OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Transitory Queueing Networks

Harsha Honnappa
School of Industrial Engineering, Purdue University, West Lafayette IN 47906. Email: honnappa@purdue.edu

Rahul Jain
EE & ISE Departments, University of Southern California, Los Angeles, CA 90089. Email: rahul.jain@usc.edu

Queueing networks are notoriously difficult to analyze *sans* both Markovian and stationarity assumptions. Much of the theoretical contribution towards performance analysis of time-inhomogeneous single class queueing networks has focused on Markovian networks, with the sole exception of recent work in Liu and Whitt (2011). In this paper, we introduce *transitory queueing networks* as a model of inhomogeneous queueing networks, where a large, but finite, number of jobs arrive at queues in the network over a fixed time horizon. The queues offer FIFO service, and we assume that the service rate can be time-varying. The non-Markovian dynamics of this model complicate the analysis of network performance metrics, necessitating approximations. In this paper we develop fluid and diffusion approximations to the number-in-system performance metric by scaling up the number of external arrivals to each queue, following Honnappa et al. (2014). We also discuss the implications for bottleneck detection in tandem queueing networks.

Key words: Strategic arrivals, Population games, Game theory, Queueing Networks. *OR/MS subject classification:* Games/group decisions: Bidding/auctions, Natural resources: Energy, Communications.

Area of Review: Revenue Management

History: Submitted:.

1. Introduction

Single class queueing networks (henceforth 'queueing networks') have been studied extensively in the literature, with much effort focused on understanding the steady-state joint distribution of the

state of the network (typically defined as the number of jobs in each queue). In this paper, we consider a variation of the generalized Jackson network model ((Chen and Yao 2001, Chapter 7)) of a single class queueing network where a finite, but large, number of jobs arrive at the network from an extraneous source. We characterize fluid and diffusion approximations to the queue length state process, as the population size scales to infinity. The motivation for this model include transportation networks, manufacturing and service networks. To the best of our knowledge, this is the first time transitory networks have been studied in the literature, despite its wide applicability.

Bottleneck detection and prediction is likely the most important question that a system operator faces. Heavy traffic theory has been immensely succesful at characterizing steady state bottlenecks in very general queueing networks, under minimal network data assumptions. However, there are many circumstances, ranging from manufacturing, to healthcare, transportation and computing, where transient bottleneck detection and analysis is critical. For example, consider a facility that manufactures a jet engine. Each part of the engine is produced and assembled in a separate machine that requires some human supervision. Typically, there is a fixed, finite, number of jobs that need to be completed in a shift spanning a few hours. Furthermore, jobs cannot be carried over to the next shift. It is typically the case that the shift horizon is not long enough for the system to reach a steady state. In this purely transient or 'transitory' setting, it is common for the bottleneck node to change over the shift horizon. As a consequence the plant manager moves workers around trying to ease bottlenecks, increasing costs and increasing the likelihood of job overages. Another example is in the healthcare setting where patient diagnosis relies on a number of tests that must be done with different machines. Furthermore, in many time critical settings, the horizon within which tests must be conducted is fixed. An important question in these situations is whether transient bottlenecks can be accurately predicted, given network data.

Note that the standard definition of a bottleneck is a 'capacity level' one, defined in terms of long-term averages. This definition, of course, is not satisfactory in the transitory setting where steady states might not be reached. Second, standard heavy-traffic analysis completely ignores

non-stationarities in the network data, which is generally prevalent in most systems, making it an inappropriate analytical tool to use when non-stationarities are prevalent. To address these issues, we introduce the *transitory queueing network* as a model of a single class queueing network where a finite number of jobs arrive in a finite time horizon. The finite population effect implies that a steady state analysis is not feasible, and instead we must focus on transient distributions. Since transient analysis is non-trivial, we characterize the transient distribution of the network state (defined as the vector of queue lengths at each node) by developing fluid and diffusion approximations in an appropriately defined high-intensity regime. This provides first- and second-order approximate characterizations of the network performance.

Transitory queueing networks consist of a number of infinite buffer, FIFO, single server queues (a.k.a. 'nodes') interconnected by customer routes. We assume that the routing matrix satisfies a so-called *Harrison-Reiman (H-R)* condition that the matrix has a spectral radius of less than one. On completion of service at a particular node, a customer is routed to another node or exits the network altogether. Jobs enter nodes at random time epochs modeled as the ordered statistics of independent and identically distributed (i.i.d.) random variables. The arrival times at different nodes can be correlated. We assume that the service processes at different nodes are independent with time-inhomogeneous service rates, and modeled as a time change of a unit rate renewal counting process, generalizing the construction of a time-inhomogeneous Poisson process. The transient analysis of generalized Jackson networks is non-trivial, as noted before. The conventional heavy-traffic diffusion approximation that relied on long-run average rates has been used to approximate the evolution of the state process. However, these rates do not exist in transitory queueing networks. In this paper we develop a 'population acceleration' approximation, by increasing the number of jobs arriving at the network in the interval of interest to infinity, and suitably scaling (or 'accelerating') the service process in each queue by the population size.

To be precise, we consider a sequence of queueing networks wherein n jobs arrive at each node that receives external traffic in the nth network. We first establish a functional strong law of large

numbers (FSLLLN) to the arrival process, as the population size scales to infinity, by generalizing the Glivenko-Cantelli Theorem to multiple dimensions. Similarly, for the functional central limit theorem (FCLT) we introduce the notion of a multidimensional Gaussian bridge process, and show that the diffusion scaled arrival process converges to a Gaussian bridge in the large population limit. We also assume the service processes satisfy a FSLLN and FCLT in the population acceleration scale. The queue length fluid limit is shown to be equal to the oblique reflection of the difference of the arrival and service processes (or the 'netput' process). The diffusion limit turns out to be complicated, and it is shown to be a reflection of a multidimensional diffusion bridge process - however, the reflection is through a directional derivative of the oblique reflection of the netput in the direction of the diffusion limit of the netput process. This is a highly non-standard result. Indeed, it is only in the recent past that Mandelbaum and Ramanan Mandelbaum and Ramanan (2010) have investigated the existence of a directional derivative to the oblique reflection map.

Leveraging the results of Mandelbaum and Ramanan (2010), we can only establish a pointwise diffusion limit for an arbitrary transitory generalized Jackson network. This is due to the fact that the directional derivative limit can have sample paths with discontinuities that are both right- and left-discontinuous. Thus, establishing convergence in a sample path space under a suitably weak topology such as M_1, for instance, is not straightforward. Instead, we focus on the case of tandem queueing networks, with uniform and unimodal arrival time distribution functions. In this case, we show that the discontinuities in the limit are either right or left continuous, and hence we can establish M_1 convergence. Using these approximations, we next address the question of bottleneck prediction in a transitory network in a high intensity regime. We generalize the standard definition of a bottleneck in a single class network, defined as the queues whose fluid arrival rate exceeds the fluid service capacity to the transitory setting.

Our results complement the existing literature on the analysis of single-class queueing networks by establishing the following results:

(i) we develop a large population approximation framework for studying single class queueing networks in a transitory setting, complementing and extending Markovian network analyses to non-Markovian queueing networks,

(ii) Our diffusion approximations use the recently developed directional derivative oblique reflection map in Mandelbaum and Ramanan (2010) to establish a diffusion scale approximation; this is substantially different from the conventional heavy-traffic approximations used to study single-class queueing networks, and

(iii) we study the evolution of the bottleneck process over the time horizon, identifying the bottleneck station as time progresses. This analysis extends the standard bottleneck analyses, where bottlenecks are identified in terms of the long-term average arrival and service rates.

1.1. Related Literature

There has been significant interest in the analysis of single class queueing networks. Under the assumption of Poisson arrival and service processes, Jackson Jackson (1957) showed that the steady state distribution of the state of the network (the number of jobs waiting in each node) is equal to the product of the distribution of the state of each node in the network. This desirable property implies that, in steady state, the network exhibits a nice independence property. This property does not extend to networks with general arrival and service processes; these are also known as generalized Jackson networks.

Reiman first established the heavy-traffic diffusion approximation to open generalized Jackson networks in Reiman (1984). In particular, the diffusion approximation is shown to be a multi-dimensional reflected Brownian motion in the non-negative orthant, reflected through the oblique reflection mapping. Such reflection maps have come to be called as Harrison-Reiman maps following the work in Harrison and Reiman (1981). Chen and Mandelbaum Chen and Mandelbaum (1991a,b) characterize a homogeneous fluid network, as well as establishing fluid and diffusion approximations. The analysis of non-stationary and time inhomogeneous queueing systems is non-trivial in general. For single server queues, see Keller (1982), Massey (1985), Mandelbaum and Massey (1995) among others. In Honnappa et al. (2014, 2013) we develop fluid and diffusion models of transitory single server queues. For networks of queues, Mandelbaum et al. (1998) develops strong approximations to queueing networks with nonhomogeneous Poisson arrival and service processes.

In Duffield et al. (2001), the authors study the offered load process in a bandwidth sharing network, with nonstationary traffic and general bandwidth requirements. More recently, Liu and Whitt Liu and Whitt (2011) study a network of non-Markovian fluid queues with time-varying traffic and customer abandonments. To be precise, they consider a $(G_t/M_t/s_t + GI_t)^m/M_t$ network with m nodes, time-varying arrivals, staffing and abandonments, and inhomogeneous Poisson service and routing, and characterize the performance of the network as a direct extension of the single-server queue case.

