A FUNCTIONAL CENTRAL LIMIT THEOREM FOR MARKOV ADDITIVE ARRIVAL PROCESSES AND ITS APPLICATIONS TO QUEUEING SYSTEMS

HONGYUAN LU, GUODONG PANG, AND MICHEL MANDJES

Abstract. We prove a functional central limit theorem for Markov additive arrival processes (MAAPs) where the modulating Markov process has the transition rate matrix scaled up by n^α ($\alpha > 0$) and the mean and variance of the arrival process are scaled up by n. It is applied to an infinite-server queue and a fork-join network with a non-exchangeable synchronization constraint, where in both systems both the arrival and service processes are modulated by a Markov process. We prove FCLTs for the queue length processes in these systems joint with the arrival and departure processes, and characterize the transient and stationary distributions of the limit processes. We also observe that the limit processes possess a stochastic decomposition property.

1. Introduction

Markov additive arrival processes (MAAP) have been used to model the arrival processes of many stochastic systems, for example, telecommunication and service systems in random environments [2]. Its usefulness lies in capturing burstiness in the arrival processes, thus departing from the usual renewal-type assumptions. MAAPs are described by a couple (A, X), where the process A is the counting process of arrivals, and the process X is a modulating Markov process. A popular model is the Markov-modulated Poisson process (MMPP), which has been widely used to model a variety of relevant stochastic systems [2, 25]. For a broad range of queues with MMPP input analytical results have been derived, often employing the matrix computational approach. The main objective of this paper is to generate functional central limit theorems (FCLTs) for MAAPs, in particular, the counting process A, and their applications in specific, practically relevant, queueing systems.

FCLTs for MMPPs have been studied in the literature under two types of scalings. In the first scaling, time is scaled up by a parameter n while space is scaled down by $\sqrt{n}$, and thus the transition times of the modulating Markov process are implicitly accelerated by a factor n. Under this scaling, assuming that the modulating Markov process has a finite number of states and is irreducible, an FCLT can be proven for the scaled arrival process, where the limit process is a Brownian motion (reviewed in (2.1)–(2.5)). This has been applied to prove heavy-traffic limits for single-server queueing (network) models; see, e.g., [25, Ch. 9]. Under this same scaling, Steichen [23] considered an MAAP where the arrival process in each state can be non-Poisson, and proved an FCLT with a Brownian motion limit. That result was also applied to study some single-server queueing networks in [23].

In the second scaling, time is not scaled, but the arrival rates in each state are scaled up by n and the space is scaled down by $\sqrt{n}$, while at the same time the transition rates of the modulating Markov process are scaled up by n^α for some $\alpha > 0$. Under this scaling, an FCLT has recently been proved for the scaled arrival process in [1], where the limit process is a Brownian motion (reviewed in (2.6)–(2.8)). This is then applied in [1] to prove an FCLT for the $M/M/\infty$ queue with MMPP input. This scaling is useful in many-server systems, where the demand is relatively large but service times do not scale as the demand gets larger, and the modulating Markov process may speed up or slow down.

Date: November 17, 2015.

Key words and phrases. Markov additive arrival process, functional central limit theorem, infinite-server queues, fork-join networks with non-exchangeable synchronization, Gaussian limits, stochastic decomposition.

FCLTs for MAAPs. MMPPs, in which the Poisson arrival rate jumps between several values, significantly generalize the traditional Poisson setting. Nonetheless, in many applications the assumption of the input being locally Poisson is not adequate. To remedy this we consider in this paper a general class of MAAPs, where the arrival process in each state can be a general stationary counting process, including renewal processes. We prove an FCLT for this class of MAAPs, in Theorem 2.1, under the second type of scaling and under three regimes of α values, i.e., $0 < \alpha < 1$, $\alpha = 1$ and $\alpha > 1$. The limit process is also a Brownian motion, whose variance coefficient compactly captures the variabilities in the interarrival times in each state as well as the variabilities in the modulating Markov process. We apply this FCLT to two queueing systems: a general infinite-server queue and a fork-join network with the non-exchangeable synchronization (NES) constraint.

General infinite-server queue. Several recent papers have studied infinite-server queues with MMPP input. Exact analysis and related approximations have been derived for specific infinite-server queues in random environments (Markov or semi-Markov modulated) in [3, 5, 9, 11, 12, 17, 19]. In [1], an $M/M/\infty$ queue with MMPP input is studied, leading to an FCLT for the queue length process under the second type of scaling mentioned above. [4] studies an $M/GI/\infty$ queue with MMPP input and general service times depending on the state of the modulating Markov process upon arrival. The exact mean and variance formulas for the transient and stationary distributions of the queue length process are provided, and the asymptotic results are also obtained in the regime where the arrival rates are scaled up by n and the transition rates are scaled up by $n^{1+\epsilon}$ for some $\epsilon > 0$. Central limit theorems are proved for the $M/M/\infty$ queue with both the arrival and services modulated by a finite-state Markov process in [6, 7], where the arrival rates are scaled up by n.

In Section 3, we establish an FCLT for the queue length process joint with the arrival and departure processes in the $G/G/\infty$ queue where both the arrival process and the service time distributions are modulated by a Markov process (applying Theorem 2.1), thus generalizing the existing literature substantially. The limiting queue-length and departure processes are continuous Gaussian processes, of which we characterize the transient and steady-state distributions. We also derive a stochastic decomposition property: the variabilities of the arrival process and modulating Markov process are captured in one limit component, while those of the service process are captured in a second independent limit component.

Fork-join network with NES. In our second application, we consider a fork-join network with NES, where both the arrival process and the joint service time distributions of the parallel tasks of each job are modulated by a Markov process. In the network, each job is forked into a fixed number of parallel tasks, each of which is processed in a multi-server service station, and after service completion, each task will join a buffer associated with the service station, waiting for synchronization. The NES constraint requires that synchronization occurs only when all the tasks of the same job are completed. It is important to understand the joint dynamics of the service process as well as the waiting buffers for synchronization.

Heavy-traffic limits are proved for a single-class multi-server fork-join network with NES, in the underloaded quality-driven (QD) regime [14] and the critically loaded quality-and-efficiency driven (QED) regime [16]. The setup considered is such that the arrival process is general (satisfying an FCLT), whereas the service times of the parallel tasks form i.i.d. random vectors that can be correlated. In addition, in [15], an infinite-server fork-join network with NES in a renewal alternating environment (up-down cycles) is studied, where the service vectors of parallel tasks are correlated and the service processes are interrupted during the down periods.

In this paper we study a multi-server fork-join network with NES in the QD regime, where both the arrival and service processes are modulated by a Markov process. We apply our FCLT for the MAAP to obtain a multi-dimensional Gaussian limit process for the processes representing the number of tasks in service at each station and the numbers of tasks in the waiting buffer for synchronization associated with each station, jointly with the arrival process and the process

representing the number of synchronized jobs. We characterize the transient and steady-state joint distributions of the limit queueing processes, as multivariate Gaussian distributions, and of the synchronized process, as a Gaussian distribution. We also observe a similar stochastic decomposition property as in the infinite-server queues above, where the two independent limit components capture the variabilities of the arrival and modulating Markov processes, and of the service processes separately.

1.1. **Organization of the paper.** The rest of the paper is organized as follows. We finish this section below with a summary of notations used in the paper. In Section 2, we present the general MAAP, review the existing FCLTs for MAAPs, and state the new FCLT under the second type of scaling. In Section 3, we apply the FCLT for the MAAP to a general infinite-server queueing model with both arrival and service times modulated by a Markov process. In Section 4, we apply the FCLT for the MAAP to a fork-join network with both arrival and service processes being modulated by a Markov process. The proofs of these results are presented in Section 5. We make some concluding remarks in Section 6.

1.2. **Notations.** The following notations will be used throughout the paper. $\mathbb{R}$ and $\mathbb{R}_+$ ($\mathbb{R}^d$ and $\mathbb{R}^d_+$, respectively) denote sets of real and real non-negative numbers (d-dimensional vectors, respectively, $d \geq 2$). For $a, b \in \mathbb{R}$, we denote $a \wedge b := \min(a, b)$. For any $x \in \mathbb{R}_+$, $\lfloor x \rfloor$ is used to denote the largest integer no greater than x. We use bold letter to denote a vector, e.g., $\mathbf{x} := (x_1, \ldots, x_N) \in \mathbb{R}^N$. $\mathbf{1}(A)$ is used to denote the indicator function of a set A. 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 $\mathcal{S}$ with the maximum metric 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., [8, 10, 25]). 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. Let $\mathbb{D}([0,\infty)^k, \mathbb{R}) \equiv \mathbb{D}_k$ denote the space of all "continuous from above with limits from below" real-valued functions on $[0,\infty)^k$ with the generalized Skorohod J_1 topology [18, 24] for $k \geq 2$. Weak convergence of probability measures μ_n to μ will be denoted as $\mu_n \Rightarrow \mu$.

2. An FCLT for Markov Additive Arrival Processes

Consider a Markov additive arrival process (A, X). The process $X = \{X(t) : t \geq 0\}$ is a finite-state irreducible stationary Markov process with state space $\mathcal{S} = \{1, \ldots, I\}$ and transition rate matrix $Q = (q_{ij})_{i,j=1,\ldots,I}$. The process $A = \{A(t) : t \geq 0\}$ is a counting process modulated by the Markov process X, defined as follows. Let $\pi = (\pi_1, \ldots, \pi_I)$ be the stationary distribution of the Markov process X. We assume that the process starts in stationarity at time 0.

We introduce some auxiliary notations. Let Π be a matrix with each row being the steady-state vector π, and $P(t) = (P_{ij}(t))_{i,j=1,\ldots,I}$ be the transition matrix, that is, $P_{ij}(t) := P(X(t) = j | X(0) = i)$ for each $t \geq 0$. Let $Z = (Z_{ij})_{i,j=1,\ldots,I}$ be the fundamental matrix, given by

$$Z_{ij} := \int_0^\infty (P_{ij}(t) - \pi_j) dt.$$

It holds that $Z = (\Pi - Q)^{-1} - \Pi$.

