Queueing Systems manuscript No.
(will be inserted by the editor)

March 27, 2014

A Functional Weak Law of Large Numbers for the Time-Varying $(G_t/GI/s_t + GI)^m/M_t$ Queueing Network

A. Korhan Aras · Yunan Liu

Received: date / Accepted: date

Abstract A many-server heavy-traffic functional law of large numbers is established for the $(G_t/GI/s_t + GI)^m/M_t$ open queueing network, with a finite number of queues (the superscript m), non-stationary non-Poisson external arrival processes (the G_t), non-exponential service times (the first GI), time-varying staffing levels (the s_t), and customer abandonment following non-exponential patience times (the $+GI$). Upon service completion, customers are either routed to one of the queues in the network or out of the system according to time-dependent probabilities (the M_t). The limit provides support for a previously proposed deterministic fluid approximation and extends a previously established limit for the $G_t/GI/s_t + GI$ single queue model.

Keywords Functional weak law of large numbers · deterministic fluid limit · many-server heavy-traffic limit · time-varying arrivals · non-stationary queues · customer abandonment · open queueing network · non-Markovian queues · probabilistic routing

1 Introduction

Many large-scale service systems arising in customer contact centers, communication networks, and healthcare systems can be viewed as networks of multi-server queues [1, 2, 5, 10]. The successful design and management of these

A. Korhan Aras
Operations Research Graduate Program, North Carolina State University, Raleigh, NC, 27695-7913, USA
E-mail: akaras@ncsu.edu

Yunan Liu
Department of Industrial and System Engineering, North Carolina State University, Raleigh, NC 27607-7906, USA
Tel.: +919-513-7208
E-mail: yliu48@ncsu.edu

systems requires effectively allocating available resources (e.g., nurses and beds in hospitals). However, queueing models capturing realistic features of service systems can be extremely difficult to analyze. First, arrival rates typically vary significantly over time [1,5,12,46] which is not accounted for by standard queueing models. Second, abandonment by waiting customers, which corresponds to patients leaving without being seen by a care provider, or to callers hanging up in a call center, can significantly alter the system performance [47]. Next, empirical studies show that service times are not exponentially distributed and arrival processes are not Poisson [5,40,22], which motivates us to build more general models beyond the conventional Erlang models having tractable Markovian probability structure. Finally, service systems exhibit complicated network structures. For instance, callers may choose to call back later for more service in call centers [46] and patients are routed among different units in hospitals [40].

Any one of these features presents a significant challenge. Despite the immense queueing-theory literature, the model complexity of all four features makes exact analysis far beyond existing methods. Thus it is appropriate to seek effective approximations. *Many-server heavy-traffic* (MSHT) limit theorems for queueing systems have been proven useful to yield effective engineering approximations, because they provide both analytic performance formulas and practical insights; they turn the large scale into an advantage instead of a disadvantage.

There is a large body of literature on MSHT limits of queueing systems, see [6,7,20,21,32,34,36,35,39,44] for recent developments. We hereby review the most relevant works on non-Markovian queues with time-varying parameters. The MSHT fluid and diffusion limits were developed by Mandelbaum et al. [32] for the time-varying full Markovian queueing networks having Poisson arrivals and exponential service distributions. Liu and Whitt [25] proposed a fluid approximation for the $G_t/GI/s_t + GI$ queue with time-varying arrivals and non-exponential distributions; they later extended to the framework of networks [24,29]. A *functional weak law of large numbers* (FWLLN) [26] has been established to substantiate the fluid approximation in [25] and a *functional central limit theorem* (FCLT) [28] has been developed for the $G_t/M/s_t+GI$ model with exponential service times. More recent developments have been made. Paralleling [29] and the current paper, He and Liu [14] developed a fluid approximation for the multi-class queueing network with deterministic routing paths. Extending [24] and [28], Huang and Liu [16] developed a MSHT FCLT for network of queues with exponential service times.

Our contributions. This paper is a sequel to [26,29] which are extensions of [24,25]. We aim at extending the FWLLN [26] which supports the fluid approximation [25] for the $G_t/GI/s_t + GI$ queue. This $G_t/GI/s_t + GI$ fluid approximation was later generalized to the $(G_t/GI/s_t+GI)^m/M_t$ open queueing network [29], with a finite number of queues, non-stationary non-Poisson external arrival processes, non-exponential service times, time-varying staffing levels, customer abandonment following non-exponential patience times and probabilistic routing. Following the terminologies in [29], we call this time-

varying open network fluid approximations the *fluid queue network* (FQNets). Although simulation experiments have confirmed the effectiveness of the FQNets proposed in [29], a rigorous MSHT FWLLN supporting such FQNet approximations remained an open problem.

We will now solve this open problem by establishing a FWLLN for the $(G_t/GI/s_t+GI)^m/M_t$ open queueing network (see §§2–3 for the detailed model description and model parameters). In particular, as the scale increases, we will show that all performance functions, such as the queue lengths, routing flows and waiting time processes, of a sequence of the $(G_t/GI/s_t+GI)^m/M_t$ *stochastic queueing networks* (SQNets) converge to the associated deterministic performance functions of the corresponding $(G_t/GI/s_t + GI)^m/M_t$ FQNets. We will establish such a convergence in the product space of $\mathbb{D}$, which is the space of functions that are right continuous and have limits from the left.

A key step is to establish the FWLLN for the *total arrival process* (TAP) of each queue, that is the sum of the *external arrival process* (EAP) and feedback from the *internal routing processes* (IRPs) of the network. The *total arrival rate*, that is the fluid version of the TAP, is conjectured to satisfy a functional *fixed-point equation* (FPE), see [29] and also here in (10). We hereby prove that conjecture by first showing that the prelimit TAP satisfies a stochastic analog of the FPE (see (37)) and then establishing the asymptotic equivalence of the two equations as the scale increases. Exploiting the compactness approach [42], we (i) first prove the tightness of the TAPs and (ii) next establish the uniqueness of the limits of all convergent subsequences of TAPs. Once the FWLLN of the TAPs is established, it remains to separately treat each queue of the network by adopting results from [26], which treats the FWLLN of the TAP as an assumption.

A key assumption here is to assume all queues of the FQNet alternate between *overloaded* (OL) and *underloaded* (UL) intervals, or equivalently, the *efficiency-driven* (ED) and *quality-driven* (QD) regimes. As a result, the system should not always stay in the stable *critically loaded* (CL), or *quality-and-efficiency driven* (QED) regime. This is not too restrictive because managers of service systems may not be able or willing to frequently adjust the number of servers in the face of time-varying arrivals. When the staffing intervals are long, such as 8 hours in hospitals, these systems inevitably experience periods of overloadings and underloadings [25]. Effective staffing methods have been developed [8,19,27,30,31,46] to cope with time-varying arrivals. However, the time-stable performance can be achieved only in systems with flexible staffing. We hereby assume the system will be CL only at a finite number of time points.

Organization of the rest of the paper. In §2, we construct a sequence of $(G_t/GI/s_t + GI)^m/M_t$ SQNets and define the associated performance processes. In §3, we review the $(G_t/GI/s_t + GI)^m/M_t$ FQNet proposed in [29]; we specify the model assumptions and describe the system dynamics. In §4, we present our main result. In §5, we provide the detailed proofs of the main theorem. In §6, we provide practical confirmation of the FWLLN by considering an example. Finally, we draw conclusions in §7. Additional supplementary

materials appear in the appendix. In Appendix B we review useful results on infinite-server queues in [35]; In Appendices C–F we provide additional proofs to support §5. All acronyms are summarized in Appendix G.

2 A Sequence of $(G_t/GI/s_t + GI)^m/M_t$ Queueing Networks

The $(G_t/GI/s_t + GI)^m/M_t$ SQNet has a finite number of queues in parallel (the superscript m). The i^{th} queue, $1 \leq i \leq m$, has a general non-stationary *external arrival process* (EAP), *independent and identically distributed* (i.i.d.) service times following a non-exponential *cumulative distribution function* (cdf) G_i, a time-dependent staffing function (i.e., number of servers) $s_i(t)$ (the s_t), and allows customer abandonment with i.i.d. non-exponential abandonment times following cdf F_i. The service times, abandonment times and the EAP are mutually independent. External arrivals directly enter service if there are servers available; otherwise, they wait in an infinite-capacity queue and will receive service in order of their arrivals (following the *first-come first-served* (FCFS) service discipline), if they choose not to abandon.

Right after the service is completed at time t, a customer will independently be routed either to a queue j ($1 \leq j \leq m$) with a probability $p_{i,j}(t)$ (because the customer needs more service at station j) or directly out of the network (because the customer decides to leave the system) with probability $p_{i,0}(t)$. This routing policy is called the time-dependent probabilistic (Markovian) routing (the M_t). The probabilistic routing can be useful to model routing uncertainties. In hospitals, for example, patients leaving the intensive care units may be transferred to operating rooms due to sudden health deteriorations or to regular wards due to satisfying recovery.

A standard case of the EAP is the *non-homogeneous Poisson process* (NHPP) which is characterized by a rate function. We hereby consider a more general framework by relaxing that NHPP assumption, because statistical analysis shows that the arrival processes in real service systems can be far from Poisson [22]. If the EAPs are NHPPs (i.e., the G_t simplifies to M_t), and the service-time and abandonment-time distributions are exponential (i.e., the GI and $+GI$ degenerate to M and $+M$), then this SQNet simplifies to the full Markovian $(M_t/M/s_t + M)^m/M_t$ SQNet studied in [32].

A sequence of SQNets indexed by n**.** Using the $(G_t/GI/s_t+GI)^m/M_t$ SQNet introduced above as a base model, we now construct a sequence of $(G_t/GI/s_t + GI)^m/M_t$ SQNets indexed by n, where the scaling factor n represents the size (in terms of arrival rates and number of servers) of the n^{th} SQNet.

We assume all SQNets have the same service cdf G_i, abandonment cdf F_i and the routing probabilities $p_{i,j}(t)$ (so these parameters are independent with the scale n). Let $N_n^{(0,i)}$ and $s_n^{(i)}$ be the EAP and the number of servers of station i in the n^{th} SQNet. Let $\Rightarrow$ denote convergence in distribution [4,42]. We assume the following FWLLNs for the EAPs and staffing levels.

Assumption 1 (*FWLLN for EAPs and staffing levels*) *For each* i, $1 \le i \le m$, *there exist a nondecreasing function* $\Lambda_i(t)$ *with non-negative derivative* $\lambda_i(t)$ *and a piecewisely differentiable function* $s_i(t)$ *with derivative* $\dot{s}_i(t)$, *such that*

$$\bar{N}_n^{(0,i)}(t) \equiv n^{-1} N_n^{(0,i)}(t) \Rightarrow \Lambda_i^{(0)}(t) \equiv \int_0^t \lambda_i^{(0)}(u)du,$$
$$\bar{s}_n^{(i)}(t) \equiv n^{-1} s_n^{(i)}(t) \quad \Rightarrow s_i(t) \equiv \int_0^t \dot{s}_i(u)\, du \quad \textit{in } \mathbb{D}, \quad \textit{as } n \to \infty.$$

Remark 1 (Standard case of Assumption 1) Note we do not require the EAP to have a well-defined arrival-rate function for each n, but we do require the EAP to have an asymptotic rate function $\lambda^{(0)}$ as n increases. Considering the standard case of NHPPs, we can simply let the EAPs of the n^{th} SQNet be Poisson processes with scaled arrival rates $\lambda_n^{(i)} = n\lambda_i$. Standard cases for the staffing function are (i) $s_n^{(i)}(t) = \lceil n\, s_i(t) \rceil$ and (ii) $s_n^{(i)}(t) = \lceil n\, s_i(t) + \beta\sqrt{n\, s_i(t)} \rceil$, where β is a constant and $\lceil x \rceil$ is the least integer greater than or equal to x. Case (ii) is called the *square-root staffing* (SRS), see [8,27,28,46] for discussions on SRS. We remark that the $\sqrt{n}$ term will not affect the FWLLN or the fluid limit, but it may make an impact to the FCLT and diffusion limit [15,28].

We next define the performance functions. Let $B_n^{(i)}(t,y)$ ($Q_n^{(i)}(t,y)$) be the number of customers in service (in queue) at the i^{th} station at time t that have been so for time at most y. Let $B_n^{(i)}(t) \equiv B_n^{(i)}(t,\infty)$, $Q_n^{(i)}(t) \equiv Q_n^{(i)}(t,\infty)$ and $X_n^{(i)}(t) \equiv B_n^{(i)}(t) + Q_n^{(i)}(t)$ be the number of customers in service, in queue and total number in station i at time t. Let $A_n^{(i)}(t)$, $D_n^{(i)}(t)$ and $E_n^{(i)}(t)$ count the total number of customers that have abandoned, completed service and entered service by time t. Let $R_n^{(i,j)}(t)$ $(1 \le j \le m)$ count the number of customers routed to station j by time t from station i and let $R_n^{(i,0)}(t)$ count the number of customers departed (routed out of the network) from station i by t. Let $N_n^{(i)}(t)$ be the TAP (i.e., EAP plus IRPs) at station i by time t. Finally, let $W_n^{(i)}(t)$ and $V_n^{(i)}(t)$ denote the *head-of-line waiting time* (HWT, that is the elapsed waiting time for the customer at the head of the waiting line) and the *potential waiting time* (PWT, that is the virtual waiting time of an arrival at t assuming infinite patience).

In order to establish the FWLLN for these performance functions, we define the *law-of-large-numbers-scaled* (LLN-scaled) processes:

$$\begin{aligned} &\bar{B}_n^{(i)}(t,y) \equiv n^{-1} B_n^{(i)}(t,y), \quad \bar{Q}_n^{(i)}(t,y) \equiv n^{-1} Q_n^{(i)}(t,y), \quad \bar{X}_n^{(i)}(t) \equiv n^{-1} X_n^{(i)}(t), \\ &\bar{A}_n^{(i)}(t) \equiv n^{-1} A_n^{(i)}(t), \quad \bar{E}_n^{(i)}(t) \equiv n^{-1} E_n^{(i)}(t), \quad \bar{D}_n^{(i)}(t) \equiv n^{-1} D_n^{(i)}(t), \\ &\bar{R}_n^{(i,j)}(t) \equiv n^{-1} R_n^{(i,j)}(t) \quad \text{and} \quad \bar{N}_n^{(i)}(t) \equiv n^{-1} N_n^{(i)}(t), \quad 1 \le i \le m,\ 0 \le j \le m. \end{aligned} \tag{1}$$

We remark that the waiting times $W_n^{(i)}$ and $V_n^{(i)}$ are not scaled by n because F_i and G_i are not scaled by n.

