W}^n(t), N_k\sqrt{n}(1-\rho_k^n)F_{k,e}(t),\\
&\qquad \tilde{U}_k^{n,Y}(t), \tilde{U}_k^n(t), \tilde{W}_k^n(t), \tilde{W}_k^{n,c}(t), \quad k=1,2\Big)\\
&\Rightarrow \Big(\hat{A}(t), \hat{\boldsymbol{X}}^0(t), \hat{\boldsymbol{Y}}^0(t), \hat{S}^0(t), \bar{J}(0)^{1/2}\hat{U}(t), \hat{W}(t), N_k\beta_k F_{k,e}(t),\\
&\qquad \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t)), \bar{J}(0)^{1/2}\hat{U}_k(t), \hat{W}_k(t), \hat{W}_k^c(t), \quad k=1,2\Big)
\end{aligned}$$

in $\mathbb{D}^{18}$ as $n \to \infty$, since $\big(\hat{\boldsymbol{X}}^{n,0}(t), \hat{\boldsymbol{Y}}^{n,0}(t), \hat{S}^{n,0}(t)\big)$ and the other component processes in the prelimit above are independent of each other.

Then, by (6.40) and the maximum topology we endow on the product space, we obtain that in the above weak convergence $\tilde{W}^n$, $\tilde{U}^n(\cdot)$, $\tilde{U}_k^{n,Y}$, $\tilde{U}_k^n$, $\tilde{W}_k^{n,c}$ and $\tilde{W}_k^n$, $k = 1, 2$, can be replaced by $\hat{W}^n$, $\hat{U}^n(\cdot)$, $\hat{U}_k^{n,Y}$, $\hat{U}_k^n$, $\hat{W}_k^{n,c}$ and $\hat{W}_k^n$, $k = 1, 2$. Recall from Lemma 6.11 that $\hat{V}_k^{n,0}$, $k = 1, 2$, $\hat{V}^{n,0}$ and $\hat{M}^{n,0}$ weakly converge to 0 as $n \to \infty$. Following from Theorem 11.4.5 of [58], we have completed the proof of Lemma 6.13. □

Let Γ be a continuous distribution function on $\mathbb{R}_+$ and let $a \in \mathbb{R}$. For each $x \in \mathbb{D}$, we define the mapping $\phi_\Gamma^a : \mathbb{D} \to \mathbb{D}$ by $\phi_\Gamma^a(x) = z$ for $x \in \mathbb{D}$, where $z \in \mathbb{D}$ is a solution to the following

$$z(t) = x(t) + \int_0^t (z(t-s) + a)^+ d\Gamma(s), \quad t \geq 0. \tag{6.51}$$

The existence and uniqueness of the solution to (6.51) are proved in Proposition 3.1 of [51]. We also define the mapping $\psi : \mathbb{D}^3 \to \mathbb{D}^3$ by

$$\psi(x_1, x_2, x_3) := (\phi_{F_1}^0(x_1), \phi_{F_2}^0(x_2), x_3), \quad (x_1, x_2, x_3) \in \mathbb{D}^3. \tag{6.52}$$

Recall that we endow the product metric space with the maximum metric of each component metric space. Since the mappings $\phi_{F_1}^0$ and $\phi_{F_2}^0$ are both continuous in $\mathbb{D}$, we immediately have the following.

LEMMA 6.14. *The mapping ψ defined in (6.52) is continuous in $(\mathbb{D}^3, J_1)$.*

Recall the definition of $\hat{\mathcal{E}}^n(\cdot, \cdot)$ in (4.12). We then give a representation for $\hat{\Psi}^n := \{\hat{\Psi}^n(t) : t \geq 0\}$ where $\hat{\Psi}^n(t)$, $t \geq 0$, is defined in (6.18).

LEMMA 6.15. *The process $\{\hat{\Psi}^n(t) : t \geq 0\}$ has the following representation:*

$$\hat{\Psi}^n(t) = \int_0^t \int_0^t \hat{\mathcal{E}}^n(t - s_1, t - s_2) dF(s_1, s_2), \quad t \geq 0.$$

PROOF. Note that by the definition of E_k^n, $k = 1, 2$, for $t \geq 0$,

$$\hat{\Psi}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \int_0^t \int_0^t \mathbf{1}(s_j \leq t - \tau_i^n - w_j^{n,i}, \forall j) dF(s_1, s_2) - \frac{\lambda^n}{\sqrt{n}} \int_0^t \int_0^t \Big(\min_{k=1,2} \{t - s_k\} \Big) dF(s_1, s_2)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{A^n(t)} \int_0^t \int_0^t \mathbf{1}(\tau_i^n + w_j^{n,i} \le t - s_j, \forall j) dF(s_1, s_2)$$
$$- \frac{\lambda^n}{\sqrt{n}} \int_0^t \int_0^t \Big(\min_{k=1,2} \{t - s_k\} \Big) dF(s_1, s_2)$$
$$= \frac{1}{\sqrt{n}} \int_0^t \int_0^t \Big(\min_{k=1,2} \{E_k^n(t - s_k)\} \Big) dF(s_1, s_2)$$
$$- \frac{\lambda^n}{\sqrt{n}} \int_0^t \int_0^t \Big(\min_{k=1,2} \{t - s_k\} \Big) dF(s_1, s_2)$$
$$= \int_0^t \int_0^t \hat{\mathcal{E}}^n(t - s_1, t - s_2) dF(s_1, s_2). \qquad \square$$