When the process A is an MMPP, that is, arrivals follow a Poisson process with rate λ_i when $X = i$, $i \in \mathcal{S}$, FCLTs are proved for the process A in two different scalings. In the first scaling that

was introduced in Section 1, both time and space are scaled by n, and the diffusion-scaled process $\tilde{A}^n = \{\tilde{A}^n(t) : t \geq 0\}$ is defined by

$$\tilde{A}^n(t) := n^{-1/2} \left(A(nt) - \sum_{i=1}^{I} \pi_i \lambda_i nt \right), \quad t \geq 0. \tag{2.1}$$

By Theorem 2.3.4 in [26], one can show that

$$\tilde{A}^n \Rightarrow \tilde{A} \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{2.2}$$

where $\tilde{A} = \{\tilde{A}(t) : t \geq 0\}$ is a driftless Brownian motion with variance coefficient

$$\sigma^2 := \bar{\lambda} + \bar{\beta}, \tag{2.3}$$

with

$$\bar{\lambda} := \sum_{i=1}^{I} \pi_i \lambda_i, \tag{2.4}$$

and

$$\bar{\beta} := 2 \sum_{i=1}^{I} \sum_{j=1}^{I} \lambda_i \lambda_j \pi_i Z_{ij}. \tag{2.5}$$

See also the discussion in Example 9.6.2 in [25]. Note that under this scaling, the transition rates of the modulating Markov process are scaled up by n.

In the second scaling introduced in Section 1, time is not scaled, but the arrival rates λ are scaled by n and transition rate matrices are scaled by n^α for $\alpha > 0$. Namely, we consider a sequence of the processes (A^n, X^n) indexed by n, and write the corresponding quantities by a superscript n. Assume that $\lambda_i^n/n \to \lambda_i > 0$ for $i \in \mathcal{S}$ as $n \to \infty$ and $Q^n = n^\alpha Q$ for some $\alpha > 0$. Note that the stationary distribution of X^n remains the same, π. Define the diffusion-scaled process $\hat{A}^n = \{\hat{A}^n(t) : t \geq 0\}$ by

$$\hat{A}^n(t) := \frac{1}{n^\delta} \left(A^n(t) - \sum_{i=1}^{I} \pi_i \lambda_i^n t \right), \quad \text{for} \quad \delta > 0, \quad t \geq 0. \tag{2.6}$$

Then it is shown in [1] that

$$\hat{A}^n \Rightarrow \hat{A} \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{2.7}$$

where the limit process $\hat{A} = \{\hat{A}(t) : t \geq 0\}$ is a driftless Brownian motion with variance coefficient

$$\sigma^2(\alpha) := \begin{cases} \bar{\beta}, & \alpha < 1, \ \delta = 1 - \alpha/2, \\ \bar{\lambda} + \bar{\beta}, & \alpha = 1, \ \delta = 1/2, \\ \bar{\lambda}, & \alpha > 1, \ \delta = 1/2, \end{cases} \tag{2.8}$$

with $\bar{\lambda}$ and $\bar{\beta}$ being defined in (2.4) and (2.5), respectively.

Remark 2.1. When $\alpha = 1$, the limit processes under both scalings in fact coincide, as the arrival process and the modulating Markov process are sped up at the same rate. When $\alpha > 1$, the modulating Markov process is sped up at a faster rate than the arrival process in each state, and thus the variability in the limit comes only from the Poisson processes with the spatial scaling $n^{-1/2}$. When $0 < \alpha < 1$, the modulating Markov process is sped up at a slower rate than the arrival process in each state, and thus the variability in the limit comes only from the modulating Markov process with the spatial scaling $n^{-(1-\alpha/2)}$.

In this paper, we consider the second type of scaling and prove an FCLT for the diffusion-scaled processes $\hat{A}^n$ when the process A^n is general, including renewal process, in each state of X^n.

Let τ_k^n be the k^{th} jump time of X^n for $k = 1, 2, 3, \ldots$ and $\tau_0^n \equiv 0$. For each $i \in \mathcal{S}$, define

$$\lambda_i^n := E\left[\frac{A^n(\tau_k^n + s) - A^n(\tau_k^n)}{s}\Big| X^n(u) = i \quad \text{for} \quad \tau_k^n \le u \le \tau_k^n + s\right], \tag{2.9}$$

and

$$\nu_i^n := E\left[\frac{(A^n(\tau_k^n + s) - A^n(\tau_k^n))^2 - (\lambda_i^n)^2 s^2}{s}\Big| X^n(u) = i \quad \text{for} \quad \tau_k^n \le u \le \tau_k^n + s\right]. \tag{2.10}$$

Assume that $\boldsymbol{\lambda}^n = (\lambda_1^n, \ldots, \lambda_I^n)$ and $\boldsymbol{\nu}^n = (\nu_1^n, \ldots, \nu_I^n)$ are positive vectors. Note that when the process A^n is Poisson in each state of X^n, we have that $\lambda_i^n = \nu_i^n$, $i = 1, \ldots, I$. Note also that if the arrival process is renewal in each state, the parameter $\nu_i^n = \lambda_i^n (c_{a,i}^n)^2$, where $c_{a,i}^n$ is the coefficient of variation (CV) of the interarrival times when the Markov process X^n is in state i.

Then, we can write, for each $t \ge 0$,

$$E[A^n(t)|X^n(s),\ 0 \le s \le t] = \int_0^t \lambda_{X^n(s)}^n ds = \sum_{i=1}^I \int_0^t \lambda_i^n \mathbf{1}(X^n(s) = i) ds, \tag{2.11}$$

and

$$Var[A^n(t)|X^n(s),\ 0 \le s \le t] = \int_0^t \nu_{X^n(s)}^n ds = \sum_{i=1}^I \int_0^t \nu_i^n \mathbf{1}(X^n(s) = i) ds. \tag{2.12}$$

We make the following assumption on the parameters.

Assumption 1. *The parameters $\boldsymbol{\lambda}^n$ and $\boldsymbol{\nu}^n$ satisfy*

$$\frac{\boldsymbol{\lambda}^n}{n} \to \boldsymbol{\lambda} \in \mathbb{R}_+^I, \quad \frac{\boldsymbol{\nu}^n}{n} \to \boldsymbol{\nu} \in \mathbb{R}_+^I \quad \text{as} \quad n \to \infty.$$

The transition rate matrix $Q^n = n^\alpha Q$ for some $\alpha > 0$.

We now state the main result of this section. Its proof, as well as the proofs of all results presented in Sections 3 and 4, are provided in Section 5.

Theorem 2.1. *Under Assumption 1, for the diffusion-scaled process $\hat{A}^n$ in (2.6), (2.7) holds where the limit process $\hat{A}$ is a driftless Brownian motion with variance coefficient*

$$\sigma^2(\alpha) := \begin{cases} \bar{\beta}, & 0 < \alpha < 1,\ \delta = 1 - \alpha/2, \\ \bar{\nu} + \bar{\beta}, & \alpha = 1,\ \delta = 1/2, \\ \bar{\nu}, & \alpha > 1,\ \delta = 1/2, \end{cases} \tag{2.13}$$

with $\bar{\beta}$ being defined in (2.5) and

$$\bar{\nu} := \sum_{i=1}^I \pi_i \nu_i. \tag{2.14}$$

Remark 2.2. When the modulating Markov process X^n is in state i, if the process A^n is renewal, we obtain $\bar{\nu} = \sum_{i=1}^I \pi_i \lambda_i c_{a,i}^2$, where λ_i is the arrival rate and $c_{a,i}$ is the CV of the interarrival times in the limit and $\nu_i = \lambda_i c_{a,i}^2$.

3. Application to Infinite-Server Queues

In this section, we apply the FCLT of the MAAP process, Theorem 2.1, to $G/GI/\infty$ queues with Markov modulated arrival and service processes. It is shown in [13] that an FCLT for the number of customers/jobs in a $G/GI/\infty$ queue holds, provided that the arrival process satisfies an FCLT and the service times are i.i.d. with a general distribution. Our FCLT below extends the existing results in [1] and [4] by allowing more general arrival process, in that in each state of the underlying Markov process, the arrival process can be a general stationary point process, including a renewal process. It also proves the joint convergence of arrivals, queue length and departure processes (rather than just queue length). The limiting queue-length process is a continuous Gaussian process, and possesses a stochastic decomposition property. Our results also generalize [13] for $G/G/\infty$ queues in a Markov random environment.

Consider a sequence of $G/G/\infty$ queues modulated by a Markov process X^n, behaving as described in Section 2. The arrival process A^n is an MAAP with τ_k^n denoting the arrival time of job k, $k \geq 1$. The service times $\{\eta_{k,i} : k \geq 1\}$ are i.i.d. with a general distribution F_i (independent of n) when the underlying Markov process X^n is in state i upon the customer's arrival. Namely, we assume that the service time distribution of a customer is determined at the epoch of the arrival time, according to the state of the underlying Markov process X^n. We also assume that conditional on the modulating Markov process X^n, the arrival and service processes are independent, and that the system starts empty. Let $F_i^c := 1 - F_i$, $i = 1, \ldots, I$. Let $Q^n = \{Q^n(t) : t \geq 0\}$ be the queue length process describing the evolution of the number of customers in the system. Let $D^n = \{D^n(t) : t \geq 0\}$ be the departure process counting the number of completed jobs. We have the following balance equation:

$$D^n(t) = A^n(t) - Q^n(t), \quad t \geq 0. \tag{3.1}$$

Define the diffusion-scaled processes $\hat{Q}^n = \{\hat{Q}^n(t) : t \geq 0\}$ and $\hat{D}^n = \{\hat{D}^n(t) : t \geq 0\}$ by

$$\hat{Q}^n(t) := n^{-\delta}(Q^n(t) - nq(t)), \quad \hat{D}^n(t) := n^{-\delta}(D^n(t) - n\mathsf{d}(t)) = \hat{A}^n(t) - \hat{Q}^n(t), \quad t \geq 0, \tag{3.2}$$

where

$$q(t) := \sum_{i=1}^{I} \lambda_i \pi_i \int_0^t F_i^c(s)ds, \quad \text{and} \quad \mathsf{d}(t) := \sum_{i=1}^{I} \lambda_i \pi_i \int_0^t F_i(s)ds, \quad t \geq 0.$$

Theorem 3.1. *For the sequence of $G/GI/\infty$ models with Markov modulated arrival and service processes described above,*

$$(\hat{A}^n, \hat{Q}^n, \hat{D}^n) \Rightarrow (\hat{A}, \hat{Q}, \hat{D}) \quad \textit{in} \quad \mathbb{D}^3 \quad \textit{as} \quad n \to \infty,$$

where $\hat{A}$ is the arrival limit defined in Theorem 2.1, the process $\hat{D} = \{\hat{D}(t) : t \geq 0\}$ is defined by $\hat{D}(t) := \hat{A}(t) - \hat{Q}(t)$, $t \geq 0$, and

$$\hat{Q} := \begin{cases} \hat{Q}_1, & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ \hat{Q}_1 + \hat{Q}_2, & \delta = 1/2, \ \alpha \geq 1. \end{cases}$$