We assume the following FWLLN holds for the initial number of customers.

Assumption 2 (*FWLLN for initial numbers*) *There exist nondecreasing functions* $B_i(0,x)$ *and* $Q_i(0,x)$ *with non-negative densities* $b_i(0,x)$ *and* $q_i(0,x)$*, such that* $(s_i(t) - B_i(0,\infty))Q_i(0,\infty) = 0$ *for* $1 \le i \le m$ *and*

$$\bar{B}_n^{(i)}(0,x) \Rightarrow B_i(0,x) \equiv \int_0^x b_i(0,y)\,dy,$$
$$\bar{Q}_n^{(i)}(0,x) \Rightarrow Q_i(0,x) \equiv \int_0^x q_i(0,y)\,dy \quad \textit{in } \mathbb{D}, \quad \textit{as } n \to \infty.$$

3 The $(G_t/GI/s_t + GI)^m/M_t$ Fluid Network

In this section, we review the $(G_t/GI/s_t+GI)^m/M_t$ FQNet [29]. First, in §3.1, we introduce the FQNet system and its parameters. In §3.2, we describe the performance of the FQNet in two steps. First, we characterize the performance of each queue of the FQNet in §3.2.1 assuming the TAR is a given parameter. Second, we discuss how to compute the m-dimensional vector of the TAR in §3.2.2.

3.1 The FQNet and Its Parameters

The deterministic FQNet is a legitimate dynamical system. There are m parallel fluid stations in the FQNet. At each station i, $1 \le i \le m$, external fluid arrives with rate $\lambda_i^{(0)}(t)$. Upon arrival, fluid immediately enters the service facility with a finite capacity $s_i(t)$, if there is space available. Otherwise, fluid flows into a waiting queue with an infinite capacity. Abandonment occurs for the fluid that is waiting in queue; in particular, a proportion $F_i(x)$ of fluid abandons (leaving the queue before entering the service facility) x units of time after its arrival. If not abandoning, fluid enters the service facility following the FCFS discipline. A proportion $G_i(x)$ of the fluid completes service x units of time after it enters service. A proportion $P_{i,j}(t)$ of the fluid completing service at time t is routed to station j $(1 \le j \le m)$ and a proportion $P_{i,0}(t) \equiv 1 - \sum_{j=1}^m P_{i,j}(t)$ is routed out of the system.

Let $R_{i,j}(t)$ be the amount of fluid routed from i to j with rate $r_{i,j}(t)$. Let $\Lambda_i^{(0)}(t)$ and $\Lambda_i(t)$ be the external fluid arrival and total fluid arrival of queue i, with *external arrival rate* (EAR) $\lambda_i^{(0)}(t)$ and TAR $\lambda_i(t)$. We have the following traffic-flow equations

$$\Lambda_i(t) \equiv \Lambda_i^{(0)}(t) + \sum_{j=1}^m R_{k,i}(t) \quad \text{and} \quad \lambda_i(t) \equiv \lambda_i^{(0)}(t) + \sum_{j=1}^m r_{k,i}(t), \tag{2}$$

where

$$\Lambda_i(t) \equiv \int_0^t \lambda_i(u)du, \quad \Lambda_i^{(0)}(t) \equiv \int_0^t \lambda_i^{(0)}(u)du,$$
$$R_{i,j}(t) = \int_0^t r_{i,j}(u)du, \quad r_{i,j}(t) = P_{i,j}(t)\sigma_i(t),$$

and σ_i is the service-completion rate of queue i, defined later in (9).

Let the two-parameter function $B(t,y)$ ($Q(t,y)$) be the quantity of fluid in service (in queue) at time t that has been so for at most y time units. We assume $B(t,y)$ and $Q(t,y)$ have densities $b(t,y)$ and $q(t,y)$, namely,

$$B(t,y) = \int_0^y b(t,x)dx \quad \text{and} \quad Q(t,y) = \int_0^y q(t,x)dx, \qquad y \geq 0, \tag{3}$$

Let $Q(t) \equiv Q(t,\infty)$, $B(t) \equiv B(t,\infty)$ and $X(t) \equiv Q(t) + B(t)$. We impose two constraints: (i) $B(t) \leq s(t)$ (capacity constraint) and (ii) $Q(t)(B(t) - s(t)) = 0$ (non-idling constraint).

In order to fully characterize the dynamics of the FQNet, we have to specify the model input $(\mathcal{P}, \mathcal{I})$, with

$$\mathcal{P} \equiv \left(m, \lambda_i^{(0)}, s_i, F_i, G_i, P_{i,j}, 1 \leq i,j \leq m\right), \quad \mathcal{I} \equiv \left(b_i(0,\cdot), q_i(0,\cdot), 1 \leq i \leq m\right), \tag{4}$$

where the six-tuple $\mathcal{P}$ has all model parameters of the FQNet and the pair $\mathcal{I}$ provides complete information on the initial state of the FQNet. We point out that the TAR λ_i is not part of the model input because it includes the internal routing rates $r_{i,j}$, which is to be determined. We assume the cdf's F_i and G_i have *probability density functions* (pdf's) f_i and g_i, and hazard-rate functions $h_{G_i}(x) \equiv g_i(x)/G_i^c(x)$ and $h_{F_i}(x) \equiv f_i(x)/F_i^c(x)$, where $G_i^c \equiv 1 - G_i$ and $F_i^c \equiv 1 - F_i$ are the *cumulative cdf's* (ccdf's) of G_i and F_i. We assume the service capacity function $s_i(t)$ is piecewise continuously differentiable and is feasible such that no fluid is forced out of service if s_i decreases. See [24,25] for more discussions and sufficient conditions on the feasibility of the service capacity function.

3.2 Performance Functions of the FQNet

In this subsection, we provide the performance formulas of the $(G_t/GI/s_t + GI)^m/M_t$ FQNet. Algorithms based on these formulas can be used to compute effective approximations for the corresponding FQNet [24,29]. In §3.2.1, we describe the performance of the i^{th} fluid queue as a function of the TAR. In §3.2.2, we characterize the TAR using a multi-dimensional functional FPE.

3.2.1 Performance of the i^{th} fluid queue given the TAR λ_i.

We now provide the performance functions for station i with its TAR λ_i regarded as a given parameter. For the sake of ease, we drop the subscript i in this subsection. We first describe the *overloaded* (OL) and *underloaded* (UL) intervals and the switching criterion of these intervals.

OL and UL periods. A fluid queue is said to be OL at time t if (i) $Q(t) > 0$ or (ii) $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(t) > \dot{s}(t) + \sigma(t)$, where $\dot{s}(t)$ is the derivative of $s(t)$. An OL period ends at time $T_1 \equiv \inf\{u \geq t : Q(u) = 0,\ \lambda(u) \leq \dot{s}(u) + \sigma(u)\}$. On the other hand, a fluid queue is said to be UL at time t if (i)

$B(t) < s(t)$ or (ii) $B(t) = s(t)$, $Q(t) = 0$ and $\lambda(t) \le \dot{s}(t) + \sigma(t)$. A UL period ends at time $T_2 \equiv \inf\{u \ge t : B(u) = s(u),\ \lambda(u) > \dot{s}(u) + \sigma(u)\}$. We say the queue is *critically loaded* (CL) if $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(u) = \dot{s}(u) + \sigma(u)$. Following [24–26, 29], we make the following assumption.

Assumption 3 (*Finite number of switches between UL and OL*) *In any finite interval* $[0, T]$*, all queues of the* $(G_t/GI/s_t{+}GI)^m/M_t$ *FQNet switches between OL and UL status for a finite number of times.*

See [24, 25] for sufficient conditions of this assumption. We next characterize the performance of the density functions $b(t, x)$ and $q(t, x)$ in (3) for UL and OL intervals.

Performance in a UL interval. In a UL interval, there is no fluid waiting in queue or abandonment from the queue, so we have $q = Q = w = v = 0$ and the abandonment cdf F plays no role. As a result, the $G_t/GI/s_t + GI$ fluid queue is equivalent to the $G_t/GI/\infty$ fluid model with an infinite service capacity. According to Proposition 2 of [25], the service density

$$b(t, x) = G^c(x)\lambda(t - x)\mathbf{1}_{\{x \le t\}} + \frac{G^c(x)}{G^c(x - t)} b(0, x - t)\mathbf{1}_{\{x > t\}}. \tag{5}$$

Performance in an OL interval. In an OL interval, the service density

$$b(t, x) = b(t - x, 0)G^c(x)\mathbf{1}_{\{x \le t\}} + \frac{G^c(x)}{G^c(x - t)} b(0, x - t)\mathbf{1}_{\{x > t\}}, \tag{6}$$

where the initial service density $b(0, y)$ is part of the initial condition descriptor $\mathcal{I}$ in (4), and the rate fluid enters service $b(t, 0)$ uniquely solves the FPE

$$b(t, 0) = \hat{a}(t) + \int_0^t b(t{-}x, 0)g(x)dx \quad \text{with} \quad \hat{a}(t) \equiv \dot{s}(t) + \int_0^\infty \frac{b(0, y)g(t + y)}{G^c(y)} dy. \tag{7}$$

See Theorem 2 in [25] for more details of the FPE (7).

We next determine the queue density function $q(t, x)$ in an OL interval. Let $w(t)$ and $v(t)$ be the head-of-line waiting time and potential waiting time at t, which are the deterministic analogs of the BWT $W(t)$ and PWT $V(t)$ of the corresponding SQNet in §2. According to Corollary 3 of [25], the queue content density

$$q(t, x) = \tilde{q}(t, x \wedge w(t)), \quad \tilde{q}(t, x) \equiv \lambda(t{-}x, 0)F^c(x)\mathbf{1}_{\{x \le t\}} + q(0, x{-}t)\frac{F^c(x)}{F^c(x - t)}\mathbf{1}_{\{t < x\}} \tag{8}$$

where $x \wedge y \equiv \min(x, y)$, the initial queue density $q(0, x)$ is part of the initial condition descriptor $\mathcal{I}$ in (4), and $w(t)$ and $v(t)$ uniquely solves the following *ordinary differential equations* (ODEs)

$$\dot{w}(t) = 1 - \frac{b(t, 0)}{\tilde{q}(t, w(t))} \quad \text{and} \quad \dot{v}(t) = 1 - \frac{\tilde{q}(t + v(t), v(t))}{b(t + v(t), 0)},$$

where $b(t,0)$ satisfies (7) and $\tilde{q}(t,x)$ is given in (8). See Theorems 3 and 5 in [25] for details.

Fluid flows. Let $A(t)$, $D(t)$ and $E(t)$ be the amount of fluid that has abandoned, completed service and entered service by time t, with rates $\alpha(t)$, $\sigma(t)$ and $b(t,0)$. Define

$$\begin{aligned} A(t) &\equiv \int_0^t \alpha(u)du, \quad \alpha(t) \equiv \int_0^\infty q(t,x)h_F(x)dx, \\ D(t) &\equiv \int_0^t \sigma(u)du, \quad \sigma(t) \equiv \int_0^\infty b(t,x)h_G(x)dx, \\ E(t) &\equiv \int_0^t b(u,0)du, \quad t \geq 0, \end{aligned} \tag{9}$$

where $q(t,x)$ and $b(t,x)$ satisfy (5),(6) and (8), and $b(t,0)$ solves (7).

3.2.2 Characterizing the TAR for the FQNet.

In the previous subsection, we described the performance for each queue in the FQNet assuming the TAR is given. We now characterize the vector of TAR for the entire FQNet.

Consider an interval $[0,\tau]$ during which no fluid queue changes status (switching between UL an OL). Let $\mathcal{U}(t) \equiv \{1 \leq i \leq m : B_i(t) \leq s_i(t),\ Q_i(t) = 0\}$ and $\mathcal{O}(t) \equiv \{1 \leq i \leq m : B_i(t) = s_i(t),\ Q_i(t) > 0\}$ be the sets of the indices of UL and OL queues in the FQNet. Note that the indices do not change with time, i.e., the sets $\mathcal{U} \equiv \mathcal{U}(t)$ and $\mathcal{O} \equiv \mathcal{O}(t)$, in the interval $[0,\tau]$.

The TAR $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_m)$ satisfies the multi-dimensional FPE

$$\boldsymbol{\lambda} = \Psi(\boldsymbol{\lambda}), \tag{10}$$

where for $\boldsymbol{u} \equiv (u_i, \ldots, u_m) \in \mathbb{D}^m$, the operator $\Psi : \mathbb{D}^m \to \mathbb{D}^m$ is defined as

$$\Psi(\boldsymbol{u})_i(t) \equiv \gamma_i(t) + \sum_{i \in \mathcal{U}} P_{i,j}(t) \int_0^t g_i(x)u_i(t-x)dx \tag{11}$$

$$\text{where} \quad \gamma_i(t) \equiv \lambda_i^{(0)}(t) + \sum_{k \in \mathcal{O}} P_{k,i}(t)\sigma_k(t) + \sum_{j \in \mathcal{U}} P_{j,i}(t) \int_0^\infty \frac{b_j(0,x)g_j(t+x)}{G_j^c(x)}dx.$$

By Theorem 1 of [29], Ψ is a contraction operator in $\mathbb{D}^m$, so that (10) has a unique solution $\boldsymbol{\lambda}$ in the interval $[0,\tau]$.