The rest of the paper is organized as follows. We start with a description of the transitory generalized Jackson network model in Section 3, and we develop fluid and diffusion approximations to the network primitives. In Section 4, we develop functional strong law of large numbers approximations to the queueing equations, and identify the fluid model corresponding to the transitory network. We identify the diffusion network model in Section 5, and establish a weak convergence result for a tandem network with unimodal arrival time distribution. We end with conclusions and future research directions in Section 7

2. Notation

Following standard notation, $\mathcal{C}^K$ represents the space of continuous $\mathbb{R}^K$-valued functions, and $\mathcal{D}^K$ the space of functions that are right continuous with left limits and are $\mathbb{R}^K$-valued. The space $\mathcal{D}^K_{l,r}$ consists of $\mathbb{R}^K$-valued functions that are either right- or left-continuous at each point in time, while $\mathcal{D}^K_{\lim}$ is the space of $\mathbb{R}^K$-valued functions that have right and left limits at all points in time. The space and mode of convergence of a sequence of stochastic elements is represented by (X, Y), where X is the space in which the stochastic elements take values and Y the mode of convergence. In this paper our results will be proved under the uniform mode of convergence and occasionally in the "strong" M_1 (SM_1) topology (see (Whitt 2001b, Chapter 11)). Weak convergence of measures will be represented by $\Rightarrow$. Finally, $\mathrm{diag}(x_1, \ldots, x_K)$ represents a $K \times K$ diagonal matrix with entries $x_1, \ldots, x_K$.

3. Transitory Queueing Network

We consider a single class queueing network with K single server FIFO nodes. Each node starts service at some fixed time, which could be different from the other nodes. We assume every job is served independently of the others and that the servers are non-preemptive and non-idling. The network is assumed to offer Markovian routing between the nodes. Thus, the routing can be represented by a sub-stochastic *routing matrix*, $\mathbf{P}$. Furthermore, we assume that the network is open implying that all arriving users eventually depart the network. In this section we present (and prove, where necessary) functional strong law of large numbers and functional central limit theorem results for the network data; that is, the arrival process $\mathbf{A}$, the service process $\mathbf{S}$ and the routing process $\mathbf{R}$, in the limit of a large number of arrivals n by rescaling the service process appropriately by the population size. We call this the 'population acceleration' approximation regime, analogous to the 'uniform acceleration' regime used in Mandelbaum et al. (1998).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an appropriate probability space on which we define the requisite random elements. Let $\mathcal{K} := \{1, \ldots, K\}$ be the set of nodes in the network, and $\mathcal{E} \subset \mathcal{K}$ the set of nodes where exogeneous traffic enters the network. Each node in $\mathcal{E}$ receives n jobs that arrive exogeneously to the node. We assume a very general model of the traffic: let $\mathbf{T}_m := (T_{1,m}, \ldots, T_{J,m})$, $m \leq n$, represent the tuple of arrival epoch random variables of the mth job to each of the nodes (here $J := |\mathcal{E}|$). By assumption $T_{j,m} \in [0, T]$ for all $j \in \mathcal{E}$ and $1 \leq m \leq n$. We also assume that $\{\mathbf{T}_m; m = 1, \ldots, n\}$ forms a sequence of independent random vectors. Let F_j be the distribution function of the arrival epochs to node $j \in \mathcal{E}$; that is $\mathbb{E}[\mathbf{1}_{\{T_{j,m} \leq t\}}] = F_j(t)$ with support $[0, T]$. Users sample a time epoch to arrive at the node and enter the queue in order of the sampled arrival epochs; thus the arrival process to each node is a function of the ordered statistics of the arrival epoch random variables. In many situations, it is plausible that there is correlation between the arrival processes to the nodes in $\mathcal{E}$. To model such phenomena, we assume that the joint distribution of the arrival epochs all the nodes in the network are fully specified. To be precise, we assume that $\mathbb{P}(T_{1,m} \leq t, \ldots, T_{J,m} \leq t)$ for

all $m \in \{1, \ldots, n\}$ is well defined. This implies that the arrival epochs of the mth job to each node in the network can be correlated. Let $\mathbf{a}_m(t) := (\mathbf{1}_{\{T_{1,m} \leq t\}}, \ldots, \mathbf{1}_{\{T_{J,m} \leq t\}}) \in \mathcal{D}^J[0, \infty)$ and

$$A_j := \sum_{m=1}^{n} \mathbf{1}_{\{T_{j,m} \leq t\}} \text{ for } 1 \leq j \leq J, \tag{1}$$

then $\mathbf{A}_n(t) := \sum_{m=1}^{n} \mathbf{a}_m(t) = (A_1(t), \ldots, A_J(t)) \in \mathcal{D}^J[0, \infty)$ is the vector of cumulative arrival processes to the nodes in $\mathcal{E}$. Then, $\mathbb{E}[\mathbf{A}_n(t)] = \mathbf{F}_n(t) = n(F_1(t), \ldots, F_J(t))$ and $\mathbb{E}[\mathbf{A}_n(t)\mathbf{A}_n(t)^T] = [nF_{i,j}(t) + n(n-1)F_i(t)F_j(t)]$, where $F_{i,j}(t) := \mathbb{P}(T_{i,m} \leq t, T_{j,m} \leq t)$. This 'multi-variate empirical process' representation for the arrival process affords a very natural model of correlated traffic in networks, and stands in contrast with generalized Jackson networks where external traffic is assumed to be independent to nodes in $\mathcal{E}$.

Recall from Donsker's Theorem (for empirical sums) that $\sqrt{n}(A_i - F_i) \Rightarrow W_i^0 \circ F_i$, where W_i^0 is a standard Brownian bridge process. The Brownian bridge process is also well defined as a 'tied-down' Brownian motion process equal in distribution to $(W_i(t) - tW_i(1),\ t \in [0,1])$, for all $t \in [0,1]$, where W_i is a standard Brownian motion process. Recall that we also assume that the arrival epochs to the different nodes in the network can be correlated so that the random vector $\mathbf{T}_m$ has covariance matrix R.

Definition 1. Let $\mathbf{W} = (W_1, \ldots, W_J)$ be a J-dimensional standard Brownian motion process with identity covariance matrix. If R a $J \times J$ positive-definite matrix with lower-triangular Cholesky factor L, then $W_R = LW$ is a K-dimensional Brownian motion with covariance matrix R. By directly extending the definition of a one-dimensional Brownian bridge process,

$$\left(\mathbf{W}^0(t) = \mathbf{W}_R(t) - t\mathbf{W}_R(1),\ t \in [0,1]\right)$$

is a J-dimensional Brownian bridge process with covariance matrix R.

It is straightforward to see that $\mathbb{E}[\mathbf{W}^0(t)] = 0$ for all $t \in [0,1]$ and $\mathbb{E}[\mathbf{W}^0(t)\mathbf{W}^0(s)] = t(1-s)R$. For notational simplicity, we denote component-wise composition by $\mathbf{W}^0 \circ \mathbf{F} = (W_1^0 \circ F_1, \ldots, W_K^0 \circ F_K)$. Assume that the covariance function $R(t) = \mathbb{E}[(\mathbf{A}_n(t) - \mathbb{E}[A_n(t)])(\mathbf{A}_n(t) - \mathbb{E}[A_n(t)])^T] \in \mathcal{C}^{J \times J}$ is well defined. Then, Theorem 1 below establishes multivariate generalizations of the classical Glivenko-Cantelli and Donsker's theorems.

THEOREM 1. *Consider the triangular array of i.i.d. random vectors* $\{\mathbf{T}_m,\ m \le n\}\ n \ge 1$, *and let* $\mathbf{a}_m(t) := (\mathbf{1}_{\{T_{1,m} \le t\}}, \ldots, \mathbf{1}_{\{T_{J,m} \le t\}})$, *for* $t \in [0,\infty)$. *Then,*

(i) $\frac{1}{n}\sum_{m=1}^{n} \mathbf{a}_m \to \mathbf{F}$ *in* $(\mathcal{C}^J, U)$ *a.s. as* $n \to \infty$, *and*

(ii) $\hat{\mathbf{A}}_n := \sqrt{n}\left(\frac{1}{n}\sum_{m=1}^{n} \mathbf{a}_m - \mathbf{F}\right) \Rightarrow \mathbf{W}^0 \circ \mathbf{F}$ *in* $(\mathcal{C}^J, U)$ *as* $n \to \infty$, *where* $\mathbf{W}^0 \in \mathcal{C}^J[0,\infty)$ *is a* J*-dimensional Brownian bridge process from Definition 1.*

It is straightforward to show that the covariance function of $\mathbf{W}^0 \circ \mathbf{F}$ equals $R(t)$ as defined above. The proof of the theorem is available in the appendix.