The limit process $\hat{Q}_1 = \{\hat{Q}_1(t) : t \geq 0\}$ is a continuous Gaussian process, defined by

$$\hat{Q}_1(t) := \sigma(\alpha) \int_0^t \Big(\sum_{i=1}^{I} \pi_i F_i^c(t-s) \Big) dW(s), \quad t \geq 0,$$

where W is a standard Brownian motion and $\sigma^2(\alpha)$ is defined in (2.13). The limit process $\hat{Q}_2 = \{\hat{Q}_2(t) : t \geq 0\}$ is a continuous Gaussian process defined by

$$\hat{Q}_2(t) := \sum_{i=1}^{I} \int_0^t \int_0^\infty \mathbf{1}(s + x_i > t) d\hat{K}_i(\pi_i \lambda_i s, x_i) \tag{3.3}$$

where the processes $\hat{K}_i = \{\hat{K}_i(s,x) : s, x \geq 0\}$, $i = 1, \ldots, I$, *are independent Kiefer processes with mean* 0 *and covariance function*

$$Cov(\hat{K}_i(s,x), \hat{K}_i(t,y)) = (s \wedge t)(F_i(x \wedge y) - F_i(x)F_i(y)), \quad s,t,x,y \geq 0,$$

for each $i = 1, \ldots, I$. *The processes* W *and* $\hat{K}_i$, $i = 1, \ldots, I$, *are independent, and thus, so are the processes* $\hat{Q}_1$ *and* $\hat{Q}_2$.

Here the integrals in (3.3) are defined in the mean-square sense following [13]; see the precise definition in (5.35)–(5.36).

Remark 3.1. We remark that there is a stochastic decomposition property in the limit process, as shown in the independence of $\hat{Q}_1$ and $\hat{Q}_2$. Note that in the corresponding prelimit processes are evidently dependent because of the modulating Markov process. The limit process $\hat{Q}_1$ captures the variabilities resulting from the arrival process, as well as those resulting from the modulating Markov process. The limit process $\hat{Q}_2$ captures the variabilities from the service process, while, perhaps surprisingly, it is not affected by the variabilities of the modulating Markov process other than the steady-state distribution π. This is also shown in the following characterization of the limit processes.

Corollary 3.1. *Under the assumptions of Theorem 3.1, the limit process* $\hat{Q}$ *is Gaussian, with mean* 0 *and covariance function*

$$Cov(\hat{Q}(t), \hat{Q}(t+u)) = \begin{cases} \bar{\beta} \int_0^t \Big(\sum_{i=1}^I \pi_i F_i^c(s) \Big) \Big(\sum_{i=1}^I \pi_i F_i^c(s+u) \Big) ds, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \sigma^2(\alpha) \int_0^t \Big(\sum_{i=1}^I \pi_i F_i^c(s) \Big) \Big(\sum_{i=1}^I \pi_i F_i^c(s+u) \Big) ds & \\ \qquad + \sum_{i=1}^I \pi_i \lambda_i \int_0^t \big(F_i(s) F_i^c(u+s) \big) ds, & \delta = 1/2,\ \alpha \geq 1, \end{cases}$$

for $t, u \geq 0$, *where* $\bar{\beta}$ *and* $\sigma^2(\alpha)$ *are defined in* (2.5) *and* (2.13), *respectively. Its stationary distribution has variance*

$$Var(\hat{Q}(\infty)) = \begin{cases} \bar{\beta} \int_0^\infty \Big(\sum_{i=1}^I \pi_i F_i^c(s) \Big)^2 ds, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \sum_{i=1}^I \pi_i \lambda_i m_{s,i} + \sigma^2(\alpha) \int_0^\infty \Big(\sum_{i=1}^I \pi_i F_i^c(s) \Big)^2 ds - \sum_{i=1}^I \pi_i \lambda_i \int_0^\infty (F_i^c(s))^2 ds, & \\ \qquad \delta = 1/2,\ \alpha \geq 1, \end{cases}$$

with $m_{s,i}$ *being the mean service time associated with* F_i. *In addition, the limit process* $\hat{D}$ *is Gaussian, with mean* 0 *and covariance function*

$$Cov(\hat{D}(t), \hat{D}(t+u)) = \begin{cases} \bar{\beta} \int_0^t \Big(\sum_{i=1}^I \pi_i F_i(s) \Big) \Big(\sum_{i=1}^I \pi_i F_i(s+u) \Big) ds, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \sigma^2(\alpha) \int_0^t \Big(\sum_{i=1}^I \pi_i F_i(s) \Big) \Big(\sum_{i=1}^I \pi_i F_i(s+u) \Big) ds & \\ \qquad + \sum_{i=1}^I \pi_i \lambda_i \int_0^t \big(F_i(s) F_i^c(u+s) \big) ds, & \delta = 1/2,\ \alpha \geq 1, \end{cases}$$

for $t, u \geq 0$, *and* $\lim_{t \to \infty} t^{-1} Var(\hat{D}(t)) = \sigma^2(\alpha)$.

Remark 3.2. When F_i, $i = 1, \ldots, I$, are identical, our results establish an FCLT for $G/GI/\infty$ queues with an MAAP and i.i.d. service times. Moreover, when the arrival process is an MMPP and the service times are exponential with rate μ (independent of the modulating Markov process), our results reduce to those in [1] for $M/M/\infty$ queues.

4. Application to Fork-Join Networks

In this section, we apply the FCLT for the MAAP to a many-server fork-join network with the non-exchangeable synchronization (NES) constraint, where both the arrival process and the joint service time distribution of the parallel tasks are modulated by a Markov process.

Consider a sequence of many-server fork-join networks with NES indexed by n and let $n \to \infty$. We assume that the systems are operating in the QD regime, which is asymptotically equivalent to systems with infinite-server service stations. There is a single class of customers. Let the arrival processes A^n be an MAAP as described in Section 2. Let $\boldsymbol{\eta}^{\ell,i} = (\eta_1^{\ell,i}, \ldots, \eta_K^{\ell,i})$ be the service times that customer ℓ brings in for the K parallel tasks when the underlying Markov process X^n is in state i at the epoch of arrival. Assume that the service times $\{\boldsymbol{\eta}^{\ell,i} : \ell \geq 1\}$ are i.i.d. with a continuous joint distribution function $F^{(i)}$ and marginals $F_k^{(i)}$, $k = 1, \ldots, K$ and $i = 1, \ldots, I$. Let $F_{j,k}^{(i)}$ be the joint distribution of the service times of parallel tasks j, k for $j, k = 1, \ldots, K$ and $i = 1, \ldots, I$. Let $F_{\mathrm{m}}^{(i)}$ be the distribution of the maximum of the service times $\eta_1^{\ell,i}, \ldots, \eta_K^{\ell,i}$, i.e., $F_{\mathrm{m}}^{(i)}(x) = P(\eta_j^{1,i} \leq x, \ \forall j)$ for $x \geq 0$. Denote $G_k^{(i)} := 1 - F_k^{(i)}$ for $k = 1, \ldots, K$, and $G_{\mathrm{m}}^{(i)} := 1 - F_{\mathrm{m}}^{(i)}$, $i = 1, \ldots, I$.

Let $\boldsymbol{Q}^n = (Q_1^n, \ldots, Q_K^n)$ be the numbers of tasks in service at service stations $k = 1, \ldots, K$. Let $\boldsymbol{Y}^n = (Y_1^n, \ldots, Y_K^n)$ be the numbers of tasks that have completed service but are waiting for service at the waiting buffers for synchronization corresponding to the service stations $k = 1, \ldots, K$. Let $S^n = \{S^n(t) : t \geq 0\}$ be the process counting the number of synchronized jobs. Define the diffusion-scaled processes $\hat{\boldsymbol{Q}}^n = (\hat{Q}_1^n, \ldots, \hat{Q}_K^n)$, $\hat{\boldsymbol{Y}}^n = (\hat{Y}_1^n, \ldots, \hat{Y}_K^n)$ and $\hat{S}^n$ by

$$\hat{Q}_k^n(t) := \frac{1}{n^\delta}(Q_k^n(t) - nq_k(t)), \quad \hat{Y}_k^n(t) := \frac{1}{n^\delta}(Y_k^n(t) - ny_k(t)), \quad k = 1, \ldots, K, \ t \geq 0,$$

and

$$\hat{S}^n(t) := \frac{1}{n^\delta}(S^n(t) - n\mathsf{s}(t)), \quad t \geq 0,$$

where

$$q_k(t) := \sum_{i=1}^I \lambda_i \pi_i \int_0^t G_k^{(i)}(s)ds, \quad y_k(t) := \sum_{i=1}^I \lambda_i \pi_i \int_0^t (G_{\mathrm{m}}^{(i)}(s) - G_k^{(i)}(s))ds, \quad k = 1, ..., K, \ t \geq 0,$$

and

$$\mathsf{s}(t) := \sum_{i=1}^I \lambda_i \pi_i \int_0^t F_{\mathrm{m}}^{(i)}(s)ds, \quad t \geq 0.$$

Theorem 4.1. *For the fork-join networks with NES and Markov modulated arrival and service processes described above,* $(\hat{A}^n, \hat{\boldsymbol{Q}}^n, \hat{\boldsymbol{Y}}^n, \hat{S}^n) \Rightarrow (\hat{A}, \hat{\boldsymbol{Q}}, \hat{\boldsymbol{Y}}, \hat{S})$ *in* $\mathbb{D}^{2K+2}$ *as* $n \to \infty$*, where* $\hat{A}$ *is the arrival limit defined in Theorem 2.1,* $\hat{\boldsymbol{Q}} = (\hat{Q}_1, \ldots, \hat{Q}_K)$, $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_K)$ *and* $\hat{S}$ *are defined as follows:*

$$\hat{Q}_k := \begin{cases} \hat{Q}_{k,1}, & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ \hat{Q}_{k,1} + \hat{Q}_{k,2}, & \delta = 1/2, \ \alpha \geq 1, \end{cases}$$

$$\hat{Y}_k := \begin{cases} \hat{Y}_{k,1}, & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ \hat{Y}_{k,1} + \hat{Y}_{k,2}, & \delta = 1/2, \ \alpha \geq 1, \end{cases}$$

$$\hat{S} := \begin{cases} \hat{S}_1, & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ \hat{S}_1 + \hat{S}_2, & \delta = 1/2, \ \alpha \geq 1. \end{cases}$$