4 FWLLN for the $(G_t/GI/s_t + GI)^m/M_t$ SQNet

In this section, we present the FWLLN of $(G_t/GI/s_t + GI)^m/M_t$ SQNet. We show that the performance functions of the sequence of SQNet defined in §2 converge to the associated deterministic performance functions of the

$(G_t/GI/s_t + GI)^m/M_t$ FQNet reviewed in §3, as the scale increases. We establish the convergence in the appropriate product space of $\mathbb{D}$ and $\mathbb{D}_{\mathbb{D}}$, where $\mathbb{D} \equiv \mathbb{D}([0,\infty),\mathbb{R})$ is the space of right continuous real-valued functions with left limits, endowed with the Skorohod J_1 topology and metric d_{J_1} [42] and $\mathbb{D}_{\mathbb{D}} \equiv \mathbb{D}([0,\infty),\mathbb{D}([0,\infty),\mathbb{R}))$ [35]. We remark that the convergence under the J_1 metric reduces to the uniform convergence over compact sets $[0,T]$ for limits that are continuous functions. The limits for the single-parameter stochastic processes $\bar{N}_n$, $\bar{D}_n$, $\bar{E}_n$, $\bar{A}_n$, $\bar{R}_n$, $\bar{X}_n$, W_n, and V_n are established in the product space of $\mathbb{D}$ whereas the limits for the two-parameter processes $\bar{Q}_n$ and $\bar{B}_n$ are established in the product space of $\mathbb{D}_{\mathbb{D}}$.

Using bold face symbols to denote vectors, we define the vectors of the prelimit LLN-processes and the associated fluid functions as

$$\begin{aligned}
\bar{\boldsymbol{N}}_n &\equiv \left(\bar{N}_n^{(0,1)},\ldots,\bar{N}_n^{(0,m)}\right), & \boldsymbol{\Lambda} &\equiv (\Lambda_1,\ldots,\Lambda_m),\\
\bar{\boldsymbol{R}}_n &\equiv \left(\bar{R}_n^{(i,j)}, 1\le i\le m,\, 0\le j\le m\right), & \boldsymbol{R} &\equiv (R_{i,j}, 1\le i\le m,\, 0\le j\le m),\\
\bar{\boldsymbol{Q}}_n &\equiv \left(\bar{Q}_n^{(1)},\ldots,\bar{Q}_n^{(m)}\right), & \boldsymbol{Q} &\equiv (Q_1,\ldots,Q_m), \qquad (12)
\end{aligned}$$

and all other prelimit processes

$$\left(\bar{\boldsymbol{N}}_n^{(0)}, \bar{\boldsymbol{s}}_n, \bar{\boldsymbol{Q}}_n(0,\cdot), \bar{\boldsymbol{B}}_n(0,\cdot), \bar{\boldsymbol{D}}_n, \bar{\boldsymbol{E}}_n, \bar{\boldsymbol{A}}_n, \bar{\boldsymbol{X}}_n, \boldsymbol{W}_n, \boldsymbol{V}_n, \bar{\boldsymbol{Q}}_n\right)$$

and fluid functions

$$\left(\boldsymbol{\Lambda}^{(0)}, \boldsymbol{s}, \boldsymbol{Q}(0,\cdot), \boldsymbol{B}(0,\cdot), \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{Q}\right)$$

defined as analogs of (12). We are now ready to state our main result.

Theorem 1 (*FWLLN for the* $(G_t/GI/s_t + GI)^m/M_t$ *queueing network*) *If Assumptions 1–3 hold, then the FWLLN established in [26] for the* $G_t/GI/s_t + GI$ *model holds for the more general* $(G_t/GI/s_t+GI)^m/M_t$ *queueing network, namely, as* $n\to\infty$,

$$\begin{aligned}
&\left(\bar{\boldsymbol{N}}_n^{(0)}, \bar{\boldsymbol{s}}_n, \bar{\boldsymbol{Q}}_n(0,\cdot), \bar{\boldsymbol{B}}_n(0,\cdot), \bar{\boldsymbol{N}}_n, \bar{\boldsymbol{D}}_n, \bar{\boldsymbol{E}}_n, \bar{\boldsymbol{A}}_n, \bar{\boldsymbol{X}}_n, \boldsymbol{W}_n, \boldsymbol{V}_n, \bar{\boldsymbol{R}}_n, \bar{\boldsymbol{Q}}_n, \bar{\boldsymbol{B}}_n\right)\\
\Rightarrow &\left(\boldsymbol{\Lambda}^{(0)}, \boldsymbol{s}, \boldsymbol{Q}(0,\cdot), \boldsymbol{B}(0,\cdot), \boldsymbol{\Lambda}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V}, \boldsymbol{R}, \boldsymbol{Q}, \boldsymbol{B}\right) \qquad (13)
\end{aligned}$$

in $\mathbb{D}^{m^2+12m} \times \mathbb{D}_{\mathbb{D}}^{2m}$, *where the vectors of the prelimit processes are defined in* (12) *and §2, and the vector of deterministic fluid limit is defined in* (12) *and §3.*

Remark 2 (Useful engineering approximations from the FWLLN)
Theorem 1 provides mathematical justification for the fluid approximations in [24,29]. Simulation experiments [25,29] show that the single sample paths of the LLN-scaled performance processes agree closely with their deterministic fluid counterparts when n is large (e.g., $n = 1000$). When n is not large, stochastic fluctuations become significant, but the mean values of the LLN-scaled performance functions remain well approximated by the fluid functions with smaller n (e.g., $n = 10$).

Remark 3 (Special case without abandonment)
The $(G_t/GI/s_t)^m/M_t$ SQNet (FQNet) without customer abandonment can be viewed a special case of the $(G_t/GI/s_t+GI)^m/M_t$ SQNet (FQNet). Queueing models without customer abandonment are important because many service systems indeed have no abandonment or very low abandonment (e.g., health care systems and airport security lines). Moreover, the addition (removal) of the element of customer abandonment can significantly alter the system performance [47]. We remark that the proof of the FWLLN of the $(G_t/GI/s_t)^m/M_t$ SQNet is similar. In order to obtain the performance functions of the $(G_t/GI/s_t)^m/M_t$ FQNet, it suffices to let $f_{F_i}(x) = F_i(x) = 0$ and $F_i^c(x) = 1$ for $x \geq 0$ in all performance formulas in §3.

Remark 4 (Joint convergence in (13) and an arbitrary interval)
According to Theorem 11.4.5 of [42], the joint convergence in (13) is equivalent to the marginal convergence of each component, because the FWLLN limits are all deterministic functions. Hence, in §5, we will prove Theorem 1 by establishing the weak convergence of each component of the performance processes in space $\mathbb{D}$ or $\mathbb{D}_{\mathbb{D}}$. The weak convergence in Theorem 1 is equivalent to uniform convergence over the finite interval $[0,T]$, because the deterministic limits are continuous functions. Our proof strategy in §5 is to partition the interval $[0,T]$ into a sequence of disjoint intervals separated by a finite number of time points $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$, such that no fluid queue of the FQNet changes its OL or UL status in each interval $[t_{i-1}, t_i]$. Therefore, it suffices to prove the FWLLN in Theorem 1 by focusing on an interval $[0,\tau]$, where all initially OL (UL) queues remain OL (UL) throughout the interval.

5 Proof of the Main Result

Outline of the proof. We prove the weak convergence in Theorem 1 following the compactness approach [4,42,35]. In particular, we first show that the pre-limit processes (indexed by n) are *tight* (see [42] for definition and conditions for tightness), which implies that every subsequence has a further convergent subsequence. We next establish the full convergence by showing all convergent subsequences converge to the same limit (i.e. having a common probability law).

According to Remark 4, we will consider an interval $[0,T]$ where no queue changes its OL (UL) status (i.e., the queues that are OL (UL) at time 0 stays OL (UL) throughout the interval $[0,T]$). We establish the FWLLN for each component of (13) in the following order: First, we show the FWLLN for all service-related processes of OL queues in §5.1, including the *service-completion process* (SCP) $\bar{D}_n^{(i)}$, *enter-service process* (ESP) $\bar{E}_n^{(i)}$, *internal routing process* (IRP) $\bar{R}_n^{(i,j)}$ and the two-parameter service content $\bar{B}_n^{(i)}$, for $i \in \mathcal{O}$ and $1 \leq j \leq m$.

Using the FWLLN of the IRPs from OL queues, in §5.2 we next establish the FWLLNs of the TAP $\bar{N}_n^{(j)}$ and IRP $\bar{R}_n^{(i,j)}$, for $i \in \mathcal{U}$ and $1 \leq j \leq m$.

Finally, in §5.3 we apply the FWLLN of TAP to develop the FWLLNs of all other processes, including the service-related processes $\bar{B}_n^{(i)}$ and $\bar{D}_n^{(i)}$ for $i \in \mathcal{U}$, and the queue-related processes $W_n^{(j)}$, $V_n^{(j)}$, $\bar{Q}_n^{(j)}$ and $\bar{A}_n^{(j)}$ $j \in \mathcal{O}$. See Figure 1 for an illustration.

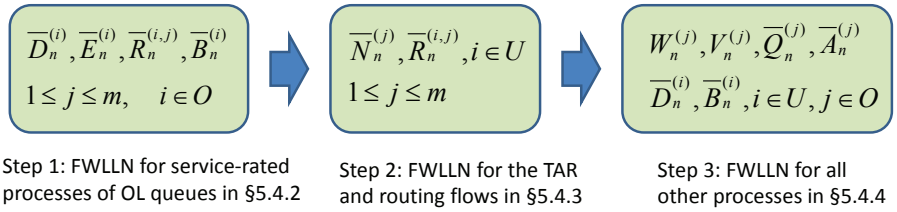


Fig. 1 Proof strategy for the FWLLN in Theorem 1.

Asymptotically UL and OL intervals. Suppose queue i of the FQNet is OL in $[0, T]$ (the argument for the UL case is similar), with the net input rate $\lambda_i(0) - \sigma_i(0) > 0$ and $\Lambda_i(t) > D_i(t)$ for $0 < t < T$. Because $\bar{Q}_n^{(i)}(0) \Rightarrow Q_i(0) > 0$, queue i of the SQNet will become asymptotically OL (that is, all servers will asymptotically become busy and remain so) throughout an interval $[t_{1,n}, t_2]$ with $0 < t_{1,n} = o(1/n) < t_2$, even though some servers could be idle in the neighborhood of 0. We next construct the performance functions for the asymptotically OL queues (with index $i \in \mathcal{O}$) and UL queues (with $i \in \mathcal{U}$). Let $|\mathcal{O}|$ and $|\mathcal{U}|$ be the numbers of OL and UL queues (i.e., the numbers of indices in sets $\mathcal{O}$ and $\mathcal{U}$).

5.1 FWLLN for service related processes at OL queues

In this subsection, we establish the FWLLN for the service related processes, including the SCP, ESP, IRP and service-content processes. In particular, we now show, for $i \in \mathcal{O}$, $1 \le j \le m$,

$$(\bar{D}_n^{(i)}, \bar{E}_n^{(i)}, \bar{R}_n^{(i,j)}, \bar{B}_n^{(i)}) \Rightarrow (D_i, E_i, R_{i,j}, B_i) \quad \text{in} \quad \mathbb{D}^3 \times \mathbb{D}_{\mathbb{D}}, \quad \text{as } n \to \infty. \tag{14}$$

We are able to prove (14) before establishing the FWLLN of the TAP because the service-related processes of the OL queues do not directly depend on the TAP. Instead, they depend on the number of existing customers in service (i.e., old service content), the total number of servers, and how fast the servers become available to serve customers at the head of the waiting line.

5.1.1 Service-completion process $\bar{D}_n^{(i)}$ and enter-service process $\bar{E}_n^{(i)}$.

Our proof in this subsection draws heavily on the results in [26]. We provide the major steps here so the paper is self contained. But we refer to [26] for some detailed proofs to avoid repetition. We also provide the omitted details in the appendix.

Flow conservation of the service content (i.e., number of customers in service) at an OL queue i implies that

$$\bar{E}_n^{(i)}(t) = (\bar{s}_n^{(i)}(t) - \bar{s}_n^{(i)}(0)) + \bar{D}_n^{(i)}(t) \quad \text{for all} \quad i \in \mathcal{O}. \tag{15}$$

Hence, by Assumption 1, the tightness and weak convergence of $\bar{D}_n^{(i)}$ will easily imply the tightness and weak convergence of $\bar{E}_n^{(i)}$. We give the tightness results in the next lemma.

Lemma 1 *The sequence of processes* $\left(\bar{D}_n^{(i)}, \bar{E}_n^{(i)}, i \in \mathcal{O}\right)$ *is C-tight in* $\mathbb{D}^{|\mathcal{O}|}$.

Proof By Theorem 11.6.7 of [42], the tightness of the big vector in Lemma 1 is equivalent to the tightness of the components $\bar{D}_n^{(i)}$ and $\bar{E}_n^{(i)}$, for all $i \in \mathcal{O}$. The proof of the C-tightness of the $\bar{D}_n^{(i)}$ closely follows from the proof in §3 of [26], so we give the proof in Appendix C. Given the C-tightness of $\bar{D}_n^{(i)}$, the C-tightness of $\bar{E}_n^{(i)}$ follows from Assumption 1, the smoothness of the limiting staffing function $s_i(t)$ and the continuous mapping theorem with addition. ∎

We next characterize the limit of a convergent subsequence of $\bar{D}_n^{(i)}$. We first split the SCP and service content into two terms, corresponding to *old customers* (initially in service at time 0) and *new customers* (arriving after time 0). In particular, we write

$$\bar{D}_n^{(i)}(t) = \bar{D}_n^{(i,o)}(t) + \bar{D}_n^{(i,\nu)}(t) \quad \text{and} \quad \bar{B}_n^{(i)}(t,x) = \bar{B}_n^{(i,o)}(t,x) + \bar{B}_n^{(i,\nu)}(t,x) \tag{16}$$

where $\bar{D}_n^{(i,o)}(t)$ ($\bar{B}_n^{(i,o)}(t,x)$) denotes the LLN-scaled number of service completions by t (customers in service at t with ages no more than x) from those already in service at time 0, and $\bar{D}_n^{(i,\nu)}(t)$ ($\bar{B}_n^{(i,\nu)}(t,x)$) denotes the LLN-scaled number of service completions by t (customers in service at t with ages no more than x) from the new arrivals in the interval $[0,t]$.