PROOF OF THEOREM 4.1. First, by Lemma 6.13 and the continuous mapping theorem [7], we have

$$\big(\hat{A}^n(t), \hat{X}_k^{n,0}(t) - N_k\sqrt{n}(1 - \rho_k^n)F_{k,e}(t) - \hat{V}_k^{n,0}(t) - \hat{U}_k^{n,Y}(t)$$
$$- \hat{U}_k^n(t) - \hat{W}_k^n(t), \quad k = 1, 2\big)$$
$$\Rightarrow \big(\hat{A}(t), \hat{X}_k^0(t) - N_k\beta_k F_{k,e}(t) - \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t))$$
$$- \bar{J}(0)^{1/2}\hat{U}_k(t) - \hat{W}_k(t), \quad k = 1, 2\big)$$

in $\mathbb{D}^3$ as $n \to \infty$. Define the mapping $g : \mathbb{D}^3 \to \mathbb{D}^5$ by

$$g(x_1, x_2, x_3) := (g_1(x_1), g_2(x_2), g_3(x_1), g_4(x_3), g_5(x_1)),$$

for $(x_1, x_2, x_3) \in \mathbb{D}^3$, where $g_1(x_1) := x_1$, $g_2(x_2) := x_2$, $g_4(x_3) := x_3$,

$$g_3(x_1)(\cdot) := \int_0^\cdot F_1^c(\cdot - s) dx_1(s), \text{ and } \quad g_5(x_1)(\cdot) := \int_0^\cdot F_2^c(\cdot - s) dx_1(s).$$

By Lemma A.9 in [51] and the metric we endow on the product metric space, it is easy to see the mapping g is continuous. Thus, by the continuous mapping theorem [7], we see that

$$\Big(\hat{A}^n(t), \hat{X}_k^{n,0}(t) - N_k\sqrt{n}(1 - \rho_k^n)F_{k,e}(t) - \hat{V}_k^{n,0}(t) - \hat{U}_k^{n,Y}(t) - \hat{U}_k^n(t) - \hat{W}_k^n(t),$$
$$\int_0^t F_k^c(t - s) d\hat{A}^n(s), \quad k = 1, 2\Big)$$
$$\Rightarrow \Big(\hat{A}(t), \hat{X}_k^0(t) - N_k\beta_k F_{k,e}(t) - \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}^c(t)) - \bar{J}(0)^{1/2}\hat{U}_k(t) - \hat{W}_k(t),$$
$$\int_0^t F_k^c(t - s) d\hat{A}(s), \quad k = 1, 2\Big)$$

in $\mathbb{D}^5$ as $n \to \infty$, which by the continuous mapping theorem [7] again implies that

$$\begin{aligned}&\Big(\hat{A}^n(t), \hat{X}_k^{n,0}(t) - N_k\sqrt{n}(1-\rho_k^n)F_{k,e}(t) - \hat{V}_k^{n,0}(t) - \hat{U}_k^{n,Y}(t) - \hat{U}_k^n(t)\\ &\qquad - \hat{W}_k^n(t) + \int_0^t F_k^c(t-s)d\hat{A}^n(s), \quad k=1,2\Big)\\ \Rightarrow &\Big(\hat{A}(t), \hat{X}_k^0(t) - N_k\beta_k F_{k,e}(t) - \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(t)) - \bar{J}(0)^{1/2}\hat{U}_k(t)\\ &\qquad - \hat{W}_k(t) + \int_0^t F_k^c(t-s)d\hat{A}(s), \quad k=1,2\Big)\end{aligned}$$

in $\mathbb{D}^3$ as $n \to \infty$. Thus, by applying the continuous mapping theorem for the mapping ψ defined in (6.52), we obtain

$$\big(\hat{A}^n, \hat{X}_1^n, \hat{X}_2^n\big) \Rightarrow \big(\hat{A}, \hat{X}_1, \hat{X}_2\big) \quad \text{in} \quad \mathbb{D}^3 \quad \text{as} \quad n \to \infty, \tag{6.53}$$

where, for $k = 1, 2$,

$$\begin{aligned}\hat{X}_k(\cdot) = \phi_{F_k}^0\Big(&\hat{X}_k^0(\cdot) - N_k\beta_k F_{k,e}(\cdot) - \bar{J}(0)^{1/2}\hat{U}_k(\cdot) - \bar{Y}_{k'}(0)^{1/2}\hat{B}_{0,k}(F_{k,e}(\cdot))\\ &- \hat{W}_k(\cdot) + \int_0^{\cdot} F_k^c(\cdot - s)d\hat{A}(s)\Big),\end{aligned}$$

which implies $\hat{X}_k$ is the unique solution to (4.23) by the definition of $\phi_{F_k}^0$ in (6.51).

Now, by the definition of E_k^n, $k = 1, 2$, we have the balanced equation: for $t > 0$,

$$E_k^n(t) + \sum_{i=1}^{Q_k^n(0)} \mathbf{1}(\tilde{w}_k^{n,i} \le t) = (X_k^n(0) - N_k^n)^+ + A^n(t) - (X_k^n(t) - N_k^n)^+.$$

Recall the definition of $\hat{\mathcal{E}}_k^{n,e}$ in (6.41). We further have

$$\hat{E}_k^n(t) = \hat{A}^n(t) + \hat{\mathcal{E}}_k^{n,e}(t) - (\hat{X}_k^n(t))^+,$$

where the processes $\hat{E}_k^n$, $k = 1, 2$,are defined in (4.11). By Lemmas 6.13 and 6.9, Theorem 11.4.5 of [58] and (6.53), we obtain

$$\big(\hat{\boldsymbol{X}}^n(0), \hat{A}^n, \hat{X}_1^n, \hat{X}_2^n, \hat{\mathcal{E}}_1^{n,e}, \hat{\mathcal{E}}_2^{n,e}\big) \Rightarrow \big(\hat{\boldsymbol{X}}(0), \hat{A}, \hat{X}_1, \hat{X}_2, 0, 0\big)$$

in $\mathbb{R}^2 \times \mathbb{D}^5$ as $n \to \infty$. Together with the continuous mapping theorem [7], we immediately obtain

$$(\hat{E}_1^n, \hat{E}_2^n) \Rightarrow (\hat{E}_1, \hat{E}_2) \quad \text{in} \quad \mathbb{D}^2 \quad \text{as} \quad n \to \infty, \tag{6.54}$$

where the processes $\hat{E}_k$, $k = 1, 2$, are defined in (4.26).