The limit processes $\hat{Q}_{k,1} = \{\hat{Q}_{k,1}(t) : t \geq 0\}$, $\hat{Y}_{k,1} = \{\hat{Y}_{k,1}(t) : t \geq 0\}$ *and* $\hat{S}_1 = \{\hat{S}_1(t) : t \geq 0\}$ *are continuous Gaussian processes defined by*

$$\hat{Q}_{k,1}(t) := \sigma(\alpha) \int_0^t \left(\sum_{i=1}^I \pi_i G_k^{(i)}(t-s) \right) dW(s), \quad t \geq 0,$$

$$\hat{Y}_{k,1}(t) := \sigma(\alpha) \int_0^t \left(\sum_{i=1}^I \pi_i (F_k^{(i)}(t-s) - F_{\mathrm{m}}^{(i)}(t-s)) \right) dW(s), \quad t \geq 0,$$

$$\hat{S}_1(t) := \sigma(\alpha) \int_0^t \left(\sum_{i=1}^I \pi_i F_{\mathrm{m}}^{(i)}(t-s) \right) dW(s), \quad t \geq 0,$$

where W is a standard Brownian motion with variance coefficient $\sigma^2(\alpha)$ as defined in Theorem 2.1. The limit processes $\hat{Q}_{k,2} = \{\hat{Q}_{k,2}(t) : t \geq 0\}$, $\hat{Y}_{k,2} = \{\hat{Y}_{k,2}(t) : t \geq 0\}$ and $\hat{S}_2 = \{\hat{S}_2(t) : t \geq 0\}$ are continuous Gaussian processes defined by

$$\hat{Q}_{k,2}(t) := \sum_{i=1}^I \int_0^t \int_{\mathbb{R}_+^K} \mathbf{1}(s + x_k > t) d\hat{\boldsymbol{K}}_i(\pi_i \lambda_i s, \mathbf{x}), \quad t \geq 0,$$

$$\hat{Y}_{k,2}(t) := \sum_{i=1}^I \int_0^t \int_{\mathbb{R}_+^K} \big(\mathbf{1}(s + x_k \leq t) - \mathbf{1}(s + x_j \leq t, \ \forall j) \big) d\hat{\boldsymbol{K}}_i(\pi_i \lambda_i s, \mathbf{x}), \quad t \geq 0,$$

$$\hat{S}_2(t) := \sum_{i=1}^I \int_0^t \int_{\mathbb{R}_+^K} \big(\mathbf{1}(s + x_j \leq t, \ \forall j) \big) d\hat{\boldsymbol{K}}_i(\pi_i \lambda_i s, \mathbf{x}), \quad t \geq 0,$$

where $\hat{\boldsymbol{K}}_i(s, \mathbf{x})$ are independent multiparameter Kiefer processes (Gaussian random field) with mean 0 and covariance $Cov(\hat{\boldsymbol{K}}_i(s, \mathbf{x}), \hat{\boldsymbol{K}}_i(t, \mathbf{y})) = (s \wedge t)(F^{(i)}(\mathbf{x} \wedge \mathbf{y}) - F^{(i)}(\mathbf{x})F^{(i)}(\mathbf{y}))$ for $s, t \geq 0$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^K$. The integrals in $\hat{Q}_{k,2}(t)$, $\hat{Y}_{k,2}(t)$ and $\hat{S}_2(t)$ are defined in the mean squared sense. The Brownian motion W is independent from $\hat{\boldsymbol{K}}_i$, $i = 1, \ldots, I$, and thus, $\hat{Q}_{k,1}$ and $\hat{Q}_{j,2}$ are independent, and so are $\hat{Y}_{k,1}$ and $\hat{Y}_{j,2}$ for each $k, j = 1, \ldots, K$. $\hat{S}_1$ and $\hat{S}_2$ are also independent.

Remark 4.1. We remark that there is also a stochastic decomposition property (as the one we have seen for the infinite-server queueing model). The variabilities in the arrival process and the Markov process are captured in $\hat{Q}_{k,1}$, $\hat{Y}_{k,1}$ and $\hat{S}_1$ for each k, while the variabilities in the service process are captured in $\hat{Q}_{k,2}$, $\hat{Y}_{k,2}$ and $\hat{S}_2$.

Corollary 4.1. *Under the assumptions of Theorem 4.1, the limit process $(\hat{\boldsymbol{Q}}, \hat{\boldsymbol{Y}})$ is a multidimensional Gaussian process with mean zero and covariance functions: for $j, k = 1, \ldots, K$, $t, t' \geq 0$,*

$$Cov(\hat{Q}_j(t), \hat{Q}_k(t')) = \begin{cases} Cov(\hat{Q}_{j,1}(t), \hat{Q}_{k,1}(t')), & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ Cov(\hat{Q}_{j,1}(t), \hat{Q}_{k,1}(t')) + Cov(\hat{Q}_{j,2}(t), \hat{Q}_{k,2}(t')), & \delta = 1/2, \ \alpha \geq 1, \end{cases}$$

$$Cov(\hat{Y}_j(t), \hat{Y}_k(t')) = \begin{cases} Cov(\hat{Y}_{j,1}(t), \hat{Y}_{k,1}(t')), & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ Cov(\hat{Y}_{j,1}(t), \hat{Y}_{k,1}(t')) + Cov(\hat{Y}_{j,2}(t), \hat{Y}_{k,2}(t')), & \delta = 1/2, \ \alpha \geq 1, \end{cases}$$

$$Cov(\hat{Q}_j(t), \hat{Y}_k(t')) = \begin{cases} Cov(\hat{Q}_{j,1}(t), \hat{Y}_{k,1}(t')), & \delta = 1 - \alpha/2, \ 0 < \alpha < 1, \\ Cov(\hat{Q}_{j,1}(t), \hat{Y}_{k,1}(t')) + Cov(\hat{Q}_{j,2}(t), \hat{Y}_{k,2}(t')), & \delta = 1/2, \ \alpha \geq 1, \end{cases}$$

where

$$Cov(\hat{Q}_{j,1}(t), \hat{Q}_{k,1}(t')) = \sigma^2(\alpha) \int_0^{t \wedge t'} \left(\sum_{i=1}^I \pi_i G_j^{(i)}(t-s) \right) \left(\sum_{i=1}^I \pi_i G_k^{(i)}(t'-s) \right) ds,$$

$$Cov(\hat{Y}_{j,1}(t), \hat{Y}_{k,1}(t')) = \sigma^2(\alpha) \int_0^{t \wedge t'} \left(\sum_{i=1}^I \pi_i \left(F_j^{(i)}(t-s) - F_{\mathrm{m}}^{(i)}(t-s) \right) \right) \times \left(\sum_{i=1}^I \pi_i \left(F_k^{(i)}(t'-s) - F_{\mathrm{m}}^{(i)}(t'-s) \right) \right) ds,$$

$$Cov(\hat{Q}_{j,1}(t), \hat{Y}_{k,1}(t')) = \sigma^2(\alpha) \int_0^{t \wedge t'} \left(\sum_{i=1}^I \pi_i G_j^{(i)}(t-s) \right) \left(\sum_{i=1}^I \pi_i \left(F_k^{(i)}(t'-s) - F_{\mathrm{m}}^{(i)}(t'-s) \right) \right) ds,$$

$$Cov(\hat{Q}_{j,2}(t),\hat{Q}_{k,2}(t')) = \sum_{i=1}^{I}\pi_i\lambda_i\int_0^{t\wedge t'}\left(F_{j,k}^{(i)}(t-s,t'-s)-F_j^{(i)}(t-s)F_k^{(i)}(t'-s)\right)ds,$$

$$\begin{aligned}Cov(\hat{Y}_{j,2}(t),\hat{Y}_{k,2}(t')) = \sum_{i=1}^{I}\pi_i\lambda_i\int_0^{t\wedge t'}&\Big(F_{j,k}^{(i)}(t-s,t'-s)-F_j^{(i)}(t-s)F_k^{(i)}(t'-s)\\ &-F_{j,\mathrm{m}}^{(i)}(t-s,t'-s)+F_j^{(i)}(t-s)F_{\mathrm{m}}^{(i)}(t'-s)-F_{k,\mathrm{m}}^{(i)}(t'-s,t-s)\\ &+F_k^{(i)}(t'-s)F_{\mathrm{m}}^{(i)}(t-s)+F_{\mathrm{m}}^{(i)}((t-s)\wedge(t'-s))-F_{\mathrm{m}}^{(i)}(t-s)F_{\mathrm{m}}^{(i)}(t'-s)\Big)ds,\end{aligned}$$

$$\begin{aligned}Cov(\hat{Q}_{j,2}(t),\hat{Y}_{k,2}(t')) = \sum_{i=1}^{I}\pi_i\lambda_i\int_0^{t\wedge t'}&\Big(F_{j,\mathrm{m}}^{(i)}(t-s,t'-s)-F_j^{(i)}(t-s)F_{\mathrm{m}}^{(i)}(t'-s)\\ &-F_{j,k}^{(i)}(t-s,t'-s)+F_j^{(i)}(t-s)F_k^{(i)}(t'-s)\Big)ds,\end{aligned}$$

with $\sigma^2(\alpha)$ being defined in (2.13)*, and for $j=1,\ldots,K$ and $x,y\geq 0$, $F_{j,\mathrm{m}}(x,y):=F(\mathbf{z})$ for $\mathbf{z}\in\mathbb{R}_+^K$ satisfying $z_j=x\wedge y$ and $z_{j'}=y$ for $j'\neq j$.*

In addition, the limit process $\hat{S}$ is a continuous Gaussian process with mean zero and covariance functions: for $t,t'\geq 0$,

$$Cov(\hat{S}(t),\hat{S}(t')) = \begin{cases} Cov(\hat{S}_1(t),\hat{S}_1(t')), & \delta = 1-\alpha/2,\ 0<\alpha<1,\\ Cov(\hat{S}_1(t),\hat{S}_1(t'))+Cov(\hat{S}_2(t),\hat{S}_2(t')), & \delta=1/2,\ \alpha\geq 1,\end{cases}$$

where

$$Cov(\hat{S}_1(t),\hat{S}_1(t')) = \sigma^2(\alpha)\int_0^{t\wedge t'}\left(\sum_{i=1}^{I}\pi_i F_{\mathrm{m}}^{(i)}(t-s)\right)\left(\sum_{i=1}^{I}\pi_i F_{\mathrm{m}}^{(i)}(t'-s)\right)ds,$$

$$Cov(\hat{S}_2(t),\hat{S}_2(t')) = \sum_{i=1}^{I}\pi_i\lambda_i\int_0^{t\wedge t'}\left(F_{\mathrm{m}}^{(i)}((t-s)\wedge(t'-s))-F_{\mathrm{m}}^{(i)}(t-s)F_{\mathrm{m}}^{(i)}(t'-s)\right)ds,$$

with $\sigma^2(\alpha)$ being defined in (2.13)*, and $\lim_{t\to\infty}t^{-1}Var(\hat{S}(t))=\sigma^2(\alpha)$.*