Next, we consider a sequence of service processes indexed by the population size $n \ge 1$, $S_{k,n} : \Omega \times [0,\infty) \to \mathbb{N}$ for all $k \in \mathcal{K}$. We assume that for each $k \in \mathcal{K}$ the function $\mu_{k,n} : [0,\infty) \to [0,\infty)$ is Lebesgue-integrable and that $M_{k,n}(t) := \int_0^t \mu_{k,n}(s)ds$ satisfies $M_{k,n} \to M_k$ in (C, U) as $n \to \infty$, where $M_k : [0,\infty) \to [0,\infty)$ is non-decreasing and continuous. We also assume that

$$\mathbf{M}_n := (M_{1,n}, \ldots, M_{K,n}) \to \mathbf{M} := (M_1, \ldots, M_K) \text{ in } (C^K, U) \text{ as } n \to \infty. \tag{2}$$

Let $\mathbf{S}_n := (S_{1,n}, \ldots, S_{K,n})$ represent the 'network' service process, where the component processes are independent of each other and $S_{k,n}$ models the cumulative service process at node k. We assume that $\mathbf{S}_n$ satisfies the following fluid and diffusion approximations

ASSUMPTION 1. *The scaled service processes* $\{\mathbf{S}_n,\ n \ge 1\}$ *satisfies*

(i) $\left[\frac{\mathbf{S}_n}{n} - \mathbf{M}_n\right] \to 0$ *in* $(\mathcal{C}^K, U)$ *a.s. as* $n \to \infty$, *and*

(ii) $\hat{\mathbf{S}}_n(t) := \sqrt{n}\left(\frac{\mathbf{S}_n}{n} - \mathbf{M}\right) \Rightarrow \mathbf{W} \circ \mathbf{M}$ *in* $(\mathcal{C}^K, U)$ *as* $n \to \infty$, *where* $\mathbf{W} := (W_1, \ldots, W_K)$ *is a* $K-$*dimensional Brownian motion process with covariance matrix* $diag(-\mu_1 c_1^2, \ldots, -\mu_K c_K^2)$ *and* c_k^2 *is the squared coefficient of variation of the service times in the* k*th queue.*

Note that the service process proposed in Theorem 1 is analogous to the time-dependent 'general' traffic process G_t proposed in Liu and Whitt (2014). It's possible to anticipate a proof of this result when the centered service process $\mathbf{S}_n - \mathbf{M}_n$ is a martingale. This would be the case when $\mathbf{S}_n$ is a K-dimensional stochastic process where the marginal processes are nonhomogeneous Poisson

processes and $M_{k,n} = \mathbb{E}[S_{k,n}]$. We leave the development of a general result here to a separate paper, and instead assume that such a service process exists.

On completion of service at node i, a job will join node j with probability $p_{i,j} \geq 0$, $i,j \in \{1,\ldots,K\}$, or exit the network with probability $1-\sum_j p_{i,j}$. Thus, the routing matrix $P := [p_{i,j}]$ is sub-stochastic. Note that, we also allow feedback of jobs to the same node; i.e., $p_{i,i} \geq 0$. Let $\phi_l^i : \Omega \to \{1,\ldots,K\}$, $\forall k \in \{1,\ldots,K\}$ and $\forall l \in \mathbb{N}$, be a measurable function such that $\phi_l^i = j$ implies that the lth job at node i will be routed to node j and $\mathbb{E}[\mathbf{1}_{\{\phi_l^i=j\}}] = p_{i,j}$. Define the random vector $R_l(m) := \sum_{i=1}^m e_{\phi_i^l}$, where e_i is the ith K-dimensional unit vector and the kth component of $R_l(m)$, denoted $R_l^k(m)$, represents the number of departures from node l to node k out of m departures from that node. Then, $\mathbf{R}(m) := (R_1(m),\ldots,R_K(m))$ is a $K \times K$ matrix whose columns are the routing vectors from the nodes in the network.

PROPOSITION 1. *The stationary routing process* $\{R(m),\ m \geq 1\}$ *satisfies the following functional limits:*

(i) $\frac{1}{n}\mathbf{R}(ne) \to \mathbf{P}\ e$ *in* $(\mathcal{C}^{K\times K}, U)$ *a.s. as* $n \to \infty$, *where* $e : [0,\infty) \to [0,\infty)$ *is the identity function, and*

(ii) $\hat{\mathbf{R}}_n := \sqrt{n}\left(\frac{1}{n}\mathbf{R}(ne) - \mathbf{P}e\right) \Rightarrow \hat{\mathbf{R}}$, *in* $(\mathcal{C}^{K\times K}, U)$ *as* $n \to \infty$, *where* $\hat{\mathbf{R}} = [W_{i,j}]$ *and* $W_{i,j}$ *are independent Brownian motion processes with mean zero and diffusion coefficient* $p_{i,j}(1-p_{i,j})$.

The proof of this result is a straightforward application of Donsker's theorem, and we omit it. As a direct consequence of Proposition 1 we have the following corollary, which will prove useful in our analysis of the network state process in the next section.

COROLLARY 1. *The routing process* $\mathbf{R}$ *also satisfies the following fCLT:*

$$\hat{\mathbf{R}}_n^T\mathbf{1} \Rightarrow \hat{\mathbf{R}}^T\mathbf{1} = \tilde{\mathbf{W}} \text{ in } (C^K, U) \text{ as } n \to \infty, \tag{3}$$

where $\mathbf{1} = (1,\ldots,1)$ *is a* K*-dimensional vector of one's and* $\tilde{\mathbf{W}} = (\sum_{k=1}^K W_{1,k},\ldots,\sum_{k=1}^K W_{K,k})$ *is a* K*-dimensional Brownian motion with mean zero and covariance matrix*

$$\Sigma = diag\left(\sum_{k=1}^K p_{1,k}(1-p_{1,k}),\ldots,\sum_{k=1}^K p_{K,k}(1-p_{K,k})\right).$$

Finally, we claim the following joint convergence result that summarizes and generalizes the convergence results in the afore-mentioned theorems.

PROPOSITION 2. *Assume that for each $n \geq 1$, $\mathbf{A}_n$, $\mathbf{S}_n$ and $\mathbf{R}(n)$ are mutually independent. Then,*
(i) $\left(\frac{1}{n}\mathbf{A}_n, \frac{1}{n}\mathbf{S}_n, \frac{1}{n}\mathbf{R}(ne)\right) \to (\mathbf{F}, \mathbf{M}, \mathbf{P}'e)$ *in* $(\mathcal{C}^J \times \mathcal{C}^K \times \mathcal{C}^{K\times K}, U)$ *a.s. as* $n \to \infty$, *and*
(ii) $\left(\hat{\mathbf{A}}_n, \hat{\mathbf{S}}_n, \hat{\mathbf{R}}_n\right) \Rightarrow \left(\mathbf{W}^0 \circ \mathbf{F}, \mathbf{W} \circ \mathbf{M}, \hat{\mathbf{R}}\right)$ *in* $(\mathcal{C}^J \times \mathcal{C}^K \times \mathcal{C}^{K\times K}, U)$ *as* $n \to \infty$.

The joint convergence follows from the assumed independence of the pre-limit random variables, and is straightforward to establish under the uniform convergence criterion.

3.1. Network Parameters

Let $Q_k(t) = E_k(t) - D_k(t)$ be the queue length sample path at node k, where $E_k(t) := A_k(t) + \sum_{l=1}^K R_l^k(S_l(B_l(t)))$ is the total number of jobs arriving at node k in the interval $[0,t]$ and D_k is the cumulative departure process. We assume that the server is non-idling implying that $D_k(t) = S_k(B_k(t))$, where $B_k(t) := \int_0^t \mathbf{1}_{\{Q_k(s)>0\}} ds$ is the total busy time of the server. Therefore, the queue length process is

$$Q_k(t) := A_k(t) + \sum_{l=1}^K R_l^k(S_l(B_l(t))) - S_k(B_k(t)). \tag{4}$$

Let $V_k(m)$ be the cumulative service time requirement of m arrivals to the kth node. As defined earlier, let ν_i^k be the workload presented by the ith arrival to the kth node in the network. Then, by definition,

$$V_k(m) := \sum_{i=1}^m \nu_i^k.$$

The instantaneous workload measured in units of time, at node k at time t is a function of the cumulative service time requirement of all arrivals at the node k, including both arrivals from the external stream and from internal routing, and the amount of time the server has been busy up to the instant of interest. Thus, we have

$$Z_k(t) := V_k(A_k(t) + \sum_{l=1}^K R_l^k(S_l(B_l(t)))) - B_k(t).$$

4. Functional Strong Law of Large Numbers

Recall the queue length process sample path (of node $k \in \mathcal{K}$) defined in (4). The K-dimensional multivariate stochastic process $\mathbf{Q} := (Q_1, \ldots, Q_K)$ represents the network state. Our first result establishes a fluid limit approximation to a rescaled version of $\mathbf{Q}$ by establishing a functional strong law of large number result as the exogeneous arrival population size n scales to infinity. Consider the queue length in the kth node, Q_k. Rescaling by the population size n, the fluid-scaled queue length process at node k is

$$Q_{k,n}(t) = A_{k,n}(t) + \sum_{l=1}^{K} R_l^k(S_{l,n}(B_{l,n}(t))) - S_{k,n}(B_{k,n}(t)),$$

where $A_{k,n}$ is defined as in (1), $S_{k,n}$ satisfies Theorem 1 and $B_{k,n}(t) := \int_0^t \mathbf{1}_{\{Q_{k,n}(s)>0\}} ds$ is the scaled busy time process. Centering each term on the right hand side by the corresponding fluid limits (and subtracting those terms), and introducing $\int_0^t \mu_{k,n}(s)ds$, we obtain $n^{-1}Q_{k,n}(t)$

$$\begin{aligned}
&= \left(\frac{1}{n}A_{k,n}(t) - F_k(t)\right) + \left(\frac{1}{n}\sum_{l=1}^{K}\left[R_l^k(S_{l,n}(B_{l,n}(t))) - p_{l,k}S_{l,n}(B_{l,n}(t))\right]\right) \\
&\quad - \left(\frac{S_{k,n}(B_{k,n}(t))}{n} - \int_0^{B_{k,n}(t)} \mu_{k,n}(s)ds\right) \\
&\quad + \left(F_k(t) - \int_0^{B_{k,n}(t)} \mu_{k,n}(s)ds + \frac{1}{n}\sum_{l=1}^{K} p_{l,k}S_{l,n}(B_{l,n}(t))\right) \\
&= \left(\frac{1}{n}A_{k,n}(t) - F_k(t)\right) + \left(\frac{1}{n}\sum_{l=1}^{K}\left[R_l^k(S_{l,n}(B_{l,n}(t))) - p_{l,k}S_{l,n}(B_{l,n}(t))\right]\right) \\
&\quad - \left(\frac{S_{k,n}(B_{k,n}(t))}{n} - \int_0^{B_{k,n}(t)} \mu_{k,n}(s)ds\right) \\
&\quad + \left(F_k(t) - \int_0^{t} \mu_{k,n}(s)ds\right) + (1-p_{k,k})\int_{B_{k,n}(t)}^{t} \mu_{k,n}(s)ds \\
&\quad + \left(\frac{1}{n}\sum_{l=1}^{K} p_{l,k}\left[S_{l,n}(B_{l,n}(t)) - n\int_0^{B_{l,n}(t)} \mu_{l,n}(s)ds\right]\right) \\
&\quad + \sum_{l=1}^{K} p_{l,k}\left(\int_0^t \mu_{l,n}(s)ds\right) - \sum_{l\neq k} p_{l,k}\int_{B_{l,n}(t)}^{t} \mu_{l,n}(s)ds.
\end{aligned} \tag{5}$$