By the definition of $\hat{\mathcal{E}}^n(\cdot, \cdot)$ in (4.12) and the joint weak convergence of $(\hat{E}_1^n, \hat{E}_2^n)$, we easily obtain the weak convergence of $\hat{\mathcal{E}}^n(t_1, t_2)$ for each fixed $t_1, t_2 \geq 0$:

$$\hat{\mathcal{E}}^n(t_1, t_2) \Rightarrow \hat{\mathcal{E}}(t_1, t_2) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty,$$

where $\hat{\mathcal{E}}(\cdot, \cdot)$ is defined in (4.31).

Before we establish the weak convergence of $\hat{S}^n$, we first show the convergence of $\hat{\Psi}^n$:

$$\hat{\Psi}^n \Rightarrow \hat{\Psi} \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty. \tag{6.55}$$

We first note that by the continuity of F and the definition of $\hat{\Psi}$ in (4.30), $\hat{\Psi}$ has continuous sample paths. In order to prove (6.55), it suffices to show that for any $T > 0$ and $\epsilon > 0$,

$$\lim_{n \to \infty} P\Big(\sup_{0 \leq t \leq T} |\hat{\Psi}^n(t) - \hat{\Psi}(t)| > \epsilon \Big) = 0.$$

Note that, by Lemma 6.15,

$$\begin{aligned}
&P\Big(\sup_{0 \leq t \leq T} |\hat{\Psi}^n(t) - \hat{\Psi}(t)| > \epsilon \Big) \\
&\quad = P\Big(\sup_{0 \leq t \leq T} \Big| \int_0^t \int_0^t [\hat{\mathcal{E}}^n(t - s_1, t - s_2) - \hat{\mathcal{E}}(t - s_1, t - s_2)] dF(s_1, s_2) \Big| > \epsilon \Big) \\
&\quad \leq P\Big(F_{\mathrm{m}}(T) \sup_{0 \leq s_1, s_2 \leq T} |\hat{\mathcal{E}}^n(s_1, s_2) - \hat{\mathcal{E}}(s_1, s_2)| > \epsilon \Big) \\
&\quad \leq P\Big(\sup_{0 \leq s_1, s_2 \leq T} |\hat{\mathcal{E}}^n(s_1, s_2) - \hat{\mathcal{E}}(s_1, s_2)| > \epsilon \Big) \\
&\quad \leq P\Big(\sup_{0 \leq s_1 \leq s_2 \leq T} |\hat{\mathcal{E}}^n(s_1, s_2) - \hat{\mathcal{E}}(s_1, s_2)| > \epsilon/2 \Big) \\
&\qquad\qquad + P\Big(\sup_{0 \leq s_2 \leq s_1 \leq T} |\hat{\mathcal{E}}^n(s_1, s_2) - \hat{\mathcal{E}}(s_1, s_2)| > \epsilon/2 \Big).
\end{aligned} \tag{6.56}$$

In the rest of the proof, we only focus on the first term on the RHS of (6.56), as the way to deal with the second term is similar. By the definition

of $\hat{\mathcal{E}}^n(\cdot,\cdot)$ in (4.12) and $\hat{\mathcal{E}}(\cdot,\cdot)$ in (4.31), we present an upper bound for the first term

$$P\Big(\sup_{0\le s_1\le s_2\le T}|\hat{\mathcal{E}}^n(s_1,s_2)-\hat{\mathcal{E}}(s_1,s_2)|>\epsilon/2\Big)\le\alpha_1^n+\alpha_2^n,$$

where α_1^n and α_2^n are defined as follows:

$$\alpha_1^n:=P\Big(\sup_{0\le s_1\le s_2\le T}\Big|(\hat{E}_1^n(s_1)-\hat{E}_1(s_1))\mathbf{1}(s_1<s_2)+\Big(\min_{k=1,2}\hat{E}_k^n(s_k)-\min_{k=1,2}\hat{E}_k(s_k)\Big)\mathbf{1}(s_1=s_2)\Big|>\epsilon/2\Big),$$

$$\alpha_2^n:=P\Big(\hat{\mathcal{E}}^n(s_1,s_2)=\hat{E}_1^n(s_1)\mathbf{1}(s_1<s_2)+(\hat{E}_1^n(s_1)\wedge\hat{E}_2^n(s_2))\mathbf{1}(s_1=s_2),\ 0\le s_1\le s_2\le T\Big).$$

By (6.54) and the definition of $\hat{\mathcal{E}}^n(\cdot,\cdot)$ in (4.12), we immediately see that $\alpha_2^n\to 0$ as $n\to\infty$. By (5.30), we obtain that

$$\begin{aligned}\alpha_1^n&=P\Big(\sup_{0\le s_1,s_2\le T}\Big|(\hat{E}_1^n(s_1)-\hat{E}_1(s_1))\mathbf{1}(s_1<s_2)\\&\qquad+\frac{1}{2}\mathbf{1}(s_1=s_2)\Big[\sum_{k=1}^{2}(\hat{E}_k^n(s_k)-\hat{E}_k(s_k))-|\hat{E}_1^n(s_1)-\hat{E}_2^n(s_2)|\\&\qquad\qquad+|\hat{E}_1(s_1)-\hat{E}_2(s_2)|\Big]\Big|>\epsilon/2\Big)\\&\le P\Big(4\sup_{0\le s\le T}|\hat{E}_1^n(s)-\hat{E}_1(s)|+2\sup_{0\le s\le T}|\hat{E}_2^n(s)-\hat{E}_2(s)|>\epsilon\Big).\end{aligned}$$

By (6.54) and the fact that $\hat{E}_k$ has continuous sample paths, $k=1,2$, we obtain that for the fixed ϵ,

$$P\Big(\sup_{0\le s\le T}|\hat{E}_k^n(s)-\hat{E}_k(s)|>\epsilon/8\Big)=0,\quad k=1,2,$$

which implies that $\alpha_1^n\to 0$ as $n\to\infty$. Therefore, we have shown (6.55) holds.