5. Proofs

5.1. **Proof of Theorem 2.1.** In this section, we prove Theorem 2.1. First of all, we write the process $\hat{A}^n$ as

$$\hat{A}^n(t) \;=\; \hat{A}_1^n(t)+\hat{A}_2^n(t),\quad t\geq 0, \tag{5.1}$$

where

$$\hat{A}_1^n(t) := \frac{1}{n^\delta}\left(A^n(t)-\int_0^t\lambda^n_{X^n(s)}ds\right), \tag{5.2}$$

and

$$\hat{A}_2^n(t) := \frac{1}{n^\delta}\left(\int_0^t\lambda^n_{X^n(s)}ds-\sum_{i=1}^{I}\pi_i\lambda_i^n t\right). \tag{5.3}$$

We now focus on proving the convergence of $\hat{A}_1^n$. Without loss of generality, we pick state 1 as the reference state. Let $\tilde{T}_0^n$ be the first time that $X^n(t)$ reaches state 1 from the initial state, and T_k^n be the $(k+1)^{\text{th}}$ jump time of $X^n(t)$ reaching state 1 (i.e., the k^{th} excursion time). Define a counting process associated with the sequence $\{T_k^n : k=1,2,\ldots\}$:

$$N^n(t) := \max\{k : T_k^n\leq t,\ k=0,1,2,\ldots\},\quad t\geq 0,\quad \text{and}\quad T_0^n := 0.$$

Then we can decompose the process $\hat{A}_1^n$ into three processes:

$$\hat{A}_1^n(t) \quad = \quad \hat{A}_{1,1}^n(t) + \hat{A}_{1,2}^n(t) + \hat{A}_{1,3}^n(t), \quad t \geq 0, \tag{5.4}$$

where

$$\hat{A}_{1,1}^n(t) := \frac{1}{n^\delta}\left(A^n(t \wedge \tilde{T}_0^n) - \int_0^{t\wedge\tilde{T}_0^n} \lambda_{X^n(s)}^n ds \right), \tag{5.5}$$

$$\hat{A}_{1,2}^n(t) := \frac{1}{n^\delta}\sum_{k=1}^{N^n(t)} \left(A^n(T_k^n) - A^n(T_{k-1}^n) - \int_{T_{k-1}^n}^{T_k^n} \lambda_{X^n(s)}^n ds \right), \tag{5.6}$$

$$\hat{A}_{1,3}^n(t) := \frac{1}{n^\delta}\left(A^n(t) - A^n(T_{N^n(t)}^n) - \int_{T_{N^n(t)}^n}^{t} \lambda_{X^n(s)}^n ds \right). \tag{5.7}$$

We will prove the convergence of the three processes $\hat{A}_{1,1}^n$, $\hat{A}_{1,2}^n$ and $\hat{A}_{1,3}^n$ in the following lemmas.

Before proving the convergence of the three processes $\hat{A}_{1,1}^n$, $\hat{A}_{1,2}^n$ and $\hat{A}_{1,3}^n$, we present some properties on the processes N^n and the sequence $\{T_k^n : k = 0, 1, 2, \ldots\}$. Let

$$\breve{T}_k^n := T_k^n - T_{k-1}^n$$

for $k = 1, 2, \ldots$. Then $\{\breve{T}_k^n : k = 1, 2, \ldots\}$ forms an i.i.d. sequence of random variables. Let $\gamma^n := E[\breve{T}_1^n]$. It is evident that $\gamma^n < \infty$ and there exists $\gamma > 0$ such that $\gamma^n = n^{-\alpha}\gamma$, since X^n has transition rate matrix $Q^n = n^\alpha Q$. Thus, it follows from the FLLN for delayed renewal processes that

$$\frac{1}{n^\alpha} N^n \Rightarrow \gamma^{-1} e \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \to \infty, \tag{5.8}$$

where $e(t) \equiv t$ for $t \geq 0$.

Lemma 5.1. *For any $\epsilon > 0$ and fixed $T > 0$,*

$$\lim_{n\to\infty} P\left(\sup_{t\in[0,T]} \left|\hat{A}_{1,1}^n(t)\right| > \epsilon \right) = 0. \tag{5.9}$$

Proof. It suffices to show that

$$\lim_{n\to\infty} E\left[\sup_{t\in[0,T]} \left|\hat{A}_{1,1}^n(t)\right| \right] = 0. \tag{5.10}$$

By (5.5), we obtain the following upper bound:

$$\begin{aligned}
E\left[\sup_{t\in[0,T]} \left|\hat{A}_{1,1}^n(t)\right| \right] &\leq \frac{1}{n^\delta} E\left[\sup_{t\in[0,T]} A^n(t \wedge \tilde{T}_0^n) \right] + \frac{1}{n^\delta} E\left[\sup_{t\in[0,T]} \int_0^{t\wedge\tilde{T}_0^n} \lambda_{X^n(s)}^n ds \right] \\
&\leq \frac{1}{n^\delta} E\left[E\left[A^n(T \wedge \tilde{T}_0^n) \;\middle|\; X^n(s) : s \leq T \right] \right] + \frac{1}{n^\delta} E\left[\int_0^{T\wedge\tilde{T}_0^n} \lambda_{X^n(s)}^n ds \right] \\
&= \frac{2}{n^\delta} E\left[\int_0^{T\wedge\tilde{T}_0^n} \lambda_{X^n(s)}^n ds \right] \\
&\leq \frac{2}{n^\delta} \Big(\max_{i\in\mathcal{S}} \lambda_i^n \Big) E[\tilde{T}_0^n].
\end{aligned} \tag{5.11}$$

By Assumption 1, we have that $\frac{1}{n}\max_{i\in\mathcal{S}} \lambda_i^n \to \max_{i\in\mathcal{S}} \lambda_i < \infty$ as $n \to \infty$. Since $Q^n = n^\alpha Q$, it is evident that $E[\tilde{T}_0^n] = O(1/n^\alpha)$ (see, e.g., [21, pp. 256–257]). Thus it follows that

$$\frac{2}{n^\delta} \Big(\max_{i\in\mathcal{S}} \lambda_i^n \Big) E[\tilde{T}_0^n] \to 0 \quad \text{as} \quad n \to \infty,$$

and we have proved (5.10). □

Lemma 5.2. *For any $\epsilon > 0$ and fixed $T > 0$,*

$$\lim_{n\to\infty} P\left(\sup_{t\in[0,T]} \left|\hat{A}^n_{1,3}(t)\right| > \epsilon\right) = 0. \tag{5.12}$$

Proof. For each $k = 1, 2, 3, \ldots$, define

$$\breve{A}^n_k := \sup_{0\le t\le \breve{T}^n_k} \frac{1}{n^\delta}\left|A^n(T^n_{k-1}+t) - A^n(T^n_{k-1}) - \int_{T^n_{k-1}}^{T^n_{k-1}+t} \lambda^n_{X^n(s)} ds\right|. \tag{5.13}$$

To prove (5.12), it suffices to prove that

$$\lim_{n\to\infty} E\big[\breve{A}^n_{N(T)+1}\big] = 0. \tag{5.14}$$

By (5.13) and conditioning, we obtain that

$$\begin{aligned} E\big[\breve{A}^n_{N^n(T)+1}\big] &\le \frac{1}{n^\delta} E\left[\left|A^n(T^n_{N^n(T)+1}) - A^n(T^n_{N^n(T)})\right|\right] + \frac{1}{n^\delta} E\left[\int_{T^n_{N^n(T)}}^{T^n_{N^n(T)+1}} \lambda^n_{X^n(s)} ds\right] \\ &\le \frac{2}{n^\delta} E\left[\int_{T^n_{N^n(T)}}^{T^n_{N^n(T)+1}} \lambda^n_{X^n(s)} ds\right] \le \frac{2}{n^\delta}\Big(\max_{i\in\mathcal{S}} \lambda^n_i\Big) E[\breve{T}^n_1] \to 0 \quad \text{as} \quad n\to\infty, \end{aligned}$$

where the convergence follows from Assumption 1 and $E[\breve{T}^n_1] = n^{-\alpha}\gamma$. Thus, the lemma is proved. □

Lemma 5.3.

$$\hat{A}^n_{1,2} \Rightarrow \begin{cases} 0, & \delta = 1-\alpha/2,\ 0<\alpha<1, \\ \hat{A}_1, & \delta = 1/2,\ \alpha \ge 1, \end{cases} \tag{5.15}$$

in $\mathbb{D}$ as $n\to\infty$, where the limit process $\hat{A}_1$ is a driftless Brownian motion with variance coefficient $\bar{\nu}$ defined in (2.14).

To prove this lemma, we need the following lemma, whose proof follows from a direct generalization of Theorem 2.7 in [22].