To treat the *first* term of the SCP in (16), we follow [26] by writing

$$\bar{D}_n^{(i,o)}(t) = \bar{B}_n^{(i)}(0) - \bar{B}_n^{(i,o)}(t) \quad \text{and} \quad \bar{B}_n^{(i,o)}(t) = \frac{1}{n} \sum_{k=1}^{B_n^{(i)}(0)} \mathbf{1}\left(\eta_{k,i}(\tau_{n,i}^{(k)}) > t\right), \tag{17}$$

where $\mathbf{1}(\cdot)$ is the indicator random variable, $\left\{0 < \tau_{n,i}^{(1)} \le \tau_{n,i}^{(2)} \le \ldots\right\}$ is an ordered sequence of the ages (i.e., elapsed service times) of the old customers (i.e., those in service at time 0), and $\{\eta_{1,i}(x), \eta_{2,i}(x), \ldots\}$ is an i.i.d. sequence of random variables following ccdf

$$P(\eta_{1,i}(x) > t) = \frac{G_i^c(t+x)}{G_i^c(x)}, \qquad x \ge 0. \tag{18}$$

By Theorem 2 of [26] and Assumption 2, we have, as $n \to \infty$,

$$\bar{B}_n^{(i,o)}(t) \Rightarrow B^{(i,o)}(t) \equiv \int_0^\infty b_i(0,x) \frac{G_i^c(t+x)}{G_i^c(x)} dx \quad \text{in } \mathbb{D}, \tag{19}$$

which, together with (17) and Assumption 2, concludes the FWLLN of $\bar{D}_n^{(i,o)}$. Namely, as $n \to \infty$, for $i \in \mathcal{O}$,

$$\bar{D}_n^{(i,o)}(t) \Rightarrow D^{(i,o)}(t) \equiv B^{(i,o)}(0) - B^{(i,o)}(t) = \int_0^\infty b_i(0,x) \left(1 - \frac{G_i^c(t+x)}{G_i^c(x)}\right) dx. \tag{20}$$

To treat the *second* term of the SCP in (16), we write

$$\bar{D}_n^{(i,\nu)}(t) = n^{-1} \sum_{k=1}^{E_n^{(i)}(t)} \mathbf{1}\left(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \le t\right), \tag{21}$$

where $\mathcal{E}_k^{(i,n)}$ denotes the time the k^{th} customer enters service, and $\left\{\mathcal{S}_1^{(i)}, \mathcal{S}_2^{(i)}, \ldots\right\}$ are the i.i.d. service times following the cdf G_i. By (15), (16), (17) and (21), we have

$$\begin{aligned}\bar{E}_n^{(i)}(t) = (\bar{s}_n^{(i)}(t) - \bar{s}_n^{(i)}(0)) &+ \frac{1}{n} \sum_{k=1}^{n\,\bar{B}_n^{(i)}(0)} \mathbf{1}\left(\eta_{k,i}(\tau_{n,i}^{(k)}) \le t\right) \\ &+ \frac{1}{n} \sum_{k=1}^{n\,\bar{E}_n^{(i)}(t)} \mathbf{1}(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \le t). \end{aligned} \tag{22}$$

Following the proofs in §6 of [26] (the details omitted here and provided in Appendix D), we have the convergence

$$\bar{D}_n^{(i,\nu)}(t) \Rightarrow D^{(i,\nu)}(t) \equiv \int_0^t G_i(t-s) b_i(s,0) ds \quad \text{and} \tag{23}$$

$$\bar{E}_n^{(i)}(t) \Rightarrow E_i(t) \equiv \int_0^t b_i(s,0) ds \quad \text{in} \quad \mathbb{D}, \tag{24}$$

where $b(0,\cdot)$ solves the FPE (7). The full convergence of $\{\bar{D}_n^{(i)}\}$ and $\{\bar{E}_n^{(i)}\}$ immediately follows from Lemma 1, (16), (20), (23), (24) and the continuous mapping theorem with addition.

5.1.2 Two-parameter service content $\bar{B}_n^{(i)}$.

Extending the sums in (17) and (21), we give the two-parameter representations for the LLN-scaled new service content $\bar{B}_n^{(i,\nu)}$ and old service content $\bar{B}_n^{(i,o)}$:

$$\bar{B}_n^{(i,\nu)}(t,y) = \frac{1}{n} \sum_{k=E_n^{(i)}((t-y)^+)+1}^{E_n^{(i)}(t)} \mathbf{1}\left(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} > t\right), \tag{25}$$

$$\bar{B}_n^{(i,o)}(t,y) = \frac{1}{n} \sum_{k=1}^{B_n^{(i)}\left(0,(y-t)^+\right)} \mathbf{1}\left(\eta_{k,i}(\tau_{n,i}^{(k)}) > t\right). \tag{26}$$

The FWLLN of (25) follows from (24) and Theorem 3.1 of [35]. In particular,

$$\bar{B}_n^{(i,\nu)}(t,y) \Rightarrow B^{(i,\nu)}(t,y) \equiv \int_{(t-y)^+}^{t} G_i^c(t-s)dE_i(s) \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}. \tag{27}$$

The FWLLN of (26) follows from Assumption 2 and Theorem 2 of [26] (here Lemma 4 in the appendix). We have

$$\bar{B}_n^{(i,o)}(t,y) \Rightarrow B^{(i,o)}(t,y) \equiv \int_{0}^{(y-t)^+} b_i(0,x)\frac{G_i^c(t+x)}{G_i^c(x)}dx \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}. \tag{28}$$

Together with (16), (27) and (28), we apply the continuous mapping theorem with addition to obtain the FWLLN of the two-parameter service content $\bar{B}_n^{(i)}$ with the limit B_i given in (3) and (6) .

Remark 5 (FWLLN for processes related to old content in service)
Although the FWLLN for the processes related to the old service content $\bar{B}_n^{(i,o)}$ and $\bar{D}_n^{(i,o)}$ are developed here for an OL queue i (i.e., $i \in \mathcal{O}$), the same arguments (of the FWLLN and fluid limit) hold for an UL queue i (i.e., $i \in \mathcal{U}$). Because we assume no customer is forced out of service before completing service when the staffing level decreases (if ever), the dynamics of the old customers in service (those already in service at time 0) does not depend on if the queue is OL or UL; namely, their behavior is not affected by the ESP $E_n^{(i)}$ or the number of servers $s_n^{(i)}$, because they will continue to occupy the servers until their services are completed (in some sense they have higher priorities comparing with new customers).

However, at an OL queue, the performance of processes related to new content (e.g., $\bar{B}_n^{(i,\nu)}$ and $\bar{D}_n^{(i,\nu)}$) are precisely controlled by the amount of available service resources (here represented by $s_n^{(i)}(t) - B_n^{(i,o)}(t)$), which determines how often new customers should enter service (reflected by $E_n^{(i)}(t)$). On the contrary, the dynamics is very different at a UL queue where there is almost no constraint on the service capacity (because a UL queue is equivalent to an infinite-server queue). Therefore, in §5.3 we will only have to establish the FWLLNs for $\bar{B}_n^{(i,\nu)}$ and $\bar{D}_n^{(i,\nu)}$ of a UL queue i, because the proofs of the FWLLNs for $\bar{D}_n^{(i,o)}$ and $\bar{B}_n^{(i,o)}$ are identical to those for an OL queue with limits in the same forms as in (20) and (28).

5.1.3 Internal routing flows $\bar{R}_n^{(i,j)}$ from OL queues.

We next establish the FWLLN for the IRP $R_n^{(i,j)}$, from an OL queue i (i.e., $i \in \mathcal{O}$) to another queue j ($0 \le j \le m$), with $j = 0$ denoting the outside world (i.e., leaving the network). First we provide the representation for the routing process using independent indicators splitting the SCP. For $s \ge 0$, let $\left\{\delta_{i,j}^{(1)}(s), \delta_{i,j}^{(2)}(s), \ldots\right\}$ and $\left\{\tilde{\delta}_{i,j}^{(1)}(s), \tilde{\delta}_{i,j}^{(2)}(s), \ldots\right\}$ be two independent i.i.d. sequences of indicator random variables with $P\left(\delta_{i,j}^{(1)}(s) = 1\right) =$

$P\left(\tilde{\delta}_{i,j}^{(1)}(s)=1\right) = 1 - P\left(\delta_{i,j}^{(1)}(s)=0\right) = 1 - P\left(\tilde{\delta}_{i,j}^{(1)}(s)=0\right) = P_{i,j}(s)$. We write

$$\bar{R}_n^{(i,j)}(t) = \bar{R}_n^{(i,j,o)}(t) + \bar{R}_n^{(i,j,\nu)}(t), \tag{29}$$

where

$$\bar{R}_n^{(i,j,o)}(t) \equiv \frac{1}{n}\sum_{k=1}^{D_n^{(i,o)}(t)} \delta_{i,j}^{(k)}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right) \quad \text{and} \quad \bar{R}_n^{(i,j,\nu)}(t) \equiv \frac{1}{n}\sum_{l=1}^{D_n^{(i,\nu)}(t)} \tilde{\delta}_{i,j}^{(l)}\left(\zeta_{n,i}^{(l)}\right) \tag{30}$$

denote the routing flows from old customers (those already in service at time 0) and new customers in service (from new arrivals in $[0,t]$), $\zeta_{n,i}^{(l)} = \mathcal{E}_l^{(i,n)} + \mathcal{S}_l^{(i)}$ is the service-completion time of the l^{th} new customer, with $\mathcal{E}_l^{(i,n)}$ and $\mathcal{S}_l^{(i)}$ defined in (21), and $\eta_{k,i}(\tau_{n,i}^{(k)})$ is the service-completion time of an old customer having the k^{th} smallest elapsed service time (age), defined in (17).

We obtain the FWLLN of the IRP using the continuous mapping theorem; in particular we express $\bar{R}_n^{(i,j)}$ as a function of the SCP $\bar{D}_n^{(i)}$. Adding and subtracting $P_{i,j}(\eta_{n,i}^{(k)})$ in the first equation and $P_{i,j}(\zeta_{n,i}^{(l)})$ in the second equation of (30) yields

$$\begin{aligned}
&\bar{R}_n^{(i,j,o)}(t) \\
&= \frac{1}{n}\sum_{k=1}^{D_n^{(i,o)}(t)} \left[\delta_{i,j}^{(k)}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right) - P_{i,j}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right)\right] + \frac{1}{n}\sum_{k=1}^{D_n^{(i,o)}(t)} P_{i,j}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right) \\
&= \frac{1}{n}\sum_{k=1}^{D_n^{(i,o)}(t)} \left[\delta_{i,j}^{(k)}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right) - P_{i,j}\left(\eta_{k,i}(\tau_{n,i}^{(k)})\right)\right] + \int_0^t P_{i,j}(u)d\bar{D}_n^{(i,o)}(u),
\end{aligned} \tag{31}$$

and

$$\begin{aligned}
\bar{R}_n^{(i,j,\nu)}(t) = &\quad \frac{1}{n}\sum_{l=1}^{D_n^{(i,\nu)}(t)} \left[\delta_{i,j}^{(l)}\left(\zeta_{n,i}^{(l)}\right) - P_{i,j}\left(\zeta_{n,i}^{(l)}\right)\right] + \frac{1}{n}\sum_{l=1}^{D_n^{(i,\nu)}(t)} P_{i,j}\left(\zeta_{n,i}^{(l)}\right) \\
= &\quad \frac{1}{n}\sum_{l=1}^{D_n^{(i,\nu)}(t)} \left[\delta_{i,j}^{(l)}\left(\zeta_{n,i}^{(l)}\right) - P_{i,j}\left(\zeta_{n,i}^{(l)}\right)\right] + \int_0^t P_{i,j}(u)d\bar{D}_n^{(i,\nu)}(u).
\end{aligned} \tag{32}$$

Convergence of the second terms in (31) *and* (32). We next show that

$$\int_0^t P_{i,j}(u)d\bar{D}_n^{(i,\nu)}(u) \Rightarrow \int_0^t P_{i,j}(u)dD^{(i,\nu)}(u), \tag{33}$$

$$\int_0^t P_{i,j}(u)d\bar{D}_n^{(i,o)}(u) \Rightarrow \int_0^t P_{i,j}(u)dD^{(i,o)}(u) \quad \text{in} \quad \mathbb{D}, \quad \text{as } n \to \infty. \tag{34}$$

We only show (33) because (34) is similar. We apply the continuous mapping theorem based on the next lemma, with its proof given in Appendix F.

Lemma 2 *For $x \in \mathbb{D}$, the function $\phi : \mathbb{D} \to \mathbb{D}$ defined as*

$$(\phi(x))(t) \equiv P_{i,j}(t)x(t) - \int_0^t x(s)dP_{i,j}(s)$$

is continuous, if the $P_{i,j}(t)$ is piecewisely differentiable.

Since $\bar{D}_n^{(i,\nu)}(\omega, t)$ is nondecreasing in t for almost all $\omega \in \Omega$, with $\bar{D}_n^{(i,\nu)}(0) = 0$ satisfying $\mathbf{E}[\bar{D}_n^{(i,\nu)}(t)] < \infty$ for all $t \in [0, \infty)$, $\bar{D}_n^{(i,\nu)}(\omega, t)$ is of bounded variation for each $n \geq 1$. Therefore, combined with the fact that $\bar{D}_n^{(i,\nu)}(t)$ is right continuous with left limits for almost all $\omega \in \Omega$, the second term in (32) is a Stieltjes integral for fixed ω. Then, by integration by parts,

$$\int_0^t P_{i,j}(s)d\bar{D}_n^{(i,\nu)}(\omega, s) = P_{i,j}(t)\bar{D}_n^{(i,\nu)}(\omega, t) - \int_0^t \bar{D}_n^{(i,\nu)}(\omega, s)dP_{i,j}(s)$$

for all $n \geq 1$. Hence, by Lemma 2 and the FWLLN of $\bar{D}_n^{(i,\nu)}$ in (23), we conclude the convergence in (33).

Asymptotic negligibility of the first terms in (31) *and* (32). We now complete the proof of the FWLLN of $\bar{R}_n^{(i,j)}$ by showing the first terms of (31) and (32) are asymptotically negligible. Because the proofs are similar, we only show the latter.