Now, we are ready to prove the weak convergence of $\hat{S}^n$. By (6.55), (6.53), (6.54) and Lemma 6.13, we have

$$\begin{aligned}&\Big(\hat{\boldsymbol{X}}^n(0), \hat{\boldsymbol{Y}}^n(0), \hat{A}^n(t), \hat{X}_1^n(t), \hat{X}_2^n(t), \hat{U}_1^{n,Y}(t), \hat{U}_2^{n,Y}(t),\\ &\qquad \hat{U}^n(t), \hat{V}^{n,0}(t), \hat{M}^{n,0}(t), \hat{W}^n(t), \hat{\Psi}^n(t), \hat{E}_1^n(t), \hat{E}_2^n(t)\Big)\\ &\Rightarrow \Big(\hat{\boldsymbol{X}}(0), \hat{\boldsymbol{Y}}(0), \hat{A}(t), \hat{X}_1(t), \hat{X}_2(t), \bar{Y}_2(0)^{1/2}\hat{B}_{0,1}(F_{1,e}(t)),\\ &\qquad \bar{Y}_1(0)^{1/2}\hat{B}_{0,2}(F_{2,e}(t)), \bar{J}(0)^{1/2}\hat{U}(t), 0, 0, \hat{W}(t), \hat{\Psi}(t), \hat{E}_1(t), \hat{E}_2(t)\Big)\end{aligned}$$

in $\mathbb{R}^4 \times \mathbb{D}^{12}$ as $n \to \infty$. By the representation of $\hat{S}^n$ in (6.11) and the continuous mapping theorem [7], we immediately see

$$\begin{aligned}&\big(\hat{\boldsymbol{X}}^n(0), \hat{\boldsymbol{Y}}^n(0), \hat{A}^n, \hat{X}_1^n, \hat{X}_2^n, \hat{S}^n, \hat{E}_1^n, \hat{E}_2^n\big)\\ &\qquad \Rightarrow \big(\hat{\boldsymbol{X}}(0), \hat{\boldsymbol{Y}}(0), \hat{A}, \hat{X}_1, \hat{X}_2, \hat{S}, \hat{E}_1, \hat{E}_2\big)\end{aligned} \tag{6.57}$$

in $\mathbb{R}^4 \times \mathbb{D}^6$ as $n \to \infty$, where the process $\hat{S}$ is defined in (4.25).

Recall the representations of $\hat{Q}_k^n$, $\hat{B}_k^n$ and $\hat{D}_k^n$ in (4.11) and $\hat{Y}_k^n$ in (4.10), $k = 1, 2$. The continuous mapping theorem [7] and (6.57) imply (4.22) holds. The uniqueness of these processes follows from the uniqueness of $\hat{X}_k$, $k = 1, 2$. □

7. Concluding remarks. In this paper we have developed a methodology to study the multi-server fork-join networks with the NES constraints in the Halfin–Whitt regime. The fluid limits are proved for the networks with an empty initial condition, in which each job is split into $K \geq 2$ parallel tasks, and the arrival rate can be time-varying. In the diffusion scale, we have restricted to the networks with a stationary initial condition, in which each job is split into $K = 2$ parallel tasks, and the arrival rate is constant. It is clear from the analysis that arbitrary initial conditions cause substantial difficulties even in order to provide a concise representation of the system dynamics when $K > 2$. We have generalized the methodology in [51] for $G/GI/N$ queues to the fork-join networks in the Halfin–Whitt regime with $K = 2$. In this framework, we require that the system starts from stationarity at time 0. Notably in [50], for the $G/GI/N$ queues, the initial conditions have been relaxed to be arbitrary, but only the finite-dimensional distribution convergence is proved. On the other hand, for $G_t/GI/N$ queues with any arbitrary initial conditions, Kaspi and Ramanan [24, 25] have established the FLLN and FCLT for the measure-valued processes that keep track of the amount of service each job has received in the many-server heavy-traffic

regimes. In future work it may be worth investigating if the measure-valued processes approach can be used or developed to study multi-server fork-join networks. In particular, it will be interesting to establish an FCLT when $K > 2$ and the system starts from empty at time 0.

APPENDIX: PROOF OF LEMMA 5.5

We first prove the martingale property of $H^{n,i}$. By the definition of $H^{n,i}$ in (6.33) and the construction of the filtration $\mathcal{H}^n$ in (5.15), $H^{n,i}$ is $\mathcal{H}^n$-adapted. Note that, for each $t \geq 0$,

$$|H^{n,i}(t)| \leq 1 + \int_0^{\eta_1^i} \int_0^{\eta_2^i} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}), \quad a.s.$$

By Lemma 4.3 in [33], we have $E[|H^{n,i}(t)|] < \infty$, for $t \geq 0$. We next show the martingale property of $H^{n,i}$, i.e., for $s < t$,

$$E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] = H^{n,i}(s).$$

It suffices to show, for $s < t$,

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] = 0, \tag{A.1}$$

and

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} \leq s) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] = H^{n,i}(s). \tag{A.2}$$

We first prove (A.1). By the construction of $\mathcal{H}^n$ in (5.15), $\hat{\tau}_j^{n,i}$ is an $\mathcal{H}^n$-stopping time, $j = 1, 2$, which implies $\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}$ is also an $\mathcal{H}^n$-stopping time. Thus, the σ-field $\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}$ is well-defined. Hence,

$$\begin{aligned}
&\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&= \mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s) E\left[E\left[H^{n,i}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right]|\mathcal{H}_s^n\right].
\end{aligned}$$