Lemma 5.4. *Let $\{\xi_{n,i} : i \ge 1\}$ be an i.i.d. sequence for each n and $U_n(t) := \sum_{i=1}^{\lfloor n^\alpha t\rfloor} \xi_{n,i}$ for each $t \ge 0$ and any $\alpha > 0$. Then $U_n \Rightarrow U$ in $\mathbb{D}$ as $n\to\infty$ where U is a stochastic process with stationary independent increments if and only if $U_n(t) \Rightarrow U(t)$ in $\mathbb{R}$ for each t as $n\to\infty$.*

Proof of Lemma 5.3. Define a process $\breve{A}^n_{1,2} = \{\breve{A}^n_{1,2}(t) : t \ge 0\}$ by

$$\breve{A}^n_{1,2}(t) := \sum_{k=1}^{\lfloor n^\alpha t/\gamma\rfloor} \left(\breve{A}^n_k - \frac{1}{n^\delta}\int_{T^n_{k-1}}^{T^n_k} \lambda^n_{X^n(s)} ds\right), \quad t \ge 0, \tag{5.16}$$

where $\breve{A}^n_k$ is defined in (5.13). We first show that, for each $t \ge 0$,

$$\breve{A}^n_{1,2}(t) \Rightarrow \begin{cases} 0, & \delta = 1-\alpha/2,\ 0<\alpha<1, \\ \breve{A}(t), & \delta = 1/2,\ \alpha \ge 1, \end{cases} \tag{5.17}$$

in $\mathbb{R}$ as $n\to\infty$, where $\breve{A}(t)$ has a normal distribution with mean 0 and variance $\bar{\nu}t$, with $\bar{\nu}$ defined in (2.14). This follows from applying CLT for doubly indexed sequences by noting that the summation terms in (5.16) are i.i.d. for each given n. It suffices to show that, as $n\to\infty$,

$$n^\alpha Var\left(\breve{A}^n_1 - \frac{1}{n^\delta}\int_{\tilde{T}^n_0}^{T^n_1} \lambda^n_{X^n(s)} ds\right) \to \begin{cases} 0, & \delta = 1-\alpha/2,\ 0<\alpha<1, \\ \bar{\nu}\gamma, & \delta = 1/2,\ \alpha \ge 1. \end{cases} \tag{5.18}$$

By conditioning, we obtain

$$\begin{aligned}
& Var\left(\breve{A}_1^n - \frac{1}{n^\delta}\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) \\
= \;& Var\big(\breve{A}_1^n\big) + Var\left(\frac{1}{n^\delta}\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) - 2Cov\left(\breve{A}_1^n, \frac{1}{n^\delta}\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) \\
= \;& E\big[Var\big(\breve{A}_1^n | X^n(s) : \tilde{T}_0^n \le s \le T_1^n\big)\big] + Var\big(E\big[\breve{A}_1^n | X^n(s) : \tilde{T}_0^n \le s \le T_1^n\big]\big) \\
& + \frac{1}{n^{2\delta}} Var\left(\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) - 2\frac{1}{n^\delta}\left(E\left[\breve{A}_1^n \int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right] - E\big[\breve{A}_1^n\big]E\left[\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right]\right) \\
= \;& \frac{1}{n^{2\delta}} E\left[\int_{\tilde{T}_0^n}^{T_1^n} \nu^n_{X^n(s)}ds\right] + \frac{2}{n^{2\delta}} Var\left(\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) - \frac{2}{n^{2\delta}} Var\left(\int_{\tilde{T}_0^n}^{T_1^n} \lambda^n_{X^n(s)}ds\right) \\
= \;& \frac{1}{n^{2\delta}} E\left[\int_{\tilde{T}_0^n}^{T_1^n} \nu^n_{X^n(s)}ds\right] = \frac{1}{n^{2\delta}} \sum_{i=1}^{I} \nu_i^n E\left[\int_{\tilde{T}_0^n}^{T_1^n} \mathbf{1}(X^n(s) = i)ds\right]. \qquad (5.19)
\end{aligned}$$

Under the assumption of the underlying Markov process X^n, we obtain that for each $i = 1, \ldots, I$ and $t \ge 0$,

$$\int_0^t \mathbf{1}(X^n(s) = i)ds \Rightarrow \pi_i t \quad \text{as} \quad n \to \infty. \qquad (5.20)$$

Since $E[\breve{T}_1^n] = n^{-\alpha}\gamma$, we obtain that as $n \to \infty$,

$$n^{\alpha-2\delta} \sum_{i=1}^{I} \nu_i^n E\left[\int_{\tilde{T}_0^n}^{T_1^n} \mathbf{1}(X^n(s) = i)ds\right] \to \begin{cases} 0, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \bar{\nu}\gamma, & \delta = 1/2,\ \alpha \ge 1. \end{cases} \qquad (5.21)$$

Thus, we have proved (5.17). By Lemma 5.4, we obtain that

$$\breve{A}^n_{1,2} \Rightarrow \begin{cases} 0, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \hat{A}_1, & \delta = 1/2,\ \alpha \ge 1, \end{cases} \qquad (5.22)$$

in $\mathbb{D}$ as $n \to \infty$. Now by (5.8), Theorem 11.4.5 of [25] and the continuous mapping theorem, we can conclude the convergence in (5.15). □

Completing the Proof of Theorem 2.1. Recall the representation of the process $\hat{A}^n$ in (5.1)–(5.3) and (5.4)–(5.7). By Lemmas 5.1 and 5.2, we obtain that $\hat{A}^n_{1,1} \Rightarrow 0$ and $\hat{A}^n_{1,3} \Rightarrow 0$ as $n \to \infty$, respectively. By Lemma 5.3, we obtain that (i) $\hat{A}^n_{1,2} \Rightarrow \hat{A}_1$ in $\mathbb{D}$ as $n \to \infty$, when $\delta = 1/2$ and $\alpha \ge 1$, where $\hat{A}_1$ is a driftless Brownian motion with variance parameter $\bar{\nu}$, and (ii) $\hat{A}^n_{1,2} \Rightarrow 0$ in $\mathbb{D}$ as $n \to \infty$, when $\delta = 1 - \alpha/2$ and $0 < \alpha < 1$.

By Proposition 3.2 in [1], we obtain that

$$\hat{A}^n_2 \Rightarrow \begin{cases} \hat{A}_2, & \delta = 1 - \alpha/2,\ 0 < \alpha < 1, \\ \hat{A}_2, & \delta = 1/2,\ \alpha = 1, \\ 0, & \delta = 1/2,\ \alpha > 1, \end{cases} \qquad (5.23)$$

in $\mathbb{D}$ as $n \to \infty$, where the limit process $\hat{A}_2 = \{\hat{A}_2(t) : t \ge 0\}$ is a Brownian motion with mean 0 and variance coefficient $\bar{\beta}$. Here, the Brownian motion $\hat{A}_1$ is independent of $\hat{A}_2$. Thus the proof is complete. □

5.2. **Proofs for applications to infinite-server queues.**

Proof of Theorem 3.1. We first note that the process Q^n can be written as

$$\begin{aligned} Q^n(t) &= \sum_{k=1}^{A^n(t)} \sum_{i=1}^{I} \mathbf{1}(\tau_k^n + \eta_{k,i} > t)\mathbf{1}(X^n(\tau_k^n) = i) \\ &= \int_0^t \int_0^\infty \sum_{i=1}^{I} \mathbf{1}(s + x_i > t)\mathbf{1}(X^n(s) = i) d\left(\sum_{k=1}^{A^n(s)} \mathbf{1}(\eta_{k,i} \le x_i) \right), \quad t \ge 0. \end{aligned} \tag{5.24}$$

From this, we obtain the following representation for the diffusion-scaled process $\hat{Q}^n$: $\hat{Q}^n(t) = \hat{Q}_1^n(t) + \hat{Q}_2^n(t)$ for $t \ge 0$, where

$$\hat{Q}_1^n(t) := \int_0^t F^c_{X^n(s)}(t-s) d\hat{A}^n(s) = \int_0^t \sum_{i=1}^{I} F_i^c(t-s)\mathbf{1}(X^n(s) = i) d\hat{A}^n(s), \tag{5.25}$$

and

$$\hat{Q}_2^n(t) := n^{1/2-\delta} \sum_{i=1}^{I} \int_0^t \int_0^\infty \mathbf{1}(s + x_i > t)\mathbf{1}(X^n(s) = i) d\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{A^n(s)} \big(\mathbf{1}(\eta_{k,i} \le x_i) - F_i(x_i) \big) \right). \tag{5.26}$$

We next prove the convergences of $\hat{Q}_1^n$ and $\hat{Q}_2^n$.

To prove the convergence of $\hat{Q}_1^n$, we show that

$$\lim_{n\to\infty} P\left(\sup_{0\le t\le T} \big| \hat{Q}_1^n(t) - \hat{Q}_1(t) \big| > \epsilon \right) = 0. \tag{5.27}$$

Note that

$$\begin{aligned} & P\left(\sup_{0\le t\le T} \big| \hat{Q}_1^n(t) - \hat{Q}_1(t) \big| > \epsilon \right) \\ \le\ & P\left(\sup_{0\le t\le T} \left| \int_0^t \sum_{i=1}^{I} F_i^c(t-s)\mathbf{1}(X^n(s) = i) d\big(\hat{A}^n(s) - \hat{A}(s) \big) \right| > \epsilon \right) \\ & + P\left(\sup_{0\le t\le T} \left| \int_0^t \sum_{i=1}^{I} F_i^c(t-s) \big(\mathbf{1}(X^n(s) = i) - \pi_i \big) d\hat{A}(s) \right| > \epsilon \right), \end{aligned} \tag{5.28}$$

where $\hat{A}$ is the limit process of the arrivals $\hat{A}^n$ as given in Theorem 2.1. The convergence to zero of the first term on the right-hand side of (5.28) follows from the convergence of $\hat{A}^n \Rightarrow \hat{A}$ in Theorem 2.1. To prove the convergence of the second term in (5.28), we first observe that the process $\hat{Q}_{1,2}^n = \{\hat{Q}_{1,2}^n(t) : t \ge 0\}$ defined by

$$\hat{Q}_{1,2}^n(t) := \int_0^t \sum_{i=1}^{I} F_i^c(t-s) \big(\mathbf{1}(X^n(s) = i) - \pi_i \big) d\hat{A}(s), \quad t \ge 0,$$

is a Markov process. It is easy to check that the generators of the processes $\hat{Q}_{1,2}^n$ converge to zero. Thus, by [10, Ch. IV, Thm. 2.5], we obtain the convergence of the second term in (5.28). To show the joint convergence

$$(\hat{A}^n, \hat{Q}_1^n) \Rightarrow (\hat{A}, \hat{Q}_1) \quad \text{in} \quad \mathbb{D}^2 \quad \text{as} \quad n \to \infty, \tag{5.29}$$

by endowing the product space with the maximum metric, we see that the convergence of $\hat{A}^n$ by assumption and $\hat{Q}_1^n$ in (5.27), as well as the continuity of their limits $\hat{A}$ and $\hat{Q}_1$, imply that (5.29) holds.