Note that we used the fact that $B_{n,k}(t) \leq t$ so that $\int_0^t \mu_{k,n}(s)ds = \int_0^{B_{k,n}(t)} \mu_{k,n}(s)ds + \int_{B_{k,n}(t)}^t \mu_{k,n}(s)ds$. Recall too that $I_{k,n}(t) := t - B_{k,n}(t) = \int_{T_{s,k}}^t \mathbf{1}_{\{Q_{k,n}(s)=0\}} ds$ is the idle time process,

which measures the amount of time in $[T_{s,k}, t]$ that the node is not serving jobs (i.e., the queue is empty). Now, $n^{-1}Q_{k,n}$ can be decomposed as the sum of two processes, $\bar{X}_{k,n}$ and $\bar{Y}_{k,n}$, where

$$\begin{aligned}\bar{X}_{k,n}(t) = &\left(\frac{1}{n}A_{k,n}(t) - F_k(t)\right) + \left(\frac{1}{n}\sum_{l=1}^{K}\left[R_l^k(nS_{l,n}(B_{l,n}(t))) - p_{l,k}S_{l,n}(B_{l,n}(t))\right]\right)\\ &- \left(\frac{S_{k,n}(B_{k,n}(t))}{n} - \int_0^{B_{k,n}(t)}\mu_{k,n}(s)ds\right)\\ &+ \left(F_k(t) - \left(\int_0^t \mu_{k,n}(s)ds\right)\mathbf{1}_{\{t\geq T_{s,k}\}}\right) + \sum_{l=1}^{K} p_{l,k}\left(\int_0^t \mu_{l,n}(s)ds\right)\mathbf{1}_{\{t\geq T_{s,l}\}}\\ &+ \left(\frac{1}{n}\sum_{l=1}^{K} p_{l,k}\left[S_{l,n}(B_{l,n}(t)) - n\int_0^{B_{l,n}(t)}\mu_{l,n}(s)ds\right]\right), \text{ and}\end{aligned} \tag{6}$$

$$\bar{Y}_{k,n}(t) = (1-p_{k,k})\int_{B_{k,n}(t)}^{t}\mu_{k,n}(s)ds - \sum_{l\neq k} p_{l,k}\int_{B_{l,n}(t)}^{t}\mu_{l,n}(s)ds. \tag{7}$$

While this expression appears formidable, the analysis is simplified significantly by the fact that $\bar{\mathbf{Q}}_n := n^{-1}(Q_{1,n},\ldots,Q_{K,n})$ and $\bar{\mathbf{Y}}_n := (\bar{Y}_{1,n},\ldots,\bar{Y}_{K,n})$ are solutions to the K-dimensional Skorokhod/oblique reflection problem. First, recall the definition of the oblique reflection problem.

THEOREM 2. *[Oblique Reflection Problem] Let* $\mathbf{R}$ *be a* $K\times K$ *M-matrix1. , also known as the reflection matrix. Then, for every* $x\in\mathcal{D}_0^K := \{x\in\mathcal{D}^K : x(0)\geq 0\}$*, there exists a unique tuple of functions* (y,z) *in* $\mathcal{D}^K\times\mathcal{D}^K$ *satisfying*

$$\begin{aligned} z &= x + \mathbf{R}y \geq 0,\\ dy &\geq 0 \text{ and } y(0) = 0,\\ z_j dy_j &= 0, \quad j = 1,\ldots,K.\end{aligned} \tag{8}$$

The process $(z,y) := (\Phi(x),\Psi(x))$ *is the so-called oblique reflection map, where* $\Phi(x) = x + \mathbf{R}\Psi(x)$.

Note that, in general, if $\mathbf{G}$ is a nonnegative $\mathbf{M}$-matrix then so is $\mathbf{R} = \mathbf{I} - \mathbf{G}$ (Lemma 7.1 of Chen and Yao (2001)). The following lemma shows that the queue length satisfies the Oblique Reflection Mapping.

LEMMA 1. *Consider* $\bar{\mathbf{X}}_n(t) = (\bar{X}_{1,n}(t),\ldots,\bar{X}_{K,n}(t))\in\mathcal{D}_0^K$*, where* $\bar{X}_{k,n}(t)$ $k\in\{1,\ldots,K\}$ *is defined in* (6)*,* $\bar{\mathbf{Q}}_n\in\mathcal{D}^K$ *and* $\bar{\mathbf{Y}}_n\in\mathcal{D}_0^K$*. Then,*

$$(\bar{\mathbf{Q}}_n, \bar{\mathbf{Y}}_n) = (\Phi(\bar{\mathbf{X}}_n), \Psi(\bar{\mathbf{X}}_n)).$$

Proof: First, by definition we have $\bar{\mathbf{Q}}_n = \bar{\mathbf{X}}_n + (I - \mathbf{P}^T)\bar{\mathbf{Y}}_n$. Note that $\mathbf{P}$ is a non-negative (sub-stochastic) matrix with spectral radius less than unity and, therefore, an M-matrix, implying that $\mathbf{I} - \mathbf{P}^T$ is also an M-matrix. Once again by definition $Q_{k,n}$ and $Y_{k,n}$ satisfy the conditions in (8) for all $k \in \mathcal{K}$. Thus, the conditions of Theorem 2 are satisfied and the lemma is proved. ∎

Next, we establish a functional strong law of large numbers result for (6), which will subsequently be used in Theorem 3 for the queue length approximation.

LEMMA 2. *The fluid-scaled netput process $\bar{\mathbf{X}}_n$ converges to a deterministic limit as $n \to \infty$:*

$$\bar{\mathbf{X}}_n(t) \to \bar{\mathbf{X}}(t) := (\bar{X}_1(t), \ldots, \bar{X}_K(t)) \text{ u.o.c. a.s.},$$

where,

$$\bar{X}_k(t) = F_k(t) - \int_0^t \mu_k(s)ds + \sum_{l=1}^{K} p_{l,k} \int_0^t \mu_l(s)ds. \tag{9}$$

Proof: The result follows by an application of part (i) of Proposition 2 to (6). Noting that $B_{k,n}(t) \leq t$, the random time change theorem (Theorem 5.5, Chen and Yao (2001)) and Theorem 1 together imply that,

$$\frac{1}{n} S_{k,n}(B_{k,n}(t)) - \int_0^{B_{k,n}(t)} \mu_{k,n}(s)ds \to 0 \text{ u.o.c. a.s. as } n \to \infty \; \forall t \in [0, \infty).$$

Similarly, applying the random time change theorem along with Corollary 1 and Theorem 1 we obtain

$$\frac{1}{n} \left(R_l^k(S_{k,n}(B_{k,n}(t))) - p_{l,k} S_{k,n}(B_{k,n}(t)) \right) \to 0 \text{ u.o.c. a.s. as } n \to \infty \; \forall t \in [0, \infty).$$

Applying these results to (6) it follows that $\bar{X}_{k,n}(t) \to \bar{X}_k(t)$ u.o.c. a.s. as $n \to \infty$. The joint convergence follows automatically from these results and Proposition 2. ∎

We can now establish the functional strong law of large numbers limit for the queue length process. The proof essentially follows from the continuity of the oblique reflection map $(\Phi(\cdot), \Psi(\cdot))$.

THEOREM 3. *Let* $\bar{\mathbf{X}}_n(t)$ *and* $\bar{\mathbf{X}}(t)$ *be as defined in* (6) *and* (9) *respectively. Then,* $(\bar{\mathbf{Q}}_n(t), \bar{\mathbf{Y}}_n(t))$ *satisfy Theorem 2 and, as* $n \to \infty$*,*

$$(\bar{\mathbf{Q}}_n(t), \bar{\mathbf{Y}}_n(t)) \to (\Phi(\bar{\mathbf{X}}(t)), \Psi(\bar{\mathbf{X}}(t))) \text{ u.o.c. a.s. } \forall t \in [0, \infty).$$

Proof: It follows by Lemma 1 that $(\bar{\mathbf{Q}}_n(t), \bar{\mathbf{Y}}_n(t))$ satisfy the oblique reflection mapping theorem. Therefore, $(\bar{\mathbf{Q}}_n(t), \bar{\mathbf{Y}}_n(t)) \equiv (\Phi(\bar{\mathbf{X}}_n(t)), \Psi(\bar{\mathbf{X}}_n(t)))$. Now, the reflection regulator map, $\Psi(\cdot)$, is Lipschitz continuous under the uniform metric (Theorem 7.2, Chen and Yao (2001)). By the Continuous Mapping Theorem and Lemma 2 it follows that, $(\Phi(\bar{\mathbf{X}}_n(t)), \Psi(\bar{\mathbf{X}}_n(t))) \to (\Phi(\bar{\mathbf{X}}(t)), \Psi(\bar{\mathbf{X}}(t)))$ *u.o.c. a.s.* as $n \to \infty$, $\forall t \in [0, \infty)$. ∎

Note that neither Theorem 2 nor Theorem 3 provide an explicit functional form for the reflection regulator $\Psi(\cdot)$. It can be shown (see (Chen and Yao 2001, Chapter 7)) that the regulator map is the unique fixed point, $y^* \in \mathcal{D}^K$, of the map $\pi(x, y)(t) := \sup_{0 \le s \le t}[-x(s) + \mathbf{G}y(s)]^+ \ \forall t \in [0, \infty)$, where $\mathbf{G}$ is an M-matrix. Note that the supremum in the definition of the regulator is applied to every dimension of $\bar{\mathbf{X}}$ simultaneously. Extracting a closed form expression for y^* is not straightforward, barring a few special cases. The following corollary shows that the reflection map and fluid limit of the queue length process for a parallel node queueing network is particularly simple and an obvious generalization of that of a single queue.