Then, we claim that

$$E\left[H^{n,i}(t)|\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}\right] = \frac{E\left[H^{n,i}(t)|\hat{\tau}_j^{n,i}, \forall j\right]}{P(\boldsymbol{\eta}^i > \mathbf{0}|\hat{\tau}_j^{n,i}, \forall j)} = 0, \tag{A.3}$$

where the last equality follows from (5.12) and the independence of $\boldsymbol{\eta}^i$ and $\hat{\tau}_j^{n,i}$, $j = 1, 2$. In order to prove the first equality in (A.3), by Lemma 3.6 in [29], we only need to show

$$\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}} \cap \{\boldsymbol{\eta}^i > \mathbf{0}\} \tag{A.4}$$

$$\subset \Big(\sigma(\boldsymbol{\eta}^r, r \geq 1, r \neq i) \vee \sigma(\hat{\tau}_j^{n,i}, j = 1,2) \vee \sigma(\xi_p^n, p \geq 1) \vee \mathcal{N}\Big) \cap \{\boldsymbol{\eta}^i > \mathbf{0}\}.$$

It is enough to check (A.4) for sets which generate $\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}}$. By the definition of $\mathcal{H}_t^n$, $t \geq 0$, in (5.15), we note that (use, e.g., the argument in Appendix A.2 of Brémaud [9])

$$\mathcal{H}^n_{\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}} = \sigma\Big(\hat{\tau}_l^{n,r}, l = 1,2, \mathbf{1}(\eta_j^r \leq s \wedge (\hat{\tau}_1^{n,r} \vee \hat{\tau}_2^{n,r}) - \hat{\tau}_j^{n,r}, \forall j), \\ s \geq 0, r = 1, ..., E_1^n(\hat{\tau}_1^{n,i}) \wedge E_2^n(\hat{\tau}_2^{n,i})\Big) \vee \sigma(\xi_r^n, r \geq 1) \vee \mathcal{N},$$

where $\mathcal{N}$ includes all the null sets. Then, for $l = i, i+1, ...$, $p = 1, 2, ...$, $s_1, s_2, ..., s_l > 0$ and Borel sets $B_1, .., B_p$, $C_1^1, ..., C_l^1$, $C_1^2, ..., C_l^2$ and $G_1, ..., G_l$, since $E_1^n(\hat{\tau}_1^{n,i}) \wedge E_2^n(\hat{\tau}_2^{n,i}) \geq l > i$, then $\hat{\tau}_j^{n,r} = \hat{\tau}_j^{n,i}, r = i+1, ..., l$, $j = 1, 2$, we have that

$$\Big(\bigcap_{r=1}^p \{\xi_r^n \in B_r\}\Big) \cap \Big\{E_1^n(\hat{\tau}_1^{n,i}) \wedge E_2^n(\hat{\tau}_2^{n,i}) \geq l\Big\} \cap \Big(\bigcap_{r=1}^l \Big\{\hat{\tau}_j^{n,r} \in C_r^j, j = 1,2\Big\}\Big) \\ \cap \Big(\bigcap_{r=1}^l \mathbf{1}\Big(\eta_j^r \leq s_r \wedge (\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}) - \hat{\tau}_j^{n,r}, j = 1,2\Big) \in G_r\Big) \cap \{\boldsymbol{\eta}^i > \mathbf{0}\} \\ = \Big(\bigcap_{r=1}^p \{\xi_r^n \in B_r\}\Big) \cap \Big(\bigcap_{r=i+1}^l \Big\{\hat{\tau}_j^{n,i} = \hat{\tau}_j^{n,r}, j = 1,2\Big\}\Big) \\ \cap \Big(\bigcap_{r=1}^{i-1} \Big\{\hat{\tau}_j^{n,r} \in C_r^j, j = 1,2\Big\}\Big) \cap \Big(\bigcap_{r=i}^l \Big\{\hat{\tau}_j^{n,r} \in C_r^j, j = 1,2\Big\}\Big) \\ \cap \Big(\bigcap_{r=1}^{i-1} \mathbf{1}\Big(\eta_j^r \leq s_r \wedge (\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i}) - \hat{\tau}_j^{n,i}, j = 1,2\Big) \in G_r\Big) \cap \{\boldsymbol{\eta}^i > \mathbf{0}\},$$

when $0 \in G_r$, $i \leq r \leq l$, and the LHS is $\emptyset$ otherwise. We show that the event on the RHS of the previous equation is in $\big(\sigma(\boldsymbol{\eta}^r, r \geq 1, r \neq i) \vee \sigma(\hat{\tau}_j^{n,i}, j = 1,2) \vee \sigma(\xi_r^n, r \geq 1) \vee \mathcal{N}\big) \cap \{\boldsymbol{\eta}^i > \mathbf{0}\}$. It is enough to prove that this holds for the event $\bigcap_{r=i+1}^l \{\hat{\tau}_j^{n,i} = \hat{\tau}_j^{n,r}, \forall j\} \cap \{\boldsymbol{\eta}^i > \mathbf{0}\}$. We then can proceed just as Lemma A.1 in [51], and we omit the details here. Thus, we have proved (A.1) holds.