Next we will show the convergence of $\hat{Q}_2^n$. Define the sequential empirical processes $\hat{K}_i^n = \{\hat{K}_i^n(t,x) : t,x \geq 0\}$ by

$$\hat{K}_i^n(t,x) := \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nt \rfloor} (\mathbf{1}(\eta_{k,i} \leq x) - F_i(x)), \quad t,x \geq 0,$$

for each $i = 1, \ldots, I$. By [13, Lemma 3.1] and the independence of $\hat{K}_i^n$, $i = 1, \ldots, I$, we know

$$\hat{K}_i^n \Rightarrow \hat{K}_i \quad \text{in} \quad \mathbb{D}_{\mathbb{D}} \quad \text{as} \quad n \to \infty, \tag{5.30}$$

where $\hat{K}_i$, $i = 1, \ldots, I$, are independent Kiefer processes defined in Theorem 3.1. We let $A_i^n = \{A_i^n(t) : t \geq 0\}$ be the process counting the number of arrivals whose service type is i, i.e.,

$$A_i^n(t) := \max\left\{ j \geq 0 : \tau_{k_j}^n \mathbf{1}(X^n(\tau_{k_j}^n) = i) \leq t \right\}, \quad t \geq 0, \tag{5.31}$$

where $k_0 := 0$ and $\tau_0^n := 0$, for $i = 1, \ldots, I$. Define the fluid-scaled processes $\bar{A}_i^n := n^{-1} A_i^n$ for each $i = 1, \ldots, I$. Thus, Theorem 2.1 directly implies the functional weak law of large numbers (FWLLNs) for A_i^n, i.e.,

$$(\bar{A}_1^n, \ldots, \bar{A}_I^n) \Rightarrow (\pi_1 \lambda_1 e, \ldots, \pi_I \lambda_I e) \quad \text{in} \quad \mathbb{D}^I \quad \text{as} \quad n \to \infty. \tag{5.32}$$

We can rewrite (5.26) as

$$\hat{Q}_2^n(t) = -n^{1/2-\delta} \sum_{i=1}^{I} \hat{Q}_{2,i}^n(t), \quad t \geq 0, \tag{5.33}$$

where the processes $\hat{Q}_{2,i}^n := \{\hat{Q}_{2,i}^n(t) : t \geq 0\}$ are defined by

$$\hat{Q}_{2,i}^n(t) := \int_0^t \int_0^\infty \mathbf{1}(s + x_i \leq t) d\hat{K}_i^n(\bar{A}_i^n(s), x_i), \quad t \geq 0. \tag{5.34}$$

Tightness of the processes $\{\hat{Q}_{2,i}^n : n \geq 1\}$ in $\mathbb{D}$ follows directly from the tightness of the corresponding processes for the $G/GI/\infty$ queues in [13], for $i = 1, \ldots, I$. Thus, we obtain the processes $\{\hat{Q}_2^n : n \geq 1\}$ are tight.

We now focus on proving the joint convergence of finite-dimensional distributions of $\hat{Q}_1^n$ and $\hat{Q}_2^n$. We only need to show the case $\delta = 1/2$, since otherwise the limit $\hat{Q}_2$ vanishes. Define the process $\hat{Q}_{2,i} = \{\hat{Q}_{2,i}(t) : t \geq 0\}$ by

$$\hat{Q}_{2,i}(t) := \int_0^t \int_0^\infty \mathbf{1}(s + x_i \leq t) d\hat{K}_i(\pi_i \lambda_i s, x_i), \tag{5.35}$$

for $t \geq 0$ and $i = 1, \ldots, I$. The integral $\hat{Q}_{2,i}$ in (5.35) is understood as a mean-square integral. Specifically, we define

$$\hat{Q}_{2,i}(t) := \text{l.i.m}_{l \to \infty} \hat{Q}_{2,i,l}(t), \quad t \geq 0, \tag{5.36}$$

where l.i.m. represents mean-square limit, that is,

$$\lim_{l \to \infty} E\left[\left(\hat{Q}_{2,i}(t) - \hat{Q}_{2,i,l}(t) \right)^2 \right] = 0,$$

and

$$\hat{Q}_{2,i,l}(t) := \int_0^t \int_0^\infty \mathbf{1}_{l,t}(s,x) d\hat{K}_i(\pi_i \lambda_i s, x_i) = \sum_{j=1}^{l} \Delta_{\hat{K}_i}((\pi_i \lambda_i s_{j-1}^l, 0); (\pi_i \lambda_i s_j^l, t - s_j^l)),$$

with $\mathbf{1}_{l,t}(\cdot,\cdot)$ defined by

$$\mathbf{1}_{l,t}(s,x) := \mathbf{1}(s=0)\mathbf{1}(x \le t) + \sum_{j=1}^{l} \mathbf{1}(s_{j-1}^l < s \le s_j^l)\mathbf{1}(x \le t - s_j^l),$$

with the points $0 = s_1^l < s_2^l < \cdots < s_l^l = t$ being chosen so that $\max_{1\le j\le l} |s_j^l - s_{j-1}^l| \to 0$ as $l \to \infty$, and for $a_1 \le a_2$, $b_1 \le b_2$ and $i = 1, \ldots, I$,

$$\Delta_{\hat{K}_i}((a_1,b_1);(a_2,b_2)) = \hat{K}_i(a_2,b_2) - \hat{K}_i(a_1,b_2) - \hat{K}_i(a_2,b_1) + \hat{K}_i(a_1,b_1). \tag{5.37}$$

We define additional processes $\hat{Q}_{2,i,l}^n = \{\hat{Q}_{2,i,l}^n(t) : t \ge 0\}$ and $\breve{Q}_{2,i,l}^n = \{\breve{Q}_{2,i,l}^n(t) : t \ge 0\}$ by

$$\hat{Q}_{2,i,l}^n(t) := \int_0^t \int_0^\infty \mathbf{1}_{l,t}(s,x) d\hat{K}_i^n(\bar{A}_i^n(s), x_i) = \sum_{j=1}^{l} \Delta_{\hat{K}_i^n}((\bar{A}^n(s_{j-1}^l), 0); (\bar{A}^n(s_j^l), t - s_j^l)),$$

$$\breve{Q}_{2,i,l}^n(t) := \int_0^t \int_0^\infty \mathbf{1}_{l,t}(s,x) d\hat{K}_i^n(\pi_i\lambda_i s, x_i) = \sum_{j=1}^{l} \Delta_{\hat{K}_i^n}((\pi_i\lambda_i s_{j-1}^l, 0); (\pi_i\lambda_i s_j^l, t - s_j^l)),$$

where $\Delta_{\hat{K}_i^n}$ is defined similar to $\Delta_{\hat{K}_i}$ in (5.37) with $\hat{K}_i$ replaced by $\hat{K}_i^n$.

By the weak convergence of $\hat{K}_i^n$ to $\hat{K}_i$ in $\mathbb{D}_{\mathbb{D}}$ as $n \to \infty$, we easily obtain that, for $i = 1, \ldots, I$,

$$\breve{Q}_{2,i,l}^n \xrightarrow{f.d.d.} \hat{Q}_{2,i,l} \quad \text{as} \quad n \to \infty,$$

where $\xrightarrow{f.d.d.}$ stands for the convergence in finite-dimensional distributions. By noting that $\breve{Q}_{2,i,l}^n$, $i = 1, \ldots, I$, and A^n are independent from each other, together with (5.29), we have

$$(\hat{A}^n, \hat{Q}_1^n, \breve{Q}_{2,1,l}^n, \ldots, \breve{Q}_{2,I,l}^n) \xrightarrow{f.d.d.} (\hat{A}, \hat{Q}_1, \hat{Q}_{2,1,l}, \ldots, \hat{Q}_{2,I,l}) \quad \text{as} \quad n \to \infty.$$

In order to establish the joint convergence of $\hat{A}^n$, $\hat{Q}_1^n$ and $\hat{Q}_{2,i}^n$ in finite-dimensional distributions, $i = 1, \ldots, I$, i.e.,

$$(\hat{A}^n, \hat{Q}_1^n, \hat{Q}_{2,1}^n, \ldots, \hat{Q}_{2,I}^n) \xrightarrow{f.d.d.} (\hat{A}, \hat{Q}_1, \hat{Q}_{2,1}, \ldots, \hat{Q}_{2,I}) \quad \text{as} \quad n \to \infty, \tag{5.38}$$

it is sufficient to show the following: for any $T > 0$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{0\le t\le T} |\hat{Q}_{2,i,l}^n(t) - \breve{Q}_{2,i,l}^n(t)| > \epsilon\right) = 0, \quad i = 1, \ldots, I, \tag{5.39}$$

and, for $t > 0$ and $\epsilon > 0$,

$$\lim_{l\to\infty} \lim_{n\to\infty} P\left(|\hat{Q}_{2,i,l}^n(t) - \hat{Q}_{2,i}^n| > \epsilon\right) = 0, \quad i = 1, \ldots, I. \tag{5.40}$$

We can easily obtain (5.39) from (5.30) and (5.32), as well as the continuity of $\hat{K}_i$, $i = 1, \ldots, I$. Following the proof of [13, Lemma 5.3], we immediately see that (5.40) also holds. Therefore, we have shown (5.38).

By the continuous mapping theorem, together with (5.33) and $\delta = 1/2$, we further have

$$(\hat{A}^n, \hat{Q}_1^n, \hat{Q}_2^n) \xrightarrow{f.d.d.} (\hat{A}, \hat{Q}_1, \hat{Q}_2) \quad \text{as} \quad n \to \infty.$$

Since $\{\hat{Q}_1^n : n \ge 1\}$ and $\{\hat{Q}_2^n : n \ge 1\}$ are tight as previously shown, we have established the weak convergence of $\hat{Q}^n$ joint with $\hat{A}^n$ when $\delta = 1/2$ and $\alpha \ge 1$. Furthermore, by noting (3.1) and (3.2), as well as the continuous mapping theorem, we obtain the weak convergence of $(\hat{A}^n, \hat{Q}^n, \hat{D}^n)$ jointly. The case when $\delta = 1 - \alpha/2$ and $0 < \alpha < 1$ can be obtained analogously by noting that the limit $\hat{Q}_2^n$ vanishes as $n \to \infty$. Therefore, the proof of Theorem 3.1 is complete. □

Proof of Corollary 3.1. The covariance functions of $\hat{Q}$ can be obtained similarly to [13, Lemma 5.1] in combination with Itô isometry as well as the fact that the Kiefer processes $\hat{K}_i$, with $i = 1, \ldots, I$, and the arrival limit $\hat{A}$ are independent of each other. The covariance functions of $\hat{D}$ can also be derived similarly. We omit the details here for brevity. □