We first condition on a realization of the sequence $\{\zeta_{n,i}^{(l)},\ l \geq 1\}$. For a fixed t, the first term of (32) is a scaled random sum of independent zero-mean random variables, each taking values in the interval $[-1, 1]$. Because the random variables are not identically distributed, we apply the law of large numbers for non-identically distributed triangular arrays, see Theorem 1 on p.307 of [9], also see Appendix E. As a result, for fixed $t \geq 0$, $i \in \mathcal{O}$ and $1 \leq j \leq m$, we have

$$\begin{aligned}\widehat{R}_n^{(i,j,\nu)}(t) &= \frac{1}{n}\sum_{l=1}^{D_n^{(i,\nu)}(t)} \left(\tilde{\delta}_{i,j}^{(l)}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)})\right) \\ &= \bar{D}_n^{(i,\nu)}(t) \sum_{l=1}^{D_n^{(i,\nu)}(t)} \frac{\tilde{\delta}_{i,j}^{(l)}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)})}{D_n^{(i,\nu)}(t)} \Rightarrow 0. \end{aligned} \tag{35}$$

in $\mathbb{R}$, where the sum in the second equation converges in distribution to 0 by (66) and $\bar{D}_n^{(i,\nu)}$ converges to $D^{(i,\nu)}$. The convergence in (35) for a fixed t can then easily extend to uniform convergence over compact sets according to Theorem 3.2.1 in the internet supplement of [42].

Repeating the same argument for $\bar{R}_n^{(i,j,o)}$ and apply the continuous mapping theorem using addition, we complete the proof of the FWLLN of (29), namely,

$$\bar{R}_n^{(i,j)}(t) \Rightarrow \int_0^t P_{i,j}(u)dD^{(i,\nu)}(u) + \int_0^t P_{i,j}(u)dD^{(i,o)}(u) \quad i \in \mathcal{O}, \quad 1 \leq j \leq m. \tag{36}$$

It is easy to see that the right-hand side of (36) agrees with $R_{i,j}(t)$ in (2), by combining (20) and (23).

5.2 FWLLN for the Total Arrival Process

We now prove the FWLLN of the TAP $\bar{\boldsymbol{N}}_n$. First, we construct equations describing the prelimits of the TAP. We next prove the full convergence of $\bar{\boldsymbol{N}}_n$ following the compactness approach [42], by establishing (i) the tightness of the TAP and (ii) showing the limit of all convergent subsequences uniquely solves the multi-dimensional FPE in (10).

Because the TAP is the sum of the EAP and IRPs, we have, for $1 \le j \le m$,

$$\bar{N}_n^{(j)}(t) = \bar{N}_n^{(0,j)}(t) + \sum_{i\in\mathcal{O}} \bar{R}_n^{(i,j)}(t) + \sum_{i\in\mathcal{U}} \bar{R}_n^{(i,j)}(t), \tag{37}$$

where $\bar{N}_n^{(0,j)}$ is the EAP of the j^{th} queue and $\bar{R}_n^{(i,j)}$ is the IRP from queue i to queue j. Because the FWLLN is obtained in §5.1.3 for $\bar{R}_n^{(i,j)}$ with $i \in \mathcal{O}$ and the FWLLN for $\bar{N}_n^{(0,j)}$ is given in Assumption 1, it remains to treat the third term in (37). Although the IRPs from an UL queue has the same representation as that in (29) and (30) for $i \in \mathcal{O}$, the SCP of new customers at a UL queue is different because the ESP is now the TAP, i.e., $E_n^{(i)} = N_n^{(i)}$. Modifying (21), we have

$$\bar{D}_n^{(i,\nu)}(t) = \frac{1}{n} \sum_{k=1}^{N_n^{(i)}(t)} \mathbf{1}(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \le t) \quad \text{for all} \quad i \in \mathcal{U} \tag{38}$$

Following the compactness approach, we first establish the tightness of the TAP in the next lemma.

Lemma 3 *The TAP* $\left(\bar{N}_n^{(1)}, \ldots, \bar{N}_n^{(m)}\right)$ *is C-tight in* $\mathbb{D}^m$.

Proof By Theorem 11.6.7 of [42], it suffices to show the C-tightness of $\bar{N}_n^{(j)}$ in $\mathbb{D}$ for all $1 \le j \le m$. Based on (38), we first bound the routing processes $R_n^{(i,j)}$ with the departures $D_n^{(i)}$, in particular, we have

$$\begin{aligned}\bar{N}_n^{(i)}(t) &\le \bar{N}_n^{(0,i)}(t) + \sum_{i\in\mathcal{O}} \bar{D}_n^{(i)}(t) + \sum_{k\in\mathcal{U}} \bar{D}_n^{(k)}(t) \\ &= \bar{N}_n^{(0,i)}(t) + \sum_{i\in\mathcal{O}} \bar{D}_n^{(i)}(t) + \sum_{k\in\mathcal{U}} \bar{D}_n^{(k,o)}(t) + \sum_{k\in\mathcal{U}} \bar{D}_n^{(k,\nu)}(t). \end{aligned} \tag{39}$$

The convergence to continuous limits of (i) $\bar{N}_n^{(0,i)}$ for $1 \le i \le m$ (Assumption 1), (ii) $\bar{D}_n^{(i)}$ for $i \in \mathcal{O}$ (§5.1.1) and (iii) $\bar{D}_n^{(i,o)}$ for $i \in \mathcal{U}$ (Remark 5), implies the C-tightness of the first three terms of (39). To complete the proof of Lemma 3, it remains to show the C-tightness of the last term in (39), because the C-tightness is preserved under addition (Chapter VI, Corollary 3.33 of [17]).

For a UL queue k (i.e., $k \in \mathcal{U}$), let $s_n^{k,\uparrow} \equiv \sup\{s_n^{(k)}(t) : 0 \le t \le T\}$ and let $Z_1(t), Z_2(t), \ldots$ be an i.i.d. sequence of renewal processes with inter-renewal

times following the cdf G_k. We can then bound the SCP $D_n^{(k,\nu)}(t)$ by the sum of $s_n^{k,\uparrow}$ independent renewal processes, in particular,

$$\bar{D}_n^{(k)}(t) \le \frac{1}{n} \sum_{r=1}^{s_n^{k,\uparrow}(k)} Z_r(t). \tag{40}$$

By the proof of Lemma 1 (see Appendix C), the right-hand side of (40) is C-tight. Therefore, by Chapter VI, Proposition 3.35 of [17], $\bar{D}_n^{(k)}$ has to be C-tight for each $k \in \mathcal{U}$. We thus conclude the proof. ∎

Since Lemma 3 implies that every subsequence of $\bar{N}_n^{(i)}$ has a further convergent subsequence $\bar{N}_{n_k}^{(i)}$, we complete the proof of the FWLLN of the TAP by showing that every convergent subsequence of $\bar{N}_n^{(i)}$ converges to $\Lambda_i(t) = \int_0^t \lambda_i(u)du$ with λ_i characterized as the unique solution to the multi-dimensional FPE in (10). For simplicity, we use $\{\bar{N}_n^{(i)}\}$ (instead of $\{\bar{N}_{n_k}^{(i)}\}$) to denote an arbitrary convergent subsequence of the TAP. Because of the C-tightness, we assume this subsequence $\bar{N}_n^{(i)} \Rightarrow N_i^*$ in $\mathbb{D}$ for some continuous limit N_i^*, $1 \le i \le m$, as $n \to \infty$.

Paralleling the proof of the FWLLN for $\bar{D}_n^{(i,\nu)}$ of an OL queue in (23), we easily obtain, from (38), that

$$\bar{D}_n^{(i,\nu)}(t) \Rightarrow D_i^{(\nu,*)}(t) \equiv \int_0^t G_i(t-s)dN_i^*(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty, \quad i \in \mathcal{U}. \tag{41}$$

Paralleling the proof of the FWLLN for $\bar{R}_n^{(i,j,\nu)}$ with $i \in \mathcal{O}$ in §5.1.3, we have

$$\bar{R}_n^{(i,j,\nu)}(t) \Rightarrow R_{i,j}^{(\nu,*)}(t) \equiv \int_0^t P_{i,j}(s)dD_i^{(\nu,*)}(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty, \tag{42}$$

for all $i \in \mathcal{U}$, $0 \le j \le m$. By Remark 5, we obtain the FWLLN of $\bar{R}_n^{(i,j,o)}$ for free, namely,

$$\bar{R}_n^{(i,j,o)}(t) \Rightarrow \int_0^t P_{i,j}(s)dD^{(i,o)}(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty, \quad i \in \mathcal{U}, 0 \le j \le m, \tag{43}$$

where $D^{(i,o)}$ is defined in (20). Finally, combining (37), (42), (43), Assumption 1 and (36), we have

$$\begin{aligned} N_j^*(t) \equiv \Lambda_j^{(0)}(t) &+ \sum_{i \in \mathcal{O}} \int_0^t P_{i,j}(u)dD_i(u) + \sum_{i \in \mathcal{U}} \int_0^t P_{i,j}(u)dD^{(i,o)}(u) \\ &+ \sum_{i \in \mathcal{U}} \int_0^t P_{i,j}(u)dD_i^{(\nu,*)}(u), \end{aligned} \tag{44}$$

where $D_i \equiv D^{(i,o)} + D^{(i,\nu)}$, $D^{(i,o)}$ is given in (20), $D^{(i,\nu)}$ is given in (23) and $D_i^{(\nu,*)}$ is defined in (41). It is not hard to see that (44) agrees with the integral

version of the FPE (10). In particular, because the proof in Appendix C also implies the limit $\bar{N}_j^*$ is Lipschitz continuous, taking the derivative of (44) with respect to t gives

$$\dot{N}_j^*(t) = \lambda_j^{(0)}(t) + \sum_{i\in\mathcal{O}} P_{i,j}(t)\sigma_i(t) + \sum_{i\in\mathcal{U}} P_{i,j}(t)\int_0^\infty \frac{b_i(0,y)g_i(t+y)}{G_i^c(y)}dy$$
$$+ \sum_{i\in\mathcal{U}} P_{i,j}(t)\int_0^t g_i(t-x)dN_i^*(x), \tag{45}$$

which coincides with the FPE (10). Since this FPE has a unique solution (see Theorem 1 of [29]) and the choice of the subsequence is arbitrary, all convergent subsequences must have the same limit, we have $P(\boldsymbol{N}^* = \boldsymbol{\Lambda}) = 1$ for $\boldsymbol{\Lambda}$ in (12) and (2). Hence we have completed the proof of the FWLLN of the TAP.

5.3 FWLLNs for Other Processes

We now complete the proof of Theorem 1 by establishing the FWLLNs of all other processes, including the service-related processes of UL queues (e.g., $\bar{D}_n^{(i)}$ and $\bar{B}_n^{(i)}$ for $i \in \mathcal{U}$) and queue-related processes of OL queues (e.g., $W_n^{(j)}$, $V_n^{(j)}$, $\bar{Q}_n^{(j)}$ and $\bar{A}_n^{(j)}$ for $i \in \mathcal{O}$). Because the FWLLN of the TAP is established, we now independently treat each queue i, $1 \le i \le m$, with a given FWLLN of its TAP $\bar{N}_n^{(i)}$. We draw heavily on the proofs in [26,28].

5.3.1 FWLLNs for service-related processes at UL queues.

Mimicking (16) and the arguments in §5.1.1, we split $\bar{D}_n^{(i)}$ ($\bar{B}_n^{(i)}$) of a UL queue i into the SCP (service content) of new customers $\bar{D}_n^{(i,\nu)}$ ($\bar{B}_n^{(i,\nu)}$) and the SCP (service content) of old customers $\bar{D}_n^{(i,o)}$ ($\bar{B}_n^{(i,o)}$). As discussed in Remark 5, the FWLLNs of $\bar{D}_n^{(i,o)}$ and $\bar{B}_n^{(i,o)}$ have been developed in §5.1.1 with limits in (20) and (28). It remains to prove the FWLLNs for $\bar{D}_n^{(i,\nu)}$ and $\bar{B}_n^{(i,\nu)}$. Modifying (21) and (25), we have for $i \in \mathcal{U}$,

$$\bar{D}_n^{(i,\nu)}(t) = \frac{1}{n}\sum_{k=1}^{N_n^{(i)}(t)} \mathbf{1}(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \le t),$$
$$\bar{B}_n^{(i,\nu)}(t,y) = \frac{1}{n}\sum_{k=N_n^{(i)}((t-y)^+)+1}^{N_n^{(i)}(t)} \mathbf{1}(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} > t).$$

By the FWLLN of the TAP in §5.2, (41) and Theorem 3.1 of [35] (here Lemma 4 in the appendix), we quickly obtain the FWLLNs for $\bar{D}_n^{(i,\nu)}$ and $\bar{B}_n^{(i,\nu)}$, in

particular,

$$\bar{D}_n^{(j,\nu)}(t) \Rightarrow D_j^{(\nu)}(t) \equiv \int_0^t G_i(t-s)d\Lambda_i(s) \quad \text{in} \quad \mathbb{D}, \tag{46}$$

$$\bar{B}_n^{(j,\nu)}(t,y) \Rightarrow B^{(j,\nu)}(t,y) \equiv \int_{t-y}^t G_i^c(t-s)d\Lambda_i(s) \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}, \quad \text{as } n \to \infty. \tag{47}$$

where Λ_i satisfies the traffic-flow equation in (2).

5.3.2 FWLLNs for queue-related processes at OL queues.

Since all service-related processes have already been treated in §5.1, it remains to prove the FWLLNs for the queue-related processes, including the two-parameter queue-content $\bar{Q}_n^{(i)}$, HWT $W_n^{(i)}$, PWT $V_n^{(i)}$ and the abandonment process $\bar{A}_n^{(i)}$, for $i \in \mathcal{O}$.

Following §§6.2–6.3 in [28], we let $Q_n^{(i,*)}(t,x)$ be the two-parameter *queue-length process ignoring flows into service* (QLIFIS). Namely, $Q_n^{(i,*)}(t,x)$ denotes the number of customers in queue at t with elapsed waiting times no more than x, assuming no customer has been allowed to enter service since time 0. To obtain a representation of the queue-length process $Q_n^{(i)}$ which allows the usual flow into service, we now bound the second argument x by the HWT $W_n^{(i)}$, because no one waits longer than $W_n^{(i)}(t)$ at time t. Namely, we have $Q_n^{(i)}(t,x) = Q_n^{(i,*)}\left(t, x \wedge W_n^{(i)}\right)$. Because $Q_n^{(i,*)}$ is continuous in the second argument [28,35], we can apply the continuous mapping theorem if we can prove the convergence of the QLIFIS $Q_n^{(i,*)}$ and the HWT $W_n^{(i)}$.