Next, we will show (A.2). The LHS of (A.2) has the following decomposition:

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} \leq s) E\left[H^{n,i}(t) | \mathcal{H}_s^n\right] \\ = \mathbf{1}(\eta_j^i \leq s - \hat{\tau}_j^{n,i}, \forall j) E\left[H^{n,i}(t) | \mathcal{H}_s^n\right] \tag{A.5}$$

$$
\begin{aligned}
&+\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&+\mathbf{1}(\eta_1^i > s-\hat{\tau}_1^{n,i} \ge 0, \eta_2^i \le s-\hat{\tau}_2^{n,i}) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&+\mathbf{1}(\eta_j^i > s-\hat{\tau}_j^{n,i} \ge 0, \forall j) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right].
\end{aligned}
$$

We start with the first term on the RHS of (A.5). Since $\mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j)$ is $\mathcal{H}_s^n$ measurable, by (5.12) we have that

$$
\begin{aligned}
&\mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) H^{n,i}(t) \\
&= \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) \\
&\quad - \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) \int_0^{\eta_1^i \wedge (s-\hat{\tau}_1^{n,i})^+} \int_0^{\eta_2^i \wedge (s-\hat{\tau}_2^{n,i})^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}).
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
&\mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&\quad = \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) \\
&\qquad - \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) \int_0^{\eta_1^i \wedge (s-\hat{\tau}_1^{n,i})^+} \int_0^{\eta_2^i \wedge (s-\hat{\tau}_2^{n,i})^+} \frac{1}{F^c(\boldsymbol{u})} dF(\boldsymbol{u}) \\
&\quad = \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) H^{n,i}(s).
\end{aligned} \tag{A.6}
$$

For the second term of the RHS of (A.5), we first observe that

$$
\begin{aligned}
&\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&\quad = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&\qquad\qquad - \mathbf{1}(\eta_j^i \le s-\hat{\tau}_j^{n,i}, \forall j) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right].
\end{aligned} \tag{A.7}
$$

Since $\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i})$ is $\mathcal{H}_s^n$-measurable, we have, by the definition of $H^{n,i}$ in (5.12),

$$
\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}) H^{n,i}(s). \tag{A.8}
$$

Combining (A.6), (A.7) and (A.8), we obtain

$$
\begin{aligned}
&\mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right] \\
&\quad = \mathbf{1}(\eta_1^i \le s-\hat{\tau}_1^{n,i}, \eta_2^i > s-\hat{\tau}_2^{n,i} \ge 0) H^{n,i}(s).
\end{aligned} \tag{A.9}
$$

Similar to (A.9), we can show for the third term on the RHS of (A.5)

$$
\mathbf{1}(\eta_1^i > s-\hat{\tau}_1^{n,i} \ge 0, \eta_2^i \le s-\hat{\tau}_2^{n,i}) E\left[H^{n,i}(t)|\mathcal{H}_s^n\right]
$$

$$= \mathbf{1}(\eta_1^i > s - \hat{\tau}_1^{n,i} \geq 0, \eta_2^i \leq s - \hat{\tau}_2^{n,i}) H^{n,i}(s),$$

and the details of its proof are omitted for brevity.

To complete the proof of (A.2), we only need to prove the following equation for the last term on the RHS of (A.5)

$$\begin{aligned} &\mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j) E\left[H^{n,i}(t) | \mathcal{H}_s^n\right] \\ &= \mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j) H^{n,i}(s). \end{aligned} \tag{A.10}$$

Observe that on the event that both tasks of job i have not completed service by time s, i.e., the event $\{\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j\}$, we have that $\boldsymbol{\eta}^i$ is independent of the entering-service processes $\hat{E}_j^n$, $j = 1, 2$, up to time s. Also, since $\boldsymbol{\eta}^i$ is independent of $\boldsymbol{\eta}^l$ for $l \neq i$, we obtain that, on the event $\{\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j\}$, by the definitions of $\mathcal{H}^n$ in (5.15) and $H^{n,i}$ in (5.12), $H^{n,i}(t)$ is dependent on $\mathcal{H}_s^n$ for $t > s$ only through $\boldsymbol{\eta}^i$ and $\hat{\tau}_j^{n,i}$, $j = 1, 2$. More precisely, let $\tilde{E}_j^n(u)$, $u \geq 0$, be the number of tasks having entered service in station j by time u that would have occured if tasks of the job with service vector $\boldsymbol{\eta}^i$ remained in service forever, $j = 1, 2$. Then, $\tilde{E}_j^n(u)$ is a Borel function of ξ_r^n, $r \geq 1$, $\boldsymbol{\eta}^p$, $p \geq 1$, $p \neq i$, on the one hand, and coincides with $E_j^n(u)$ for $u \leq s$ on the event $\{\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j\}$, on the other hand. Analogous to (A.4), we can see by the definition of $\mathcal{H}^n$

$$\begin{aligned} &\mathcal{H}_s^n \cap \{\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j\} \\ &\subset (\sigma(\xi_r^n, r \geq 1) \vee \sigma(\boldsymbol{\eta}^p, p \geq 1, p \neq i) \vee \sigma(\hat{\tau}_j^{n,i}, \forall j) \vee \mathcal{N}) \\ &\quad \cap \{\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j\}, \end{aligned}$$

where $\mathcal{N}$ includes all the null sets. Since $\mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j)$ is $\mathcal{H}_s^n$-measurable, by Lemma 3.6 of [29], we have

$$\begin{aligned} &\mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j) E\left[H^{n,i}(t) | \mathcal{H}_s^n\right] \\ &\quad = \mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i} \geq 0, \forall j) \frac{E\left[\mathbf{1}(\eta_j^i > s - \hat{\tau}_j^{n,i}, \forall j) H^{n,i}(t) | \hat{\tau}_j^{n,i}, \forall j\right]}{P(\eta_j^i > s - \hat{\tau}_j^{n,i} | \hat{\tau}_j^{n,i}, \forall j)}, \end{aligned}$$

where $0/0 = 0$. Evaluating the RHS of the above using (5.12), we have proved (A.10) holds. Thus, we have proved the martingale property of $H^{n,i}$.