5.3. **Proofs for applications to fork-join networks.**

Proof Sketch of Theorem 4.1. We first note that the processes Q_k^n, Y_k^n and S can be represented as

$$\begin{aligned}
Q_k^n(t) &= \sum_{\ell=1}^{A^n(t)} \sum_{i=1}^{I} \mathbf{1}(\tau_\ell^n + \eta_k^{\ell,i} > t)\mathbf{1}(X^n(\tau_\ell^n) = i) \\
&= \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} \mathbf{1}(s + x_k^i > t)\mathbf{1}(X^n(s) = i) d\left(\sum_{\ell=1}^{A^n(s)} \mathbf{1}(\boldsymbol{\eta}^{\ell,i} \le \mathbf{x}^i) \right), \\
Y_k^n(t) &= \sum_{\ell=1}^{A^n(t)} \sum_{i=1}^{I} (\mathbf{1}(\tau_\ell^n + \eta_k^{\ell,i} \le t) - \mathbf{1}(\tau_\ell^n + \eta_j^{\ell,i} \le t,\ \forall j))\mathbf{1}(X^n(\tau_\ell^n) = i) \\
&= \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} (\mathbf{1}(s + x_k^i \le t) - \mathbf{1}(s + x_j^i \le t,\ \forall j))\mathbf{1}(X^n(s) = i) d\left(\sum_{\ell=1}^{A^n(s)} \mathbf{1}(\boldsymbol{\eta}^{\ell,i} \le \mathbf{x}^i) \right), \\
S^n(t) &= \sum_{\ell=1}^{A^n(t)} \sum_{i=1}^{I} \mathbf{1}(\tau_\ell^n + \eta_j^{\ell,i} \le t,\ \forall j)\mathbf{1}(X^n(\tau_\ell^n) = i) \\
&= \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} \mathbf{1}(s + x_j^i \le t,\ \forall j)\mathbf{1}(X^n(s) = i) d\left(\sum_{\ell=1}^{A^n(s)} \mathbf{1}(\boldsymbol{\eta}^{\ell,i} \le \mathbf{x}^i) \right).
\end{aligned}$$

Then we can obtain the representations for the diffusion-scaled processes $\hat{Q}_k^n$, $\hat{Y}_k^n$ and $\hat{S}^n$ as follows:

$$\begin{aligned}
\hat{Q}_k^n(t) &= \sum_{i=1}^{I} \int_0^t G_k^{(i)}(t-s)\mathbf{1}(X^n(s) = i) d\hat{A}^n(s) \\
&\quad + n^{1/2-\delta} \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} \mathbf{1}(s + x_k^i > t) d\hat{\boldsymbol{K}}_i^n(\bar{A}_i^n(s), \mathbf{x}^i), \qquad (5.41) \\
\hat{Y}_k^n(t) &= \sum_{i=1}^{I} \int_0^t (F_k^{(i)}(t-s) - F_{\mathrm{m}}^{(i)}(t-s))\mathbf{1}(X^n(s) = i) d\hat{A}^n(s) \\
&\quad + n^{1/2-\delta} \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} (\mathbf{1}(s + x_k^i \le t) - \mathbf{1}(s + x_j^i \le t,\ \forall j)) d\hat{\boldsymbol{K}}_i^n(\bar{A}_i^n(s), \mathbf{x}^i), \qquad (5.42) \\
\hat{S}^n(t) &= \sum_{i=1}^{I} \int_0^t F_{\mathrm{m}}^{(i)}(t-s)\mathbf{1}(X^n(s) = i) d\hat{A}^n(s) \\
&\quad + n^{1/2-\delta} \sum_{i=1}^{I} \int_0^t \int_{\mathbb{R}_+^K} \mathbf{1}(s + x_j^i \le t,\ \forall j) d\hat{\boldsymbol{K}}_i^n(\bar{A}_i^n(s), \mathbf{x}^i), \qquad (5.43)
\end{aligned}$$

where the processes A_i^n, $i = 1, \ldots, I$, are defined in (5.31), and the multiparameter sequential empirical processes $\hat{\boldsymbol{K}}_i^n = \{\hat{K}_i^n(t, \mathbf{x}) : t \geq 0, \mathbf{x} \in \mathbb{R}_+^K\}$ are defined by

$$\hat{\boldsymbol{K}}_i^n(t, \mathbf{x}) := \frac{1}{\sqrt{n}} \sum_{\ell=1}^{\lfloor nt \rfloor} (\mathbf{1}(\boldsymbol{\eta}^{\ell,i} \leq \mathbf{x}) - F^{(i)}(\mathbf{x})), \quad t \geq 0, \ \mathbf{x} \in \mathbb{R}_+^K.$$

The weak convergence of the first terms in (5.41)–(5.43) follows analogously from the proof for (5.25) in Theorem 3.1. Note from [14, Thm. 3.1] and the independence of $\hat{\boldsymbol{K}}_i^n$, $i = 1, \ldots, I$, that

$$\hat{\boldsymbol{K}}_i^n \Rightarrow \hat{\boldsymbol{K}}_i \quad \text{in} \quad \mathbb{D}([0, \infty), \mathbb{D}_K) \quad \text{as} \quad n \to \infty, \tag{5.44}$$

where $\hat{\boldsymbol{K}}_i$, $i = 1, \ldots, I$, are independent generalized Kiefer processes with covariance functions in Theorem 4.1. With similar argument to the proof in Theorem 3.1 and [14, Section 6.2], we can also show the weak convergence of the second terms in (5.41)–(5.43), as well as the joint convergence of $(\hat{A}, \hat{\boldsymbol{Q}}, \hat{\boldsymbol{Y}}, \hat{S})$. The details are omitted for brevity. □

Proof of Corollary 4.1. The covariance functions of $\hat{Q}_j(t)$ and $\hat{Y}_k(t')$, and $\hat{S}(t)$ and $\hat{S}(t')$ are analogous to [14, Theorem 3.4] for $j, k = 1, \ldots, K$, and $t, t' \geq 0$, together with the fact that the generalized Kiefer processes $\hat{\boldsymbol{K}}_i$'s are independent, $i = 1, ..., I$. We omit the details for brevity. □

6. Concluding Remarks

We have studied a large class of MAAPs that can capture more burstiness and variabilities than MMPPs. Under mild conditions on the parameters, we have established an FCLT for the MAAPs. The FCLT is applied to non-Markovian infinite-server systems and fork-join networks with NES. It can be also similarly applied to obtain two-parameter heavy-traffic limits for infinite-server systems as in [20]. The FCLT can be potentially applied to study large-scale service systems in Markov random environments, for example, queueing networks in which all stations are modulated by the same Markov process. The results can be also used to study resource allocation and system design problems for such queueing and network models. It may be also interesting to study the (sample-path) large deviation problems for queueing systems with MAAPs.

Acknowledgments. Hongyuan Lu and Guodong Pang acknowledge the support from the NSF grant CMMI-1538149. Michel Mandjes acknowledges the support from Gravitation project NETWORKS, grant number 024.002.003, funded by the Netherlands Organization for Scientific Research (NWO).

References

[1] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck. (2016) A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability.* Vol. 18, No. 1, 153–168.

[2] S. Asmussen. (2003) *Applied Probability and Queues.* 2nd edition. Springer, Berlin.

[3] M. Baykal-Gursoy and W. Xiao. (2004) Stochastic decomposition in $M/M/\infty$ queues with Markov modulated service rates. *Queueing Systems.* Vol. 48, No.1, 75–88.

[4] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir. (2014) Markov-modulated infinite-server queues with general service times. *Queueing Systems.* Vol. 76, No. 4, 403–424.

[5] J. Blom, M. Mandjes, and H. Thorsdottir. (2013) Time-scaling limits for Markov-modulated infinite-server queues. *Stochastic Models.* Vol. 29, No.1, 112–127.

[6] J. Blom, K. de Turck, and M. Mandjes. (2015) Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences.* Vol. 29, No. 3, 433–459.

[7] J. Blom, K. de Turck, and M. Mandjes. (2016) Functional central limit theorems for Markov-modulated infinite-server systems. *Mathematical Methods of Operations Research.* Vol. 83, No. 3, 351–372.

[8] P. Billingsley. (2009) *Convergence of Probability Measures.* Wiley, New York.

[9] B. D'Auria. (2007) Stochastic decomposition of the $M/G/\infty$ queue in a random environment. *Operations Research Letters.* Vol. 35, No. 6, 805–812.

[10] S. N. Ethier and T. G. Kurtz. (2009) *Markov Processes: Characterization and Convergence.* Wiley, New York.
[11] G. Falin. (2008) The $M/M/\infty$ queue in a random environment. *Queueing Systems.* Vol. 58, 65–76.
[12] J. Keilson and L. Servi. (1993) The matrix $M/M/\infty$ system: retrial models and Markov modulated sources. *Advances in Applied Probability.* Vol. 25, 453–471.
[13] 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.
[14] H. Lu and G. Pang. (2015a) Gaussian Limits for a fork-join network with non-exchangeable synchronization in heavy traffic. *Mathematics of Operations Research.* Vol. 41, No. 2, 560–595.
[15] H. Lu and G. Pang. (2015b) Heavy-traffic Limits for an infinite-server fork-join network with dependent and disruptive services. Submitted.
[16] H. Lu and G. Pang. (2015c) Heavy-traffic Limits for a fork-join network in the Halfin-Whitt regime. Submitted.
[17] A. Nazarov and G. Baymeeva. (2014) The $M/G/\infty$ queue in a random environment. *A. Dudlin et al. (Eds.): ITMM 2014, CCIS 487*, 312–324.
[18] G. Neuhaus. (1971) On weak convergence of stochastic processes with multidimensional time parameter. *Annals of Mathematical Statistics.* Vol. 42, No. 4, 1285–1295.
[19] C. O'Cinneide and P. Purdue. (1986) The $M/M/\infty$ queue in a random environment. *Journal of Applied Probability.* Vol. 23, No. 1, 175–184.
[20] G. Pang and W. Whitt. (2010) Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems.* Vol. 65, No. 4, 325–364
[21] S. M. Ross. (1996) *Stochastic Processes.* 2nd ed. John Wiley & Sons, Inc.
[22] A. V. Skorohod. (1957) Limit theorems for stochastic processes with independent increments. *Theory Probab. Appl.* 2, 138–171.
[23] J. L. Steichen. (2001) A functional central limit theorem for Markov additive processes with an application to the closed Lu-Kumar network. *Stochastic Models.* 17(4), 459–489.
[24] M. L. Straf. (1972) Weak convergence of stochastic processes with several parameters. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2, 187–221.
[25] W. Whitt. (2002) *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Applications to Queues.* Springer, Berlin.
[26] W. Whitt. (2002) *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Applications to Queues. Online Supplement.*

The Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, College of Engineering, Pennsylvania State University, University Park, PA 16802
E-mail address: hzl142@psu.edu, gup3@psu.edu

Korteweg-de Vries Institute (KdVI) for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

CWI, Science Park 123, P. O. Box 94079, 1090 GB Amsterdam, The Netherlands.
E-mail address: m.r.h.mandjes@uva.nl