The FWLLNs of the HWT $W_n^{(i)}$ and PWT $V_n^{(i)}$ have been established in §§6.6.1–6.6.3 of [28]. We now complete the proof by showing the convergence of the QLIFIS $Q_n^{(i,*)}$. For an OL queue i, we split $Q_n^{(i,*)}$ into two terms, corresponding to *old customers* (initially waiting in queue at time 0) and *new customers* (arrivals after time 0). In particular, we have

$$\bar{Q}_n^{(i,*)}(t,x) = \bar{Q}_n^{(i,o,*)}(t,x) + \bar{Q}_n^{(i,\nu,*)}(t,x) \tag{48}$$

where $\bar{Q}_n^{(i,o,*)}(t,x)$ denotes the LLN-scaled number of customers in queue at t with elapsed waiting times no more than x from those customers that are in queue at time 0, and $\bar{Q}_n^{(i,\nu,*)}(t,x)$ denotes the LLN-scaled number of customers in queue at t with elapsed waiting times no more than x from the new arrivals in the interval $[0,t]$. Paralleling the treatments for $\bar{B}_n^{(i,o)}(t,x)$ and $\bar{B}_n^{(i,\nu)}(t,x)$

in §5.1.2, we write

$$\bar{Q}_n^{(i,\nu,*)}(t,x) = \frac{1}{n} \sum_{k=N_n^{(i)}((t-x)^+)+1}^{N_n^{(i)}(t)} \mathbf{1}\left(\mathcal{E}_k^{(i,n)} + \mathcal{A}_k^{(i)} > t\right), \tag{49}$$

$$\bar{Q}_n^{(i,o,*)}(t,x) = \frac{1}{n} \sum_{k=1}^{Q_n^{(i)}(0,(x-t)^+)} \mathbf{1}\left(\xi_i^{(k)}(\chi_{n,i}^{(k)}) > t\right), \tag{50}$$

where $\mathcal{E}_k^{(i,n)}$ and $\mathcal{A}_k^{(i)}$ are the arrival and patience times of the k^{th} new customer (i.e., arrivals after time 0) so that $\mathcal{E}_k^{(i,n)} + \mathcal{A}_k^{(i)}$ is the time the k^{th} customer abandons from the queue if this customer does not enter service by then, $\left\{0 < \chi_{n,i}^{(1)} \le \chi_{n,i}^{(2)} \le \dots\right\}$ is the ordered sequence of elapsed waiting times of customers in queue at time 0, and $\left\{\xi_i^{(1)}(x), \xi_i^{(2)}(x), \dots\right\}$ is an i.i.d. sequence of random variables with ccdf

$$P\left(\xi_i^{(1)}(x) > t\right) = 1 - H_x^{(i)}(t) \equiv \frac{F_i^c(t+x)}{F_i^c(x)} \quad \text{for} \quad x > 0, \quad t \ge 0.$$

Because (49) and (50) are analogs of (25) and (26), we parallel the proofs for the FWLLNs of $\bar{Q}_n^{(i,\nu)}$ and $\bar{Q}_n^{(i,o)}$ in §5.1.2. Because $\bar{Q}_n^i(0,\cdot) \Rightarrow Q_i(0,\cdot)$ in $\mathbb{D}$ by Assumption 2, we have for $i \in \mathcal{O}$,

$$\bar{Q}_n^{(i,\nu,*)}(t,x) \Rightarrow Q^{(i,\nu,*)}(t,x) \equiv \int_{(t-x)^+}^{t} F_i^c(t-s)\lambda_i(s)ds, \tag{51}$$

$$\bar{Q}_n^{(i,o,*)}(t,x) \Rightarrow Q^{(i,o,*)}(t,x) \equiv \int_0^{(x-t)^+} q_i(0,y)\left(1 - H_y^{(i)}(t)\right)dy \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}. \tag{52}$$

Combining the FWLLN of $W_n^{(i)}$ and (51)–(52), we have

$$\begin{aligned}
\bar{Q}_n^{(i)}(t,x) &= \bar{Q}_n^{(i,*)}\left(t, x \wedge W_n^{(i)}(t)\right) \\
&\Rightarrow Q^{(i,*)}\left(t, x \wedge w_i(t)\right) \equiv Q^{(i,\nu,*)}\left(t, x \wedge w_i(t)\right) + Q^{(i,o,*)}\left(t, x \wedge w_i(t)\right) \\
&= \int_{(t-x\wedge w_i(t))^+}^{t} F_i^c(t-s)\lambda_i(s)ds + \int_0^{(x\wedge w_i(t)-t)^+} q_i(0,y)\left(1 - H_y^{(i)}(t)\right)dy
\end{aligned} \tag{53}$$

in $\mathbb{D}_{\mathbb{D}}$, as $n \to \infty$. Here the convergence follows from the continuous mapping theorem with composition and addition. It is not hard to see that the right-hand side of (53) coincides with the fluid limit $Q_i(t,x)$ defined in (3) and (8).

Given the FWLLNs of (i) the TAP $\bar{N}_n^{(0)}$, (ii) queue length $\bar{Q}_n^{(i)}(t,y)$ and (iii) ESP $\bar{E}_n^{(i)}$, we can easily obtain the FWLLN of the abandonment process for an OL queue $i \in \mathcal{O}$, defined as $\bar{A}_n^{(i)}(t) = \bar{Q}_n^{(i)}(0) + \bar{N}_n^{(i)}(t) - \bar{E}_n^{(i)}(t) - \bar{Q}_n^{(i)}(t)$, by the continuous mapping theorem with addition.

6 An $(M_t/H_2/s_t + E_2)^2/M_t$ Example

To provide engineering verification of Theorem 1, we now report the results of a simulation experiment. We consider a two-queue $(M_t/H_2/s_t + E_2)^2/M_t$ SQNet, with (i) NHPP arrival processes having sinusoidal arrival-rate functions $\lambda_n^{(0,i)}(t) = n\lambda_i^{(0)}(t)$, $\lambda_i^{(0)}(t) = a_i + b_i \sin(c_i\, t + \phi_i)$, (ii) two-phase *hyper-exponential* (H_2) service times with pdf $g_i(x) = p_i \cdot \mu_1^{(i)} e^{-\mu_1^{(i)}\, x} + (1 - p_i) \cdot \mu_2^{(i)} e^{-\mu_2^{(i)}\, x}$, (iii) constant staffing levels $s_n^{(i)}(t) = \lceil n s_i \rceil$, and (iv) two-phase *Erlang* (E_2) abandonment times with pdf $f_i(x) = 4\theta_i^2 x e^{-2\theta_i x}$, for $i = 1, 2$.

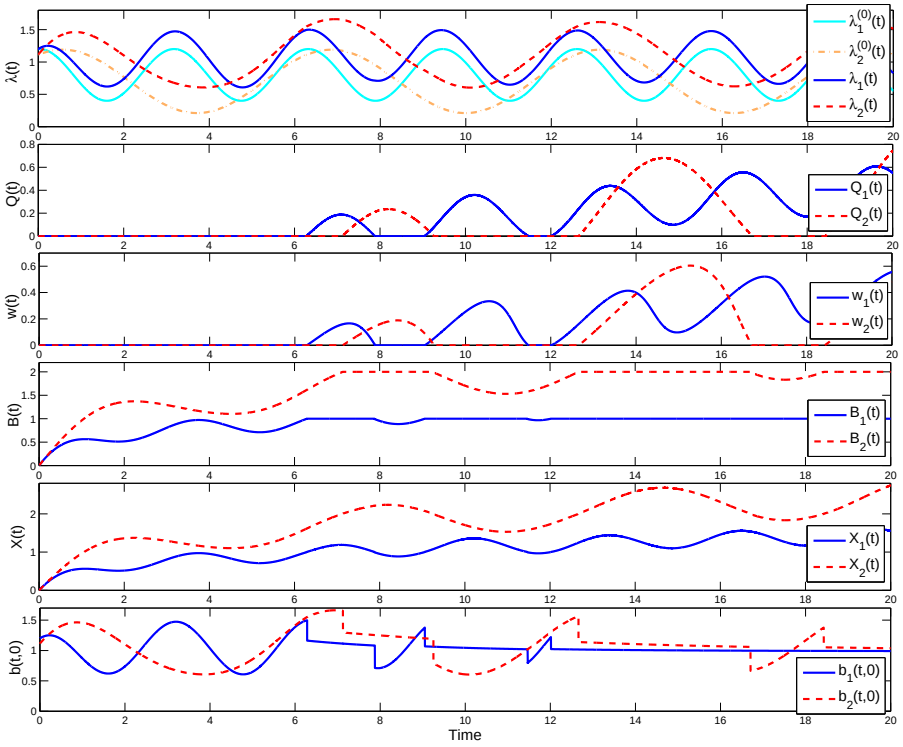


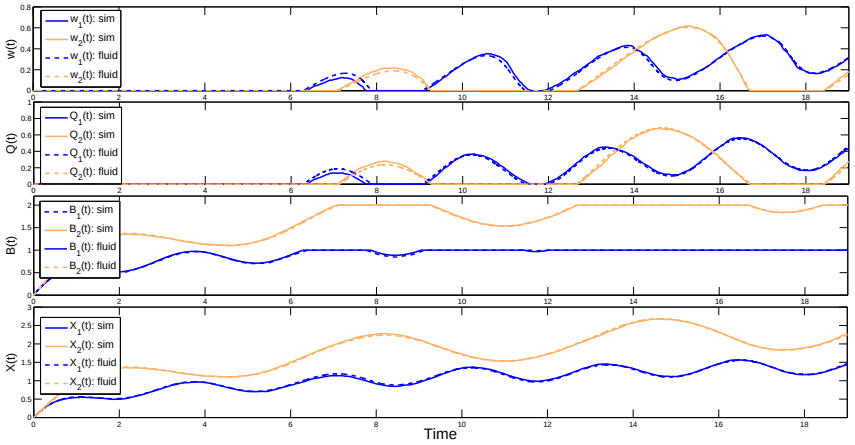
Fig. 2 Performance functions of the $(M_t/H_2/s_t + E_2)^2/M_t$ FQNet, including (i) TAR λ, (ii) queue content Q, (iii) PWT w, (iv) service content B, (v) total fluid X and (vi) rate into service $b(t, 0)$.

We let $a_1 = 0.8$, $b_1 = 0.4$, $a_2 = 0.7$, $b_2 = 0.5$, $\phi_1 = 1.5$, $\phi_2 = 1$, $c_1 = 2$, $c_2 = 1$, $\theta_1 = 0.5$, $\theta_2 = 0.3$, $s_1 = 1$, $s_2 = 2$, $\mu_1 = 1$, $\mu_2 = 0.5$, $p_1 = p_2 = 0.5(1-\sqrt{0.6})$, $\mu_1^{(i)} = 2p_i\mu_i$, $\mu_2^{(i)} = 2(1 - p_i)\mu_i$, for $i = 1, 2$. We have the service-time *squared coefficient of variation* SCV $c_s^2 = 4$ and abandonment-time SCV $c_a^2 = 1/2$. Let the routing probabilities $p_{1,1} = 0.15$, $p_{2,1} = 0.12$, $p_{1,2} = p_{2,2} = 0.2$.

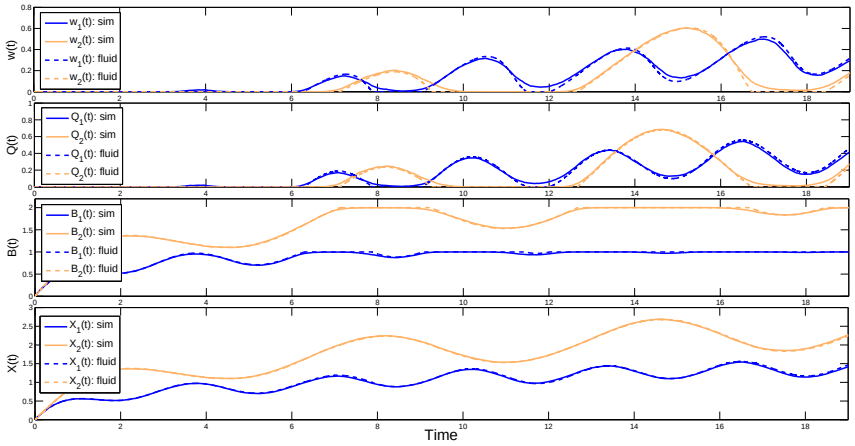
Figure 2 shows plots of key performance functions for $0 \le t \le T \equiv 20$, starting out empty, together with (i) EAPs $\lambda^{(0)}$ and TARs λ (Subplot 1), (ii) queue contents Q (Subplot 2), (iii) HWTs w (Subplot 3), (iv) service contents B (Subplot 4), (v) total fluid X (Subplot 5) and (vi) rate fluid enters

service $b(t,0)$ (last subplot). All performance functions are continuous except for $b(t,0)$: in UL intervals, $b(t,0) = \lambda(t)$; in OL intervals $b(t,0)$ is the unique solution of the FPE (7).

To verify the accuracy of the FQNet approximation, we conduct simulation comparisons in Figure 3 for LLN-scaled key performance functions of the SQNet, starting out empty (dashed lines): (i) HWT, (ii) number in queue, (iii) number in service and (iv) total number of customers.



(a) One sample path with the scale n = 3000



(b) Average of 100 sample pages with the scale n = 100

Fig. 3 A comparison of performance functions in the $(M_t/H_2/s_t + E_2)^2/M_t$ FQNet with simulation of the corresponding $(M_t/H_2/s_t + E_2)^2/M_t$ SQNet with (a) single sample paths and scale $n = 3000$, and (b) average of 100 paths and scale $n = 100$.

In Figure 3(a) we compare the fluid functions of the FQNet (the dashed lines) with the single sample paths of their corresponding LLN-scaled performance functions of the SQNet (the solid lines) with a large scale $n = 3000$. In Figure 3(b) we compare the fluid functions (the dashed lines) with the means of

the LLN-scaled performance functions of the SQNet (the solid lines, estimated by averaging 100 independent samples) with a smaller scale $n = 100$. Figure 3 verifies the remarkable performance of the FQNet approximation and provides practical confirmation of the FWLLN in Theorem 1. See [29] for additional experiments supporting the FQNet approximation.

7 Conclusion

We have established a many-server heavy-traffic limit theorem for a recently proposed deterministic fluid approximation for the $(G_t/GI/s_t + GI)^m/M_t$ queueing network [29], with a non-stationary non-Poisson arrival process, non-exponential service and abandonment times, time-varying staffing levels and Markovian (probabilistic) routing policy. Numerical analysis and simulation experiments have been developed in [25, 29] confirming the effectiveness of this deterministic fluid approximations. However, prior to this paper the functional law of large numbers of this $(G_t/GI/s_t+GI)^m/M_t$ fluid limit remained an open problem.