COROLLARY 2. *Consider a* K*-node parallel queueing network. The fluid limit to the queue length and cumulative idleness processes are* $(\bar{\mathbf{Q}}, \bar{\mathbf{Y}}) = (\Phi(\bar{\mathbf{X}}, \Psi(\bar{\mathbf{X}}))) \in \mathcal{D}^2$*, where* $\bar{X} = (X_1, \ldots, X_K)$*,* $\Psi(\bar{\mathbf{X}}(t)) = \sup_{0 \le s \le t}[-\bar{\mathbf{X}}(s)]^+$ *and* $\Phi(\bar{\mathbf{X}}) = \bar{\mathbf{X}} + \Psi(\bar{\mathbf{X}})$*.*

Proof: Note that for a parallel queueing network $\mathbf{P} = 0$. Therefore, the fixed point of the map $\pi(\cdot, \cdot)$ is simply $\sup_{0 \le s \le t}[-x(s)]^+$. It follows that the regulator map of the fluid scaled queue length process is $\Psi(\bar{\mathbf{X}}_n(t)) = \sup_{0 \le s \le t}[-\bar{\mathbf{X}}_n(s)]^+$. It follows by Theorem 3 that $\Psi(\bar{\mathbf{X}}_n(t)) \to \sup_{0 \le s \le t}[-\bar{\mathbf{X}}(s)]^+$ and $\Phi(\bar{\mathbf{X}}_n(t)) \to \bar{\mathbf{X}}(t) + \Psi(\bar{\mathbf{X}}(t))$ *u.o.c. a.s.* as $n \to \infty$. ∎

A slightly more complicated example would be a series queueing network. Corollary 3 establishes the fluid limit to the network state of a two queue tandem network, when a large, but finite,

number n of users arrive at queue 1 over a finite time horizon $[-T_0, T]$. This result can be rather straightforwardly extended to a network of more than two queues.

COROLLARY 3. *Consider a tandem queueing network where* $\mathbf{P} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, *and* $\mathbf{R} = \mathbf{I} - \mathbf{P}^T = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$. *Let* $\mathbf{F} = F_1$ *be the arrival epoch distribution with support* $[-T_0, T]$ *where* $T_0, T > 0$, *and assume that* μ_1 *and* μ_2 *are the fixed service rates. Then, the fluid limits to the queue length and cumulative idleness processes are* $(\bar{\mathbf{Q}}, \bar{\mathbf{Y}}) = (\Phi(\bar{\mathbf{X}}), \Psi(\bar{\mathbf{X}})) \in \mathcal{D}^2$, $\bar{\mathbf{X}} := (X_1, X_2) = ((F_1 - \mu_1 e), (\mu_1 - \mu_2)e)$, $\Psi(\bar{\mathbf{X}}) = (Y_1, Y_2)$ *with* $Y_1(t) = \sup_{0 \le s \le t}(-X_1(s))_+$ *and* $Y_2(t) = \sup_{0 \le s \le t}(-X_2(s) + Y_1(s))_+ = \sup_{0 \le s \le t}[-X_2(s) + \sup_{0 \le r \le s}(-X_1(r))_+]_+$, *and* $\Phi(\bar{\mathbf{X}})(t) = \bar{\mathbf{X}} + R\Psi(\bar{\mathbf{X}}) = (X_1 + Y_1,\ X_2 + Y_2 - Y_1)$.

The proof is straightforward by substitution and we omit it. Note that the queue length fluid limit to the downstream queue appears quite complicated: $\bar{Q}_2 = X_2 + Y_2 - Y_1$ where $Y_2(t) = \sup_{0 \le s \le t}(-X_2(s) + Y_1(s))_+$. By substituting in the expression for X_2 we have

$$\begin{aligned} \bar{Q}_2 &= (\mu_1 - \mu_2)e + F_1 - F_1 - Y_1 + Y_2 \\ &= (F_1 - \bar{Q}_1 - \mu_2 e) + Y_2. \end{aligned}$$

Note that $F_1 - \bar{Q}_1$ is just the cumulative fluid departure function from the upstream queue, which is precisely the input to the downstream queue. Next, we consider the fluid limit for the busy time proces when the service process is stationary; i.e., $\mu_k(t) = \mu_k$ for all $t \ge 0$ and $k \in \mathcal{K}$.

THEOREM 4. *Let* $\bar{\mathbf{B}}_n(t) = (B_{1,n}(t), \ldots, B_{K,n}(t))$. *Then, as* $n \to \infty$,

$$\bar{\mathbf{B}}_n(t) \to \underline{t} - \mathbf{M}\Psi(\bar{\mathbf{X}}(t)) \text{ u.o.c. a.s., } \forall t \in [0, \infty). \tag{10}$$

Here, $\underline{t} = (t\mathbf{1}_{\{t \ge T_{s,1}\}}, \ldots, t\mathbf{1}_{\{t \ge T_{s,K}\}})$ *and* $\mathbf{M} = diag(1/\mu_1, \ldots, 1/\mu_K)$.

Proof: By definition $\bar{\mathbf{B}}_n(t) = t\underline{1} - \bar{\mathbf{I}}_n(t)$, where $\bar{\mathbf{I}}_n(t) = (I_{1,n}(t), \ldots, I_{K,n}(t))'$. Recalling the definition of the process $\bar{\mathbf{Y}}_n(t)$ it is straightforward to see that $\bar{\mathbf{I}}_n(t) = (\mathbf{I} - \mathbf{P}')^{-1}\bar{\mathbf{Y}}_n(t)$ for all $t \ge 0$. Therefore, $\bar{B}_n(t) = \underline{t} - (\mathbf{I} - \mathbf{P}')^{-1}\bar{\mathbf{Y}}_n(t)$. Theorem 3 implies that, as $n \to \infty$, $\bar{\mathbf{B}}_n(t) \to$

$\underline{t} - \Psi(\bar{\mathbf{X}}(t))$ *u.o.c. a.s.,* $\forall t \in [0, \infty)$. ∎

The following corollary establishes the fluid busy time process for the parallel queue case. The proof follows that of Corollary 2 and we omit it.

COROLLARY 4. *Consider a K-node parallel queueing network. Then,*

$$\bar{\mathbf{B}}_n(t) \to \underline{t} - (\mathbf{I} - \mathbf{P}')^{-1} \sup_{0 \leq s \leq t} [-\bar{\mathbf{X}}(s)]^+ \text{ as } n \to \infty.$$

In the stationary case we considered here, the busyness time-scale is effectively fixed by the service rate through the matrix M. On the other hand, if the service processes are non-stationary this time-scale *itself* is time-varying. Thus, computing the busy time (or equivalently the idle time) process when the service process is non-stationary is complicated. Note that the function $\bar{\mathbf{Y}}$ represents the number of "blanks" or the amount of unused capacity in the network at each point in time, providing an indication of whether a particular queue in the network is busy or not.

Note that the population acceleration scale we use in the current analysis ensures that (in the limit) the amount of time each user spends in service is infinitesimally small, and when a queue is busy arriving jobs are almost surely going to face delays. This 'behavior' of the queue state under the population acceleration scaling is akin to the conventional heavy-traffic scaling introduced in Reiman (1984) for stationary single class queueing networks. The corresponding diffusion heavy-traffic scaling identifies the critical time-scale of the stationary queueing network. The population acceleration scaling differs from the conventional heavy-traffic scaling by the fact that the fluid limit process is non-linear in nature. This implies that queues in the network can enter idle and busy periods, and arriving jobs will only face delays in the latter time intervals. We should expect that the critical time-scale of the queue state in the diffusion scale should itself change depending on whether the queue is busy or idle, leading to a non-stationary diffusion approximation. Indeed, this is precisely what we discover in the next section.

5. Functional Central Limit Theorems

We now consider the second order refinement to the fluid limit by establishing a functional central limit theorem (FCLT) satisfied by the queue length state process. We show, in particular, that the FCLT is a reflected diffusion, where the diffusion process $\hat{\mathbf{X}}$ is a function of the multi-dimensional Brownian bridge process as defined in Definition 1. Unlike the heavy traffic limits for generalized Jackson networks (see (Chen and Yao 2001, Chapter 7) Reiman (1984)), the diffusion is *not* reflected through the oblique reflection map (see (Chen and Yao 2001, Definition 7.1)). As noted, the non-homogeneous traffic and non-stationary service processes induce a time-varying critical time-scale under the population acceleration scaling. Here, we show that this time-varying critical time-scale manifests as a time-varying reflection boundary in transitory queueing networks. To be precise, the reflection regulator for the queue length diffusion is the directional derivative of the Oblique Reflection of $\bar{\mathbf{X}}$ (from Lemma 2) in the direction of the diffusion limit $\hat{\mathbf{X}}$ to the netput process. A similar result was observed in the case of a single $\Delta_{(i)}/GI/1$ transitory queue in Honnappa et al. (2014). In that case, the directional derivative reflection map was explicitly characterized by appealing to the results in (Whitt 2001a, Chapter 9). On the other hand, the results in Mandelbaum and Ramanan (2010) characterize the directional derivative of the multidimensional oblique reflection map.