We next prove the martingale property of $\tilde{H}^{n,i}$. By the definition of $\tilde{H}_k^n$ in (5.13) and the construction of the filtration $\mathcal{H}^n$ in (5.15), $\tilde{H}^{n,i}$ is $\mathcal{H}^n$-adapted. Note that, for each $t \geq 0$,

$$|\tilde{H}^{n,i}(t)| \leq 1 + \int_0^\infty \frac{\mathbf{1}\big(\max_j(\hat{\tau}_j^{n,i} + \eta_j^i) > u\big)}{\tilde{F}^c(u)} d\tilde{F}(u), \quad a.s.$$

By Fubini's theorem, we have $E\big[|\tilde{H}^{n,i}(t)|\big] < \infty$, for $t \geq 0$. We next show the martingale property of $\tilde{H}^{n,i}$, i.e., for $s < t$,

$$E\big[\tilde{H}^{n,i}(t)|\mathcal{H}_s^n\big] = \tilde{H}^{n,i}(s).$$

It suffices to show

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} > s)E\big[\tilde{H}^{n,i}(t)|\mathcal{H}_s^n\big] = 0, \tag{A.11}$$

and

$$\mathbf{1}(\hat{\tau}_1^{n,i} \vee \hat{\tau}_2^{n,i} \leq s)E\big[\tilde{H}^{n,i}(t)|\mathcal{H}_s^n\big] = \tilde{H}^{n,i}(s). \tag{A.12}$$

The proofs of (A.11) and (A.12) follow the same argument as the proof of (A.1) and (A.2), respectively, and the details are omitted here. This completes the proof. □

Acknowledgement. This research has been partly funded by an ARL Grant W911NF-14-1-0019 and NSF Grant CMMI-1538149 and Marcus Endowment Grant at Penn State University. We thank an anonymous reviewer for the helpful comments that have improved the presentation of our results.

REFERENCES

[1] R. J. Adler and J. E. Taylor. (2007) *Random Fields and Geometry.* Springer. MR2319516

[2] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin and G. B. Yom-Tov. (2015) Patient flow in hospitals: a data-based queueing-science perspective. *Stochastic Systems.* Vol. 5, No. 1, 146–194. MR3442392

[3] R. Atar, A. Mandelbaum and A. Zviran. (2012) Control of fork-join networks in heavy traffic. *Communication, Control, and Computing (Allerton), 50th Annual Allerton Conference on. IEEE.*

[4] F. Baccelli, M. Lelarge and S. Foss. (2004) Asymptotics of subexponential max plus networks: the stochastic event graph case. *Queueing Systems.* Vol. 46, No. 1-2, 75–96. MR2072276

[5] F. Baccelli, A. M. Makowski and A. Shwartz. (1989) The fork-join queue and related systems with synchronization constraints: stochastic ordering and computable bounds. *Advances in Applied Probability.* Vol. 21, No. 3, 629–660. MR1013655

[6] F. Baccelli, W. A. Massey and D. Towsley. (1989) Acyclic fork-join queueing networks. *Journal of the ACM (JACM).* Vol. 36, No. 3, 615–642. MR1072240

[7] P. Billingsley. (2009) *Convergence of Probability Measures.* John Wiley & Sons.

[8] J. Blanchet, Y. Pei and K. Sigman. (2015) Exact sampling of some multi-dimensional queueing models with renewal input. *Working paper.* https://arxiv.org/abs/1512.07284

[9] P. Brémaud. (1981) *Point Processes and Queues.* Springer, New York.

[10] H. Dai. (2011) Exact Monte Carlo simulation for fork-join networks. *Advances in Applied Probability.* Vol. 43, No. 2, 484–503.

[11] J. Dean and S. Ghemawat. (2008) MapReduce: simplified data processing on large clusters. *Communications of the ACM.* Vol. 51, No. 1, 107–113.
[12] A. B. Dieker and M. Lelarge. (2006) Tails for (max, plus) recursions under subexponentiality. *Queueing Systems.* Vol. 53, No. 4, 213–230.
[13] R. Durrett. (2010) *Probability: Theory and Examples.* Cambridge University Press. MR2722836
[14] S. N. Ethier and T. G. Kurtz. (2009) *Markov Processes: Characterization and Convergence.* John Wiley & Sons.
[15] L. Flatto and S. Hahn. (1984) Two Parallel Queues Created by Arrivals with Two Demands I. *SIAM Journal on Applied Mathematics.* Vol. 44, No. 5, 1041–1053. MR0759714
[16] L. Flatto. (1985) Two Parallel Queues Created by Arrivals with Two Demands II. *SIAM Journal on Applied Mathematics.* Vol. 45, No. 5, 861–878.
[17] J. Gallien and L. M. Wein. (2001) A simple and effective component procurement policy for stochastic assembly systems. *Queueing Systems.* Vol. 38, No. 2, 221–248.
[18] S. Halfin and W. Whitt. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research.* Vol. 29, No. 3, 567–588. MR0629195
[19] B. G. Ivanoff. (1980) The function space $\mathrm{D}([0,\infty)^q, \mathrm{E})$. *Canadian Journal of Statistics.* Vol. 8, No. 2, 179–191.
[20] J. Jacod and A. N. Shiryaev. (1987) *Limit Theorems for Stochastic Processes.* Springer-Verlag, Berlin.
[21] A. Jean-Marie and L. Gün. (1993) Parallel queues with resequencing. *Journal of the ACM (JACM).* Vol. 40, No. 5, 1188–1208.
[22] L. Jiang and R. E. Giachetti. (2008) A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science.* Vol. 11, 248–261.
[23] I. Karatzas and S. E. Shreve. (1991) *Brownian Motion and Stochastic Calculus.* Springer. MR1121940
[24] H. Kaspi and K. Ramanan. (2011) Law of large numbers limits for many-server queues. *Annals of Applied Probability.* Vol. 21, No. 1, 33–114.
[25] H. Kaspi and K. Ramanan. (2013) SPDE limits of many-server queues. *Annals of Applied Probability.* Vol. 23, No. 1, 145–229.
[26] L. J. Klementowski. (1978) PERT/CPM and supplementary analytical techniques: an analysis of aerospace usage. *Ph.D. Thesis.* Faculty of the School of Engineering of the Air Force Institute of Technology, Air University.
[27] S. S. Ko and R. F. Serfozo. (2004) Response times in M/M/s fork-join networks. *Advances in Applied Probability.* Vol. 36, No. 3, 854–871.
[28] S. S. Ko and R. F. Serfozo. (2008) Sojourn times in G/M/1 fork-join networks. *Naval Research Logistics.* Vol. 55, No. 5, 432–443.
[29] E. V. Krichagina and A. A. Puhalskii. (1997) A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems.* Vol. 25, No. 1-4, 235–280.
[30] R. C. Larson, M. F. Cahn and M. C. Shell. (1993) Improving the New York City arrest-to-arraignment system. *Interfaces.* Vol. 23, No. 1, 76–96.
[31] M. Lin, L. Zhang, A. Wierman and J. Tan. (2013) Joint optimization of overlapping phases in MapReduce. *Proceedings of IFIP Performance.* Vol. 70, No. 10, 720–735.
[32] R. S. Liptser and A. N. Shiryaev. (1989) *Theory of Martingales.* Kluwer, Dordrecht.
[33] H. Lu and G. Pang. (2015) Gaussian limits of a fork-join network with non-exchangeable synchronization in heavy-traffic. *Mathematics of Operations Research.* Vol. 41, No. 2, 560–595.