In this paper we solve this open problem by showing that all scaled performance processes, including the queue lengths (both in queue and in service), flows (of routing, abandonment and departure), and waiting times, jointly converge in distribution to their corresponding deterministic fluid functions conjectured in [25, 29], in appropriate functional spaces. We draw heavily on the proofs in [26] which focused on the $G_t/GI/s_t + GI$ single queue model. A key step here is to show the convergence of the total arrival process for all queues in the network. Our proof follows the compactness approach by (i) establishing the tightness in the appropriate functional space and (ii) showing that all convergent subsequences of the performance functions converge to the same desired limits.

Future work. Refining the fluid approximations which can be used to estimate the mean values of the performance functions, we next provide diffusion approximations for relevant models to quantify and approximate the stochastic fluctuations around the mean values; we do so in sequel papers [3, 16]. Because the probabilistic routing policy ignores a customer's routing history (the next queue to join depends only on the current location), in a sequel paper [14] we are motivated to seek alternative routing policies which incorporate the routing history.

References

1. Aksin, Z., Armony, M., Mehrotra, V. : The modern call center: A multi-disciplinary on operations management research. *Production and Operations Management.* **16**, 665–688 (2007)
2. Armony, M., S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. Working paper. (2014)
3. Aras, A. K., Liu, Y., Whitt, W.: Many-server heavy-traffic FCLT Limits for the $G_t/GI/s_t + GI$ queue. Working paper (2014).

4. Billingsley, P.: *Convergence of Probability Measures.* 2nd edn. Wiley, New York, (1999)
5. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., Zhao, L. : Statistical analysis of telephone call center: A queueing science perspective. *J. Amer. Statist. Assoc.* **100**, 36–50 (2005)
6. Dai, J., He, S., Tezcan, T.: Many-server diffusion limits for $G/Ph/n+GI$ queues. *Annals of Applied Probability.* **20**, 1854–1890 (2010)
7. Dai, J., He, S.: Customer abandonment in many-server queues. *Mathematics of Operations Research.* **35**, 347–362 (2010)
8. Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science.* **54** 324–338 (2008).
9. Feller, W.: *An Introduction to Probability Theory and Its Applications.* v2. John Wiley and Sons, New York, (1968).
10. Gans, N., Koole, G., Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects . *Manufacturing Service Operations Management.* **5**, 79–141 (2003)
11. Garnett, O., Mandelbaum, A., Reiman, M.I.: Designing a call center with impatient customers. *Manufacturing Service Operations Management.* **4**, 208–227 (2002)
12. Green, L.V., Kolesar, P.J., Whitt, W.: Coping with time-varying demand when setting staffing requirements for a service system. *Production Operations Management.* **16**, 13–39 (2007)
13. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential service. *Operations Research.* **29**, 567–588 (1981)
14. He, B., Liu, Y.: A fluid approximation for the multi-class network of queues with time-varying arrivals and prescribed routing paths. Working paper (2014)
15. He, B., Liu, Y.: Staffing to stabilize the tail probability of delay in service systems with time-varying demand. Working paper (2014)
16. Huang, C.-C., Liu, Y.: Diffusion limit for the time-varying many-server queueing networks with time-varying parameters. Working paper (2014)
17. Jacod, J., Shiryaev, A. N.: *Limit Theorems for Stochastic Processes.* Springer (1987)
18. Jelenkovic, P., Mandelbaum, A., Momcilovic, P.: Heavy-traffic limits for queues with many deterministic servers. *Queueing Systems Theory Applications.* **47**, 53–69 (2005)
19. Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. Server Staffing to Meet Time-Varying Demand. *Management Science.* **42** 1383–1394 (1996)
20. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. *Annals of Applied Probability.* **20**, 2204–2260 (2010)
21. Kaspi, H., Ramanan, K.: Law of large number limits for many-server queues. *Annals of Applied Probability.* **21**, 33–114 (2011)
22. Kim, S.-H., W. Whitt. 2014. Are Call Center and Hospital Arrivals Well Modeled by Nonhomogeneous Poisson Processes? Columbia University, New York, NY.
23. Krichagina, E.V., Puhalskii, A.A.: A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems.* **25**, 235–280 (1997)
24. Liu, Y., Whitt, W.: A network of time-varying many-server fluid queues with customer abandonment. *Operations Research.* **59**, 835–846 (2011)
25. Liu, Y., Whitt, W.: The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems.* **71**, 405–444 (2012)
26. Liu, Y., Whitt, W.: A many-server fluid limit for the $G_t/GI/s_t + GI$ queuing model experiencing periods of overloading. *Operations Research Letters.* **40**, 307–312 (2012)
27. Liu, Y., W. Whitt. Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research.* **60** 1551–1564 (2012).
28. Liu, Y., Whitt, W.: Many-server heavy-traffic limit for queues with time-varying parameters. *Annals of Applied Probability.* **24**, 378–421 (2014)
29. Liu, Y., Whitt, W.: Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing.* **26**, 59–73 (2014)
30. Liu, Y., W. Whitt. Stabilizing Performance in Network of Queues with Time-Varying Arrival Rates. *Probability in the Engineering and Informance Sciences.* Forthcoming (2014)
31. Liu, Y., W. Whitt. Stabilizing Performance in Many-Server Queues with Time-Varying Arrivals and Customer Feedback. Working paper (2014)
32. Mandelbaum, A., Massey, W. A., Reiman, M. I.: Strong approximations for Markovian service networks. *Queueing Systems.* **30**, 149–201 (1998)

33. Mandelbaum, A., Momcilovic, P.: Queues with many servers: The virtual waiting-time process in the QED regime. *Mathematics of Operations Research.* **33**, 561–586 (2008)
34. Mandelbaum, A., Momcilovic, P.: Queues with many servers and impatient customers. *Mathetmatics of Operations Research.* **37** , 41–65 (2012)
35. Pang, G., Whitt, W.: Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems.* **4**, 325–364 (2010)
36. Puhalskii, A.A., Reed, J.: On many-server queues in heavy-traffic. *Annals of Applied Probability.* **20**, 129–195 (2010)
37. Pang, P., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Proba. Surv.* **4**, 193–267 (2007)
38. Puhalskii, A.A., Reiman, M.I.: The multiclass $GI/PH/N$ queue in Halfin-Whitt regime. *Advances in Applied Probability.* **32**, 564–595 (2000)
39. Reed, J.: The $G/GI/N$ queue in the Halfin-Whitt regime. *Annals of Applied Probability.* **20**,2211–2269 (2009)
40. Shi, P., M. Chou, J. G. Dai, D. Ding, J. Sim. Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. Working paper. (2014)
41. Skorohod, A. V.: Limit theorems for stochastic processes. *Theory of Probability and Its Applications.* **1**, 261–290 (1956)
42. Whitt, W.: *Stochastic-Process Limits.* Springer, New York (2002)
43. Whitt, W.: Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research.* **30**, 1–27 (2005)
44. Whitt, W.: Fluid models for multiserver queues with abandonments. *Operations Research.* **54**, 37–54 (2006)
45. Whitt, W.: Proofs of the martingale FCLT. *Probability Surveys.* **4**, 268–302 (2007)
46. Yom-Tov, G., A. Mandelbaum. 2014. The Erlang-R queue: Time-Varying QED Queues with Re-entrant Customers in Support of Healthcare Staffing. Working paper. The Technion, Israel. (2014)
47. Zeltyn, S., A. Mandelbaum. Call Centers with Impatient Customers: Many-Server Asymptotics of the $M/M/n+G$ Queue. *Queueing Systems.* **51** 361–402 (2005)
48. Zhang, J.: Fluid models of many-server queues with abandonment. *Queueing Systems.* **73**, 147–193 (2013)

APPENDIX

to

A Functional Weak Law of Large Numbers for the Time-Varying $(G_t/GI/s_t + GI)^m/M_t$ queueing network

by

A. Korhan Aras and Yunan Liu

A Overview

This appendix contains material supplementing the main paper. We review the FWLLN for the $G_t/GI/\infty$ infinite-server queue in Appendix B. Appendix C contains the classical tightness criteria (Theorem 2) followed by the details of the proof of Lemma 1. Appendix D contains the proof of the convergence result in (23). Next, in Appendix E, we show that the first terms in (31) and (32) are asymptotically negligible using a LLN for triangular arrays (see Theorem 9.1 of [9]; also see Theorem 3 here). In §F, we provide the proof of Lemma 2. In Appendix G, we summarize all acronyms used in the main paper.

B FWLLN for the $G_t/GI/\infty$ queue

In this section we review the FWLLN for the $G_t/GI/\infty$ queue [35]. Consider a sequence of $G_t/GI/\infty$ infinite-server queueing model indexed by n, having a non-stationary arrival process (the G_t), i.i.d. service times following a cdf G and infinite servers. For the n^{th} model, let $N_n(t)$ be the number of arrivals by time t and let $X_n(t,y)$ be the number of customers in service at time t with elapsed service times at most y.

Assumption 4 (*FWLLN for the arrival process*) *There exist a nondecreasing function $\Lambda_i(t)$ with non-negative derivative $\lambda_i(t)$, $1 \le i \le m$, such that*

$$\bar{N}_n(t) \equiv n^{-1} N_n(t) \Rightarrow \Lambda_i(t) \equiv \int_0^t \lambda_i(u)du.$$

Suppose the system is initially empty, i.e., $X_n(0,x) = 0$, $x > 0$. The two-parameter queue-length process can be represented as

$$\begin{aligned} X_n(t,x) &= \sum_{i=N_n(t-x)+1}^{N_n(t)} \mathbf{1}\left(\tau_{n,i} + \eta_i > t\right). \qquad (54) \\ &\equiv X_{n,1}(t,x) + X_{n,2}(t,x) + X_{n,3}(t,x), \quad 0 \le x \le t, \end{aligned}$$

where

$$\begin{aligned} X_{n,1}(t,x) &\equiv \sqrt{n} \int_{t-x}^t G^c(t-s)\, d\hat{N}_n(s), \\ X_{n,2}(t,x) &\equiv \sqrt{n} \int_{t-x}^t \int_0^\infty \mathbf{1}(x+s>t)\, d\hat{K}_n(\Lambda(s),x), \\ X_{n,3}(t,x) &\equiv n \int_{t-x}^t G^c(t-s)\, d\Lambda(s), \end{aligned}$$

where

$$\hat{N}_n(s) \equiv \sqrt{n}\left(\bar{N}_n(s) - \Lambda(s)\right) \quad \text{and} \quad \hat{K}_n(t,x) \equiv \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor} \mathbf{1}\left(\eta_i \le x\right) - G(x)\right)$$

is a Kiefer process, see [35,23] for details. The next lemma gives the FWLLN for the two-parameter queue-length $X_n(t,x)$. Let $\bar{X}_n \equiv X_n/n$.

Lemma 4 (*FWLLN for the $G_t/GI/\infty$ queue, Theorem 3.1 in* [35])
If Assumption 4 is satisfied, then $(\bar{N}_n, \bar{X}_n) \Rightarrow (\Lambda, X)$ in $\mathbb{D} \times \mathbb{D}_{\mathbb{D}}$ as $n \to \infty$, where

$$X(t,x) = \int_{(t-x)^+}^{t} G^c(t-s)d\Lambda(s).$$

C Proof of Lemma 1

The following theorem provides the classical characterization of C-tightness. Then we prove that the LLN-scaled service-completion processes in Lemma 1 satisfy the conditions of Theorem 2.

Theorem 2 (*Classical characterization of C-tightness, Theorem 3.2 of [45]) A sequence of stochastic processes $\{X_n, n \geq 1\}$ is tight if and only if*
(i) The sequence $\{X_n, n \geq 1\}$ is stochastically bounded in $\mathbb{D}$
and
(ii) for each $T > 0$ and $\eta > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbf{P}\left(w(X_n, \delta, T) > \eta\right) = 0, \tag{55}$$

where for $x \in \mathbb{D}$,

$$w(x, \delta, T) = \sup\left\{w(x, [t, t+\delta]) : 0 \leq t \leq (t+\delta) \wedge T\right\}, \tag{56}$$

$$w(x, I) = \sup_{s,t \in I} |x(s) - x(t)|, \quad I \subset \mathbb{R}_+. \tag{57}$$

Proof of Lemma 1. We use the classical criterion in Theorem 2 to prove the C-tightness of $\{\bar{D}_n^{(i)}, i \in \mathcal{O}\}$, i.e., we show that, for each $i \in \mathcal{O}$, the process $\bar{D}_n^{(i)}(t)$ is stochastically bounded in $\mathbb{D}$ (satisfying Condition (i) of Theorem 2) and the modulus of continuity condition holds (satisfying Condition (ii) of Theorem 2), that is, for each $T > 0$ and $\eta > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \mathbf{P}\left(w(\bar{D}_n^{(i)}, \delta, T) > \eta\right) = 0. \tag{58}$$

Consider an OL queue i (i.e., $i \in \mathcal{O}$), with $\lceil ns^{(i)}(t) \rceil$ servers for simplicity. Successive service completions from each server form a (delayed) renewal process since the service times are i.i.d. random variables with cdf GI. Hence, we can represent the service-completion process of the i^{th} queue as

$$D_n^{(i)}(t) = \sum_{k=1}^{\lceil ns^{(i)}(t) \rceil} D_n^{(i,k)}(t), \qquad t \geq 0. \tag{59}$$

where $D_n^{(i,k)}(t)$, $k = 1, \ldots, \lceil ns^{(i)}(t) \rceil$, are renewal counting processes associated with the departure processes from each server in the i^{th} queue.