Recall that $\mathbf{R}$ is a $K \times K$ M-matrix and $\mathbf{P}^T = \mathbf{I} - \mathbf{R}$. Let $x \in \mathcal{C}_0$ then, under the hypothesis of Theorem 2, there exists a unique *oblique reflection map* $(z, y) := (\Phi(x), \Psi(x)) \in \mathcal{C} \times \mathcal{C}$ such that $z = x + \mathbf{R}y$, y_j is non-decreasing and y_j grows only when z_j is zero (for all $j = 1, \ldots, K$). The directional derivative of the oblique reflection of x in the direction of the process $\chi \in \mathcal{C}$ is defined as follows (see Mandelbaum and Ramanan (2010) as well):

Definition 2. Given $(x, \chi) \in \mathcal{C}_0 \times \mathcal{C}$ and M-matrix R, the directional derivative of the oblique reflection map $\Phi(x) = x + R\Psi(x)$ in the direction of χ is the pointwise limit of

$$\mathbf{\Delta}^n_\chi(x) := \sqrt{n}\left(\Phi\left(\frac{\chi}{\sqrt{n}} + x\right) - \Phi(x)\right) \in \mathcal{C} \quad n \geq 1$$

as $n \to \infty$.

Theorem 1.1 (ii) of Mandelbaum and Ramanan (2010) identifies the limit process, which we state as a lemma for completeness. Here, $\mathcal{D}_{usc}$ is the space of RCLL functions that are upper semi-continuous as well.

LEMMA 3. *If $(x,\chi) \in \mathcal{C}_0 \times \mathcal{C}$ then the directional derivative limit $\mathbf{\Delta}_\chi(x)$ exists and convergence in Definition 2 is uniformly on compact subsets of continuity points of the limit $\mathbf{\Delta}_\chi(x)$. Further, if (z,y) solve the oblique reflection problem for x then*

$$\mathbf{\Delta}_\chi(x) = \chi + R\gamma(x,\chi),$$

where $\gamma := \gamma(x,\chi)$ lies in $\mathcal{D}_{usc}$ and is the unique solution to the system of equations

$$\gamma^i(t) = \begin{cases} \sup_{s\in\nabla_t^i}[-\chi^i(s) + [P\gamma]^i(s)]_+ & t \in [0, t_u^i], \\ \sup_{s\in\nabla_t^i}[-\chi^i(s) + [P\gamma]^i(s)] & t > t_u^i, \end{cases}$$

for $i = 1,\ldots,K$, where $\nabla_t^i := \{s \in [0,t] | z^i(s) = 0$ and $y^i(s) = y^i(t)\}$, and $t_u^i := \inf\{t \geq 0 : y^i(t) > 0\}$.

Now, the second order refinement to the netput process is $\hat{\mathbf{X}}_n := \sqrt{n}\left(\bar{\mathbf{X}}_n - \bar{\mathbf{X}}\right) \in \mathcal{D}^K$. Using Proposition 2, and the fact that the limit processes have sample paths in $\mathcal{C}^K$, the following Lemma is straightforward to establish. We abuse notation slightly and denote composition of two vector-valued functions as $x \circ y = (x_1 \circ y_1, \ldots, x_K \circ y_K)$.

LEMMA 4. *The diffusion-scaled netput process satisfies,*

$$\hat{\mathbf{X}}_n \Rightarrow \hat{\mathbf{X}} \text{ in } (\mathcal{C}^K, U) \text{ as } n \to \infty,$$

where $\hat{X}_k := W_k^0 \circ F_k - W_k \circ \int_0^t \mu_k(s)ds + \left\langle \hat{\mathbf{R}}_k \circ \mathbf{M}, \mathbf{1} \right\rangle$, $\hat{\mathbf{R}}_k$ is the kth row of the matrix valued process $\hat{\mathbf{R}}$ defined in part (ii) of Proposition 1, $\mathbf{M}$ is defined in (2), and $\langle \cdot, \cdot \rangle$ is the inner product operator and $\mathbf{1}$ is the K-dimensional vectors of ones.

The proof of the lemma is a straightforward application of part (ii) of Proposition 2 and omitted for brevity.

Returning to the queue length process, the diffusion scale process is $\hat{\mathbf{Q}}_n := \sqrt{n}\left(\bar{\mathbf{Q}}_n - \bar{\mathbf{Q}}\right) \in \mathcal{D}^K$. Recall, from Lemma 1, that $\bar{\mathbf{Q}}_n = \bar{\mathbf{X}}_n + \mathbf{R}\Psi(\bar{\mathbf{X}}_n)$ and, from Theorem 3, that $\bar{\mathbf{Q}} = \bar{\mathbf{X}} + \mathbf{R}\Psi(\bar{\mathbf{X}})$. It follows that

$$\begin{aligned}
\hat{\mathbf{Q}}_n &= \sqrt{n}\left(\bar{\mathbf{X}}_n + \mathbf{R}\Psi(\bar{\mathbf{X}}_n) - \bar{\mathbf{X}} - \mathbf{R}\Psi(\bar{\mathbf{X}})\right) \\
&= \hat{\mathbf{X}}_n + \mathbf{R}\sqrt{n}\left(\Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right) - \Psi\left(\bar{\mathbf{X}}\right)\right) + \mathbf{R}\sqrt{n}\left(\Psi(\bar{\mathbf{X}}_n) - \Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right)\right) \\
&= \Delta^n_{\hat{\mathbf{X}}_n}\left(\bar{\mathbf{X}}\right) + \mathbf{R}\sqrt{n}\left(\Psi(\bar{\mathbf{X}}_n) - \Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right)\right).
\end{aligned}$$

Our next result shows that $\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}_n}(\bar{\mathbf{X}})$ is asymptotically equal to $\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$.

LEMMA 5. *Let $\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$ and $\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}_n}(\bar{\mathbf{X}})$ be defined as in Definition 2. Then,*

$$\left\|\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}_n}(\bar{\mathbf{X}}) - \boldsymbol{\Delta}^n_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})\right\| \to 0 \text{ a.s. as } n \to \infty,$$

where $\|\cdot\|$ is the supremum norm.

Proof: First, recall that $\boldsymbol{\Delta}^n_{\hat{\mathbf{X}}_n}(\bar{\mathbf{X}}) = \hat{\mathbf{X}}_n + \mathbf{R}\ \sqrt{n}\left(\Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right) - \Psi(\bar{\mathbf{X}})\right)$. By Lemma 4 and the Skorokhod representation theorem (Durrett 2010, Chapter 8), it follows that $\|\hat{\mathbf{X}}_n - \hat{\mathbf{X}}\| \to 0$ a.s. as $n \to \infty$. The lemma is proved once we show that $\|\sqrt{n}\left(\Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right) - \Psi\left(\frac{\hat{\mathbf{X}}}{\sqrt{n}} + \bar{\mathbf{X}}\right)\right)\| \to 0$ a.s. as $n \to \infty$.

Chen and Whitt Chen and Whitt (1993) show that the oblique reflection map and the reflection regulator are Lipschitz continuous with respect to the uniform metric topology. Therefore,

$$\begin{aligned}
\left\|\sqrt{n}\left(\Psi\left(\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}}\right) - \Psi\left(\frac{\hat{\mathbf{X}}}{\sqrt{n}} + \bar{\mathbf{X}}\right)\right)\right\| &\leq K\sqrt{n}\left\|\frac{\hat{\mathbf{X}}_n}{\sqrt{n}} + \bar{\mathbf{X}} - \frac{\hat{\mathbf{X}}}{\sqrt{n}} - \bar{\mathbf{X}}\right\|, \\
&= K\|\hat{\mathbf{X}}_n - \hat{\mathbf{X}}\|,
\end{aligned}$$

where K is the Lipshitz constant associated with the oblique reflection map. The proof follows from the argument above showing that $\|\hat{\mathbf{X}}_n - \hat{\mathbf{X}}\| \to 0$ a.s. as $n \to \infty$. ∎

Lemma 5 implies it suffices to consider

$$\hat{\mathbf{Q}}_n \equiv \Delta^n_{\hat{\mathbf{X}}}\left(\bar{\mathbf{X}}\right) + \mathbf{R}\sqrt{n}\left(\Psi(\bar{\mathbf{X}}_n) - \Psi\left(\frac{\hat{\mathbf{X}}}{\sqrt{n}} + \bar{\mathbf{X}}\right)\right) \tag{11}$$

(where by an abuse of notation we call this process $\hat{\mathbf{Q}}_n$ as well). Now, if we show that

$$\left\| \sqrt{n} \left(\Psi(\bar{\mathbf{X}}_n) - \Psi\left(\frac{\hat{\mathbf{X}}}{\sqrt{n}} + \bar{\mathbf{X}} \right) \right) \right\| \to 0$$

a.s. as $n \to \infty$, then Lemma 3 implies that $\hat{\mathbf{Q}}_n$ converges to the process $\mathbf{\Delta}_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$ pointwise in the large population limit. The following lemma establishes the required result in a general setting.

LEMMA 6. *Let $x_n, x \in \mathcal{D}^K$ be stochastic processes that satisfy $\|\sqrt{n}(x_n - x)\| \to \chi$ a.s. as $n \to \infty$. Then,*

$$\left\| \sqrt{n} \left(\Psi(x_n) - \Psi\left(\frac{\chi}{\sqrt{n}} + x \right) \right) \right\| \to 0 \text{ a.s. as } n \to \infty, \tag{12}$$

where $\chi \in \mathcal{C}^K$.