[34] H. Lu, G. Pang and M. Mandjes. (2016) A functional central limit theorem for Markov additive arrival processes and its applications to queueing systems. *Queueing Systems*. Vol. 84, No. 3, 381–406.
[35] H. Lu and G. Pang. (2017) Heavy-traffic limits for an infinite-server fork-join queueing system with dependent and disruptive services. *Queueing Systems*. Vol. 85, No. 1–2, 67–115.
[36] M. Mandelker. (1982) Continuity of monotone functions. *Pacific Journal of Mathematics*. Vol. 99, No. 2, 413–418.
[37] A. W. Marshall and I. Olkin. (1967) A multivariate exponential distribution. *Journal of the American Statistical Association*. Vol. 62, No. 317, 30–44.
[38] A. Mandelbaum and P. Momčilović. (2012) Queues with many servers and impatient customers. *Mathematics of Operations Research*. Vol. 37, No. 1, 41–65. MR2891146
[39] G. Neuhaus. (1971) On weak convergence of stochastic processes with multidimensional time parameter. *Annals of Mathematical Statistics*. Vol. 42, No. 4, 1285–1295.
[40] V. Nguyen. (1993) Processing networks with parallel and sequential tasks: heavy traffic analysis and Brownian Limits. *Annals of Applied Probability*. Vol. 3, No. 1, 28–55.
[41] V. Nguyen. (1994) The trouble with diversity: fork-join networks with heterogeneous customer population. *Annals of Applied Probability*. Vol. 4, No. 1, 1–25.
[42] G. Pang, R. Talreja and W. Whitt. (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*. 4, 193–267.
[43] G. Pang and W. Whitt. (2010) Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*. Vol. 65, No. 4, 325–364.
[44] G. Pang and W. Whitt. (2012) Infinite-server queues with batch arrivals and dependent service times. *Probability in Engineering and Informational Sciences*. Vol. 26, No. 2, 197–220.
[45] G. Pang and W. Whitt. (2013) Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems*. Vol. 73, No. 2, 119–146.
[46] G. Pang and Y. Zhou. (2016) Two-parameter process limits for an infinite-server queue with arrival dependent service times. Forthcoming in *Stochastic Processes and their Applications*. http://dx.doi.org/10.1016/j.spa.2016.08.003
[47] G. Pang and Y. Zhou. (2017) Two-parameter process limits for infinite-server queues with dependent service times via chaining bounds. *Under review*.
[48] D. Pinotsi and M. A. Zazanis. (2005) Synchronized queues with deterministic arrivals. *Operations Research Letters*. Vol. 33, No. 6, 560–566.
[49] B. Prabhakar, N. Bambos and T. S. Mountford. (2000) The synchronization of Poisson processes and queueing networks with service and synchronization nodes. *Advances in Applied Probability*. Vol. 32, No. 3, 824–843.
[50] A. A. Puhalskii and J. E. Reed. (2010) On many-server queues in heavy traffic. *Annals of Applied probability*. Vol. 20, No. 1, 129–195.
[51] J. E. Reed. (2009) The G/GI/N queue in the Halfin-Whitt regime. *Annals of Applied Probability*. Vol. 19, No. 6, 2211–2269.
[52] M. Sklar. (1959) Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* 8.
[53] M. Takahashi, H. Ōsawa and T. Fujisawa. (2000) On a synchronization queue with two finite buffers. *Queueing Systems*. Vol. 36, No. 1-3, 107–123.
[54] J. Tan, X. Meng and L. Zhang. (2012) Delay tails in MapReduce scheduling. *ACM SIGMETRICS Performance Evaluation Review*. Vol. 40, No. 1, 5–16.
[55] A. W. van der Vaart and J. A. Wellner. (1996) *Weak Convergence and Empirical Processes*. Springer.

[56] S. Varma. (1990) Heavy and light traffic approximations for queues with synchronization constraints. *Ph.D. Thesis.*

[57] W. Wang, K. Zhu, L. Ying, J. Tan and L. Zhang. (2013) Map task scheduling in MapReduce with data locality: throughput and heavy-traffic optimality. *Proceedings of IEEE INFOCOM.* 1609–1617.

[58] W. Whitt. (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues.* Springer.

[59] C. J. Willits and D. C. Dietz. (2001) Nested fork-join queueing network model for analysis of airfield operations. *Journal of Aircraft.* Vol. 38, No. 5, 848–855.

[60] I. Zaied. (2012) The offered load in fork-join networks: calculations and applications to service engineering of emergency department. *M.Sc. Research Thesis.* Technion.

[61] A. Zviran. (2011) Fork-join networks in heavy traffic: diffusion approximations and control. *M.Sc. Research Thesis.* Technion.

Hongyuan Lu
The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802, USA
E-mail: hzl142@psu.edu

Guodong Pang
The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802, USA
E-mail: gup3@psu.edu