Suppose that $\{\bar{D}_n^{(i)}\}_{n \geq 1}$ is not stochastically bounded in $\mathbb{D}$. Then there exists $\epsilon_0 > 0$ such that $\mathbf{P}(\bar{D}_n^{(i)}(0) > \eta) \geq \epsilon_0$ for any $\eta > 0$. Integrating both sides with respect to η implies that $\mathbf{E}[\bar{D}_n^{(i)}(0)] = \infty$. On the other hand, by (59), we have

$$\begin{aligned}
\mathbf{E}[\bar{D}_n^{(i)}(0)] &= n^{-1} \sum_{k=1}^{\lceil ns^{(i)}(0) \rceil} \mathbf{E}[D_n^{(i,k)}(0)] \\
&\leq n^{-1} \lceil ns^{(i)}(0) \rceil \max_{1 \leq k \leq \lceil ns^{(i)}(0) \rceil} \left\{\mathbf{E}[D_n^{(i,k)}(0)]\right\} \\
&= \lceil s^{(i)}(0) \rceil \max_{1 \leq k \leq \lceil ns^{(i)}(0) \rceil} \left\{\mathbf{E}[D_n^{(i,k)}(0)]\right\} < \infty\,.
\end{aligned} \tag{60}$$

since $\mathbf{E}[D_n^{(i,k)}(t)] < \infty$ almost surely for $t \geq 0$, $k = 1, \ldots, \lceil ns^{(i)}(0) \rceil$. Hence, a contradiction. Therefore, we conclude that for any $\epsilon > 0$, there exists $\eta > 0$ such that $\mathbf{P}(\bar{D}_n^{(i)}(0) > \eta) < \epsilon$. This proves the first condition in Theorem 3.2 of [45].

Next we verify Condition (58). Since $\bar{D}_n^{(i)}(t)$ is nondecreasing in t for each $i \in \mathcal{O}$ and $n \geq 1$, (57) reduces to $\bar{D}_n^{(i)}(b) - \bar{D}_n^{(i)}(a)$ for $[a, b] \subset \mathbb{R}_+$. Consequently, (56) for the process $\bar{D}_n^{(i)}(t)$ becomes

$$w(\bar{D}_n^{(i)}, \delta, T) = \sup \left\{ \bar{D}_n^{(i)}(t+\delta) - \bar{D}_n^{(i)}(t) : 0 \leq t \leq (t+\delta) \wedge T \right\}.$$

Observe that $w(\bar{D}_n^{(i)}, \delta, T) \downarrow 0$ as $\delta \downarrow 0$ for each $n \geq 1$ since $\bar{D}_n^{(i)}(t)$ is a finite sum of renewal processes (see (59)). This implies that $\mathbf{P}(w(\bar{D}_n^{(i)}, \delta, T) > \eta) \downarrow 0$ as $\delta \downarrow 0$ for any $\eta > 0$, $n \geq 1$. Consequently,

$$\limsup_{n \to \infty} \mathbf{P}\left(w(\bar{D}_n^{(i)}, \delta, T) > \eta \right) \to 0 \quad \text{as} \quad \delta \downarrow 0.$$

The proof of the C-tightness of the sequence $\{\bar{D}_n^{(i)}\}_{n \geq 1}$ is completed. ∎

D Proof of the convergence result in (23).

Since the sequence $\{\bar{E}_n^{(i)}\}$ is C-tight in $\mathbb{D}$ by Lemma 1, every subsequence has a convergent subsequence. Suppose we have such a convergent subsequence. We do not introduce a special notation for the subsequence and, without loss of generality, we label that subsequence as $\{\bar{E}_n^{(i)}(t)\}$ so that we have $\bar{E}_n^{(i)} \Rightarrow E^{(i)}$ in $\mathbb{D}$ for all $i \in \mathcal{O}$ where the limit $E^{(i)}$ is yet to be characterized.

From (21), we see that the service-completion process of new customers has the same mathematical form as the departure process from an infinite-server queue with arrival process $E_n^{(i)}(t)$ and service times following cdf G_i. We can use directly apply Lemma 4 (also see Theorem 3.1 of [35]). Consequently, we have

$$\bar{D}_n^{(i,\nu)}(t) \Rightarrow D^{(i,\nu)}(t) \equiv \int_0^t G_i(t-s) b_i(s,0) ds, \ t \in [0,T] \quad \text{for all} \quad i \in \mathcal{O}. \tag{61}$$

Combining (22), (61), (28) with $y = \infty$ and by applying the continuous mapping theorem with addition, we obtain weak convergence of the sequence $\{\bar{E}_n^{(i)}(t)\}$ to an integral equation

$$\bar{E}_n^{(i)}(t) \Rightarrow s^{(i)}(t) - s^{(i)}(0) + B^{(i,o)}(0) - B^{(i,o)}(t) + \int_0^t G_i(t-s) b_i(s,0) ds. \tag{62}$$

For each $i \in \mathcal{O}$, the derivative of (62) satisfies the fixed point equation (7), which has a unique solution (see [25]). Since the choice of the convergent subsequence is arbitrary, the derivative of the limit of every convergent subsequence of $\{\bar{E}_n^{(i)}\}$ must satisfy (7). Hence, we have the full convergence of $\{\bar{E}_n^{(i)}\}$ and $\{\bar{D}_n^{(i,\nu)}\}$ for all $i \in \mathcal{O}$. ∎

E LLN for non-identically distributed triangular arrays

We first review an LLN results for non-identically distributed triangular arrays (e.g., see Theorem 9.1. of [9]).

Let $\{X_{k,n}\}$, $k = 1, \ldots, n$, be a general triangular array of random variables with cdf $F_{k,n}$. Assume that the random variables in each row of the triangular array are mutually independent. Define S_n as the partial sum of $X_{k,n}$, i.e., $S_n = \sum_{k=1}^n X_{k,n}$. Also define $\tau_s(X_{k,n})$ as the truncated version of $X_{k,n}$, where $\tau_s(x) = x$ if $|x| \leq s$; $\tau_s(x) = -s$ if $x < -s$;

$\tau_s(x) = s$ if $x > s$. Let $S_n^s = \sum_{k=1}^n \tau_s(X_{k,n})$. Consider the following condition: for arbitrary $\eta > 0$ and $\epsilon > 0$

$$\mathbf{P}(|X_{k,n}| > \eta) < \epsilon \qquad k = 1, \ldots, n \tag{63}$$

for all n sufficiently large. Now, we are ready to state the theorem.

Theorem 3 (*LLN for non-iid random variables, Theorem 9.1 of* [9]) *If Condition* (63) *holds, then there exist constants* b_n *such that* $S_n - b_n \to 0$ *in probability if and only if*

$$\sum_{l=1}^n \mathbf{P}\{|X_{l,n}| > \eta\} \to 0, \qquad \sum_{l=1}^n Var(\tau_s(X_{l,n})) \to 0 \qquad as \quad n \to \infty \tag{64}$$

for each $\eta > 0$ *and each truncation level* s. *In this case, one may take* $b_n = \mathbf{E}[S_n^s]$.

We next make use of Theorem 3 to prove that the first terms of (31) and (32) converge to 0. Conditioning on $\{\zeta_{n,i}^{(l)}\}$, the first terms of (31), (32) are LLN-scaled sum of independent non-identically distributed zero-mean random variables with values in $[-1, 1]$. Therefore, we can use Theorem 3 to prove convergence. We will later uncondition to obtain the desired result. The proof of the convergence of the first terms of (31) and (32) are similar. Therefore, we only provide a proof for the latter.

In our case, we have from (32)

$$X_{l,n} \equiv n^{-1}\left(\delta_{i,j}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)})\right) \qquad and \qquad S_n = \sum_{l=1}^n X_{l,n} \quad \text{for all} \quad n \geq 1. \tag{65}$$

for fixed $i \in \mathcal{O},\ j \in \{1, \ldots, m\}$. Using $\tau_s(\cdot)$, we define the truncation of $X_{l,n}$ and the partial sum of truncated variables accordingly.

Conditioning on the sequence $\{\zeta_{n,i}^{(l)}\}$, we have, by Markov inequality,

$$\mathbf{P}\{|X_{l,n}| > \eta\} \leq \frac{\mathbf{E}[|X_{l,n}|^2]}{\eta^2} = \frac{P_{i,j}(\zeta_{n,i}^{(l)})(1 - P_{i,j}(\zeta_{n,i}^{(l)}))}{n^2\eta^2} \leq \frac{1}{n^2\eta^2}$$

which implies that

$$\sum_{l=1}^n \mathbf{P}\{|X_{l,n}| > \eta\} \leq \frac{1}{n\eta^2} \to 0 \quad \text{as} \quad n \to \infty.$$

As for the second term in (64), we have that $Var(\tau_s(X_{l,n})) = \mathbf{E}[(\tau_s(X_{l,n}))^2]$ since $\mathbf{E}[X_{l,n}] = 0$ for $n \geq 1,\ 1 \leq l \leq n$. The desired result easily follows because

$$\mathbf{E}[(\tau_s(X_{l,n}))^2] \leq \mathbf{E}[(X_{l,n})^2] \leq \frac{1}{n^2} \quad \text{for all} \quad s > 0$$

which implies

$$\sum_{l=1}^n Var(\tau_s(X_{l,n})) \leq \sum_{l=1}^n \frac{1}{n^2} = \frac{1}{n} \to 0 \qquad \text{as} \quad n \to \infty.$$

for all $s > 0$. Since $\mathbf{E}[S_n^s] = 0$, we have $S_n \to 0$ in probability. More explicitly,

$$\sum_{l=1}^n \frac{\delta_{i,j}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)})}{n} \to 0 \quad \text{in probability} \quad \text{as} \quad n \to \infty. \tag{66}$$

The arguments of unconditioning follows from the arguments on p.255 of [37]. In particular, by Skorohod representation theorem, we may assume that the scaled enter-service process converges in $\mathbb{D}$ almost surely. Then we deduce that the above convergence holds whenever the enter-service process converges almost surely in $\mathbb{D}$. Therefore, the conditional convergence is obtained by applying the generalized continuous mapping theorem, e.g., Theorem 3.4.4. of [42].

F Proof of Lemma 2

Let $\{x_n\}$ be a convergent sequence such that $d(x_n, x) \to 0$ where $d(\cdot,\cdot)$ is the Skorohod J_1 metric [4,17,42]. Then we have $||x_n - x \circ \lambda_n||_T \to 0$ and $||\lambda_n - e||_T \to 0$ where e is the identity function, i.e., $e(t) = t$ for $t \geq 0$, and $||\cdot||$ is the uniform norm over the interval $[0, T]$. Let $M = \sup_{0\leq t\leq T} \big|x(t)\big|$. Our goal is to show that $d(\phi(x_n), \phi(x)) \to 0$. Consider

$$
\begin{aligned}
&\big|\phi(x_n)(t) - \phi(x)(\lambda_n(t))\big| \\
&= \left|P_{i,j}(t)x_n(t) - \int_0^t x_n(u)dP_{i,j}(u) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) + \int_0^{\lambda_n(t)} x(u)dP_{i,j}(u)\right| \\
&\leq \big|P_{i,j}(t)x_n(t) - P_{i,j}(\lambda_n(t))x(\lambda_n(t))\big| + \left|\int_0^{\lambda_n(t)} x(u)dP_{i,j}(u) - \int_0^t x_n(u)dP_{i,j}(u)\right| \\
&\leq P_{i,j}(t)\big|x_n(t) - x(\lambda_n(t))\big| + \big|P_{i,j}(t) - P_{i,j}(\lambda_n(t))\big|\big|x(\lambda_n(t))\big| \\
&+ \left|\int_0^{\lambda_n(t)} x(u)dP_{i,j}(u) - \int_0^t x_n(u)dP_{i,j}(u)\right| \\
&\leq P_{i,j}(t)\big|x_n(t) - x(\lambda_n(t))\big| + \big|P_{i,j}(t) - P_{i,j}(\lambda_n(t))\big|\big|x(\lambda_n(t))\big| \\
&+ \left|\int_0^t x(\lambda_n(s))dP_{i,j}(\lambda_n(s)) - \int_0^t x_n(s)dP_{i,j}(s)\right| \\
&\leq P_{i,j}(t)\big|x_n(t) - x(\lambda_n(t))\big| + \big|P_{i,j}(t) - P_{i,j}(\lambda_n(t))\big|\big|x(\lambda_n(t))\big| \\
&+ \left|\int_0^t x(\lambda_n(s))d(P_{i,j}(\lambda_n(s)) - P_{i,j}(s))\right| + \left|\int_0^t x(\lambda_n(s))dP_{i,j}(s) - \int_0^t x_n(s)dP_{i,j}(s)\right|
\end{aligned}
$$

$$
\begin{aligned}
&\leq P_{i,j}(t)\big|x_n(t) - x(\lambda_n(t))\big| + M\big|P_{i,j}(t) - P_{i,j}(\lambda_n(t))\big| \\
&+ M\big|(P_{i,j}(\lambda_n(t)) - P_{i,j}(t)) - (P_{i,j}(\lambda_n(0)) - P_{i,j}(0))\big| + \int_0^t \big|x(\lambda_n(s)) - x_n(s)\big|dP_{i,j}(s) \\
&\leq 2||x_n - x \circ \lambda_n(t)|| + M\big|P_{i,j}(t) - P_{i,j}(\lambda_n(t))\big| \\
&+ M\big|(P_{i,j}(\lambda_n(t)) - P_{i,j}(t)) - (P_{i,j}(\lambda_n(0)) - P_{i,j}(0))\big|.
\end{aligned}
$$

The convergence of the first term follows from the convergence of $x_n \to x$ in $\mathbb{D}$. The convergence of the second and the third terms follows from the fact that $P_{i,j}(t)$ is continuous in t and $\lambda_n \to e$ uniformly over the interval $[0, T]$. ∎

G Acronyms

We now summarize all acronyms used in the main paper in the following table.

Table 1 Summary of frequently used acronyms (in alphabetic order).

Acronym	Meaning
ccdf	complementary cumulative distribution function
cdf	cumulative distribution function
CL	critically loaded
EAP	external arrival process
EAR	external arrival rate
ESP	enter-service process
FCFS	first come first served
FCLT	functional central limit theorem
FQNet	fluid queue network
FPE	fixed-point equation
FWLLN	functional weak law of large numbers
HWT	head-of-line waiting time
i.i.d.	independent and identically distributed
IRP	internal routing process
LLN	law of large numbers
MSHT	many-server heavy-traffic
NHPP	non-homogeneous Poisson process
ODE	ordinary differential equation
OL	overloaded
pdf	probability density function
PWT	potential waiting time
QLIFIS	queue length ignoring flow into service
SCP	service-completion process
SQNet	stochastic queueing network
TAP	total arrival process
TAR	total arrival rate
UL	underloaded