Proof: The condition on x_n, x implies that $x_n \overset{a.s.}{=} x + (\sqrt{n})^{-1}\chi + o(\sqrt{n})$. Therefore, it follows that

$$\begin{aligned} \left\| \sqrt{n} \left(\Psi(x_n) - \Psi\left(\frac{\chi}{\sqrt{n}} + x \right) \right) \right\| &\overset{a.s.}{=} \| \sqrt{n} \left(\Psi\left(\frac{\chi}{\sqrt{n}} + x + o(1) \right) - \Psi\left(\frac{\chi}{\sqrt{n}} + x \right) \right) \| \\ &\leq K\sqrt{n} \| o(1) \|, \end{aligned}$$

where the last inequality follows from the Lipshitz continuity of the oblique reflection map. The final conclusion follows from the fact that the indeterminate form on the right hand side converges to 0 as $n \to \infty$. ∎

We can now state and prove the main result of this section.

THEOREM 5. *Let $\hat{\mathbf{Q}}_n = \sqrt{n}(\bar{\mathbf{Q}}_n - \bar{\mathbf{Q}})$ be the diffusion-scaled network state process. Then, for any fixed $t \in [0, \infty)$, as $n \to \infty$*

$$\hat{\mathbf{Q}}_n(t) \Rightarrow \hat{\mathbf{Q}}(t) = \Delta_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})(t), \tag{13}$$

where $\Delta_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})(t) = \hat{\mathbf{X}}(t) + R\gamma(\bar{\mathbf{X}}, \hat{\mathbf{X}})(t)$.

Proof: First, using the Skorokhod representation theorem Billingsley (1968), it follows from Proposition 4 that there exist versions of the stochastic processes $\left\{ \hat{\mathbf{X}}_n \right\}$ and $\hat{\mathbf{X}}$, referred to using the same notation, such that

$$\left\| \hat{\mathbf{X}}_n - \hat{\mathbf{X}} \right\| \to 0 \text{ a.s. as } n \to \infty.$$

It follows that $\bar{\mathbf{X}}_n \overset{a.s.}{=} \bar{\mathbf{X}} + (\sqrt{n})^{-1}\hat{\mathbf{X}} + o(1)$. Lemma 6 implies that

$$\left\| \sqrt{n} \left(\Psi(\bar{\mathbf{X}}_n) - \Psi\left(\frac{\hat{\mathbf{X}}}{\sqrt{n}} + \bar{\mathbf{X}} \right) \right) \right\| \to 0$$

a.s. as $n \to \infty$. Next, using Lemma 5 and Lemma 3, it follows that $\|\hat{\mathbf{Q}}_n(t) - \Delta_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})(t)\| \to 0$ a.s. as $n \to \infty$ for any fixed $t \in [0, \infty)$, which in turn implies weak convergence of the stochastic processes thus proving the desired result. ∎

Remarks: We include a short summary of the relevant results in Mandelbaum and Ramanan (2010) that imply that process-level convergence might be near impossible to prove (in general) in a transitory queueing network. Lemma 2 in Honnappa et al. (2014) (an extension of Theorem 3.2 in Mandelbaum and Massey (1995)) proves the process-level diffusion limit result in the M_1 topology for a single queue. The fact that the limit process has right- or left-discontinuity points that are 'unmatched' by the pre-limit process necessitates that convergence be proved in the M_1 topology as opposed to the more natural J_1 topology. On the other hand, Mandelbaum and Ramanan (2010) show that it is not possible to prove a process-level convergence result even in the WM_1 topology ('weak' M_1 topology (see Whitt (2001b)), due to the fact that the multidimensional limit process can have discontinuity points that are *both right- and left-discontinuous*. For completeness, we state the relevant portion ofTheorem 1.2 of Mandelbaum and Ramanan (2010) that encapsulates the various necessary conditions for discontinuities in the sample paths of the directional derivative limit process, $\mathbf{\Delta}_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$. First, given (z, y) as the solution to the oblique reflection problem for $x \in \mathcal{C}_0$ define, for each $t \in [0, \infty)$,

$$\begin{aligned}
\mathcal{O}(t) &:= \{i \in \{1, \ldots, K\} : z^i(t) > 0\}, \\
\mathcal{U}(t) &:= \{i \in \{1, \ldots, K\} : z^i(t) = 0,\ \Delta y^i(t+) \neq 0,\ \Delta y^i(t-) \neq 0\}, \\
\mathcal{C}(t) &:= \{1, \ldots, K\} \backslash [\mathcal{O}(t) \cup \mathcal{U}(t)], \\
\mathcal{EO}(t) &:= \{i \in \mathcal{C}(t) : \exists \delta > 0 \text{ such that } z^i(s) > 0\ \forall s \in (t - \delta, t)\}, \\
\mathcal{SU}(t) &:= \{i \in \mathcal{C}(t) : \Delta z^i(t-) = 0,\ \Delta z^i(t+) \neq 0.\}
\end{aligned}$$

When $x = \bar{\mathbf{X}}$, $\mathcal{O}(t)$ is the set of nodes in the network that are *overloaded* at time t, $\mathcal{U}(t)$ is the set of underloaded nodes, $\mathcal{C}(t)$ the set of critically loaded nodes, $\mathcal{EO}(t)$ is the set of critically loaded queues that are at the end of overloading and $\mathcal{SU}(t)$ is the set of critically loaded nodes that are at the start of under-loading. Note that the definitions of overloading, under-loading and critical loading conform to the standard notions for $G/G/1$ queues, as noted in Honnappa et al. (2014). Next, we also require the notion of critical and sub-critical chains, as in Definition 1.5 of Mandelbaum and Ramanan (2010):

Definition 3 (Def. 1.5 Mandelbaum and Ramanan (2010)). Given a $K \times K$ routing matrix $\mathbf{P}$ and the oblique reflection map Ψ and $x \in \mathcal{C}^K$ so that $y = \Psi(x)$. Then a sequence $j_0, j_1, \ldots, j_m$ with $j_k \in \{1, \ldots, K\}$ for $k = 0, 1, \ldots, m$ that satisfies $P_{j_{k-1} j_k} > 0$ for $k = 0, 1, \ldots, m$ is said to be a chain. The chain is said to be a cycle if there exist distinct $k_1, k_2 \in \{0, \ldots, m\}$ such that $j_{k_1} = j_{k_2}$, the chain is said to precede i if $j_0 = i$ and is said to be empty at t if $y_{j_k}(t) = 0$ for every $k = 1, \ldots, m$. For $i = 1, \ldots, K$ and $t \in [0, \infty)$, we consider the following two types of chains:

1. An empty chain preceding i is said to be critical at time t if it is either cyclic or j_m is at the end of overloading at t.

2. An empty chain preceding i is said to be sub-critical at time t if it is either cyclic or j_m is at the start of overloading at t.

Theorem 1.2 of Mandelbaum and Ramanan (2010) gives necessary conditions so that, in general, the sample paths of the directional derivative can have both a right *and* left discontinuity at $t \in [0, \infty)$. Simply put, the structure of the routing matrix $\mathbf{P}$ determines whether we see such a point.

Proposition 3 **(Thm. 1.2 Mandelbaum and Ramanan (2010))**. *Under the conditions of Definition 3 and given a process $\chi \in \mathcal{C}^k$, if the directional derivative $\Delta_\chi(x)$ has both a right and a left discontinuity at $t \in [0, \infty)$ then one of the following conditions must hold at time t:*

a) i is at the end of overloading, and a sub-critical chain precedes i, in which case

$$\Delta_\chi(x)^i(t-) < \Delta_\chi(x)^i(t)^i = 0 < \Delta_\chi(x)^i(t+),$$

b) i is at the start of under-loading and a critical chain precedes i, in which case

$$\Delta_\chi(x)^i(t-) > \Delta_\chi(x)^i(t) > \Delta_\chi(x)^i(t+) = 0,$$

c) i is not underloaded and there exist both critical and sub-critical chains preceding i; if, in addition, i is overloaded then the discontinuity is a separated discontinuity of the form

$$\Delta_\chi(x)^i(t) < \min\{\Delta_\chi(x)^i(t-), \Delta_\chi(x)^i(t+)\}.$$

Note that the sample paths of $\mathbf{\Delta}_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$ lie in $\mathcal{D}_{\lim}$ and establishing M_1 convergence in this space is non-trivial. Recall that the standard description of M_1 convergence is through the graphs of the functions - which can be described via linear interpolations in $\mathcal{D}$ and $\mathcal{D}_{l,r}$. However, in $\mathcal{D}_{\lim}$ no such simple description exists (see Chapter 12 of Whitt (2001b) and Chapter 6, 8 of Whitt (2001a) for further details on these issues).

Given the inherent difficulty in establishing a general process-level result, we first focus on a two queue tandem network, where the arrival time distribution is uniform on the interval $[-T_0, T]$ and $T_0, T > 0$ where the difficulties will become apparent.

THEOREM 6. *Consider a tandem queueing network with* $\mathbf{P} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, *and* $\mathbf{R} = \mathbf{I} - \mathbf{P}^T = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$. *Assume that* $\mathbf{F} = F_1$ *is uniform over* $[-T_0, T]$, *and service rate at node 1 is* μ_1 *and at node 2* μ_2. *Then,* $\hat{\mathbf{Q}}_n \Rightarrow \hat{\mathbf{Q}} := \mathbf{\Delta}_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$ *in* $(\mathcal{D}_{l,r}, SM_1)$ *as* $n \to \infty$, *where* $\hat{\mathbf{X}} = (\hat{X}_1, \hat{X}_2)$ *with* $\hat{X}_1 = W_1^0 \circ F_1 - W_1 \circ M_k$, $\hat{X}_2 = W_1 \circ M_k - W_2 \circ M_2$ *and* $M_k(\cdot) = \int_0^\cdot \mu_k(s)ds$ *for* $k \in \{1, 2\}$, $\bar{\mathbf{X}} = ((F_1 - \mu_1 e), (\mu_1 - \mu_2)e)^T$ *and* $e : \mathbb{R} \to \mathbb{R}$ *is the identity map.*

Proof: Recall that $F(t) = \frac{t+T_0}{T+T_0}$ for all $t \in [-T_0, T]$. We consider three subcases and establish the weak convergence result for each of them separately.

(i) Let $\mu_1 < \mu_2$. Then,

$$\bar{Q}^1(t) = \begin{cases} (F(t) - \mu_1 t \mathbf{1}_{\{t \geq 0\}}) & \forall t \in [-T_0, \tau_1), \\ 0 & \forall t \in [\tau_1, \infty), \end{cases} \tag{14}$$

and $\bar{Q}^2(t) = 0\ \forall t \geq 0$, where $\tau_1 := \inf\{t > 0 | F(t) = \mu_1 t\}$. These follow as a consequence of Corollary 3, and noting that $\bar{\mathbf{X}} = (F(t) - \mu_1 e, (\mu_1 - \mu_2)e)$. Thus, we have

$$\nabla_t^1 := \begin{cases} \{-T_0\} & \forall t \in [0, \tau_1), \\ \{-T_0, \tau_1\} & t = \tau_1, \\ \{t\} & \forall t > \tau_1, \text{ and} \end{cases} \tag{15}$$

$$\nabla_t^2 := \{t\}\ \forall t \in [0, \infty). \tag{16}$$

Thus, node 1 is in $\mathcal{O}(t)$ for all $t \in [-T_0, \tau_1)$, $\mathcal{C}(t)$ for $t = \tau_1$ and in $\mathcal{U}(t)$ for $t > \tau_1$, and node 2 is in $\mathcal{U}(t)$ for all t.

The limit process $\hat{\mathbf{Q}}$ has a discontinuity only in the first component at $\hat{Q}^1(\tau_1) = \hat{X}^1(\tau_1) + \max\{0, -\hat{X}^1(\tau_1)\}$. Note that $\hat{Q}^1(\tau_1-) = \hat{X}^1(\tau_1)$ and $\hat{Q}^1(\tau_1+) = 0$, implying that $\hat{Q}^1$ has either a right or left discontinuity at τ_1. If $\hat{X}^1(\tau_1) \geq 0$ then $\hat{Q}^1(\tau_1) = \hat{X}^1(\tau_1) = \hat{Q}^1(\tau_1-) > \hat{Q}^1(\tau_1+) = 0$ and has a right discontinuity. Else, if $\hat{X}^1(\tau_1) < 0$ then $\hat{Q}^1(\tau_1) = 0 = \hat{Q}^1(\tau_1+) > \hat{Q}^1(\tau_1-)$ and has a left discontinuity. Thus, the limit process $\hat{\mathbf{Q}}$ has sample paths in $\mathcal{D}_{l,r}$. The proof of convergence for $\hat{\mathbf{Q}}_n = (\hat{Q}^{n,1}, \hat{Q}^{n,2})$ in this case is simple. First, Theorem 2 of Honnappa et al. (2014) shows that $\hat{Q}^{n,1} \Rightarrow \hat{Q}^1 := \hat{X}^1 + \sup_{s \in \nabla_\cdot}^1(-\hat{X}(s))$ in $(\mathcal{D}_{l,r}, M_1)$ as $n \to \infty$, and $\hat{Q}^{n,2} \Rightarrow 0$ in $(\mathcal{D}_{l,r}, M_1)$. Recall that $Disc(\hat{Q}^1)$ and $Disc(\hat{Q}^2)$ are the (respective) sets of discontinuity point, and it is obvious that $Disc(\hat{Q}^1) \cap Disc(\hat{Q}^2) = \phi$. Therefore, by Corollary 6.7.1 of Whitt (2001a), $\hat{Q}^{n,1} + \hat{Q}^{n,2} \Rightarrow \hat{Q}^1$ in $(\mathcal{D}_{l,r}(\mathbb{R}), M_1)$ as $n \to \infty$. Consequent to Theorem 6.7.2, it follows that $\hat{\mathbf{Q}}_n \Rightarrow \hat{\mathbf{Q}} := (\hat{Q}^1, 0)^T$ in $(\mathcal{D}_{l,r}, SM_1)$ as $n \to \infty$.

(ii) Let $\mu_1 > \mu_2$. Then, $\bar{Q}^1$ and ∇_t^1 follow (14) and (15) (resp.). $\bar{Q}^2$ on the other hand, is more complex now:

$$\bar{Q}^2(t) = \begin{cases} (\mu_1 - m_2)t & \forall t \in [0, \tau_1], \\ (F_1(t) - \mu_2 t) & \forall t \in [\tau_1, \tau_2], \\ 0 & \forall t > \tau_2, \end{cases}$$

where $\tau_2 := \inf\{t > \tau_1 : F_1(t) = \mu_2 t\}$ (note that $\tau_2 > \tau_2$ since $\mu_1 > \mu_2$). It follows that

$$\nabla_t^2 = \begin{cases} \{0\} & \forall t \in [0, \tau_2), \\ \{0, \tau_2\} & t = \tau_2, \\ \{t\} & \forall t > \tau_2. \end{cases}$$

It follows that node 2 is in $\mathcal{O}(t)$ for all $t \in [0, \tau_2)$, $\mathcal{C}(t)$ at $t = \tau_2$ and $\mathcal{U}(t)$ for $t > \tau_2$.

The diffusion limit $\hat{\mathbf{Q}} := (\hat{Q}^1, \hat{Q}^2)$ has discontinuities in both components. For node 1, if $\hat{X}^1(\tau_1) \geq 0$ then $\hat{Q}^1(\tau_1)$ has a right discontinuity, while $\hat{X}^1(\tau_1) < 0$ then $\hat{Q}^1(\tau_1)$ has a left discontinuity. Similarly, if $\hat{X}^2(\tau_2) \geq 0$ then $\hat{Q}^2(\tau_2)$ has a right discontinuity, and if $\hat{X}^2(\tau_2) < 0$ it has a left discontinuity. It follows that $\hat{\mathbf{Q}}$ has sample paths in $\mathcal{D}_{l,r}$. Furthermore, it is clear that $Disc(\hat{Q}^1) \cap Disc(\hat{Q}^2) = \phi$. Therefore, the weak convergence result follows by the same reasoning as in part (i).

(iii) Assume $\mu_1 = \mu_2$. Once again, $\hat{Q}^1$ and ∇_t^1 follow (14) and (15) (resp.). On the other hand, for node 2 $\hat{Q}^2 = 0$, but unlike case (i), the queue is empty but the server operates at full capacity till τ_1, and then enters underload. Thus,

$$\nabla_t^2 = \begin{cases} [0, t] & \forall t \in [0, \tau_1], \\ \{t\} & \forall t > \tau_1. \end{cases}$$

It is clear that node 2 switches from $\mathcal{C}(t)$ in $[0, \tau_1]$ to $\mathcal{U}(t)$ for $t > \tau_1$. Furthermore, at τ_1 itself, the node is in $\mathcal{SU}(t)$ (the regulator is flat to the left of τ_1 and increasing to the right).

The diffusion limit, once again, has discontinuities in both components. However, it is clear that $Disc(\hat{Q}^1) = Disc(\hat{Q}^2) = \{\tau_1\}$. For any $\mathcal{T} > -T_0$, it is straightforward to see that $(\hat{Q}^1(t) - \hat{Q}^1(t-))(\hat{Q}^2(t) - \hat{Q}^2(t-)) \geq 0$ for all $-T_0 \leq t \leq \mathcal{T}$: clearly, for any $t < \tau_1$, $\hat{Q}^i$, $i = 1, 2$ are both continuous. On the other hand, at τ_1, $\hat{Q}^1(\tau_1) \geq \hat{Q}^1(\tau_1-)$ and $\hat{Q}^2(\tau_1) = \hat{Q}^2(\tau_1-)$. Finally, for any $t > \tau_1$, $\hat{Q}^1(\tau_1) = \hat{Q}^1(\tau_1-)$ and $\hat{Q}^2(\tau_1) = \hat{Q}^2(\tau_1-)$. Now, by Theorem 6.7.3 of Whitt (2001a), it follows that $\hat{Q}^{n,1} + \hat{Q}^{n,2} \Rightarrow \hat{Q}^1 + \hat{Q}^2$ in $(\mathcal{D}_{l,r}(\mathbb{R}), M_1)$ as $n \to \infty$. Then, by Theorem 6.7.2 of Whitt (2001a), $\hat{\mathbf{Q}}_n \Rightarrow \hat{\mathbf{Q}}$ in $(D_{l,r}, SM_1)$ as $n \to \infty$. This concludes the proof. ∎

Theorem 6 shows that in the case of a tandem network, with uniform arrival time distribution, the weak convergence result can be established in the space $\mathcal{D}_{l,r}$ and in the SM_1 topology. In fact this result is true, if F_1 is unimodal such that node 1 is overloaded in the initial phase (i.e., in the interval $[-T_0, \tau_1)$, with $T_0 \geq 0$ now). We capture this fact in the following corollary. Without loss of generality, we will assume that $T_0 = 0$.

COROLLARY 5. *Let F_1 be a unimodal distribution function with finite support $[0,T]$, and consider a tandem queue as defined in Theorem 6. Then, $\hat{\mathbf{Q}}_n \Rightarrow \hat{\mathbf{Q}} := \mathbf{\Delta}_{\hat{\mathbf{X}}}(\bar{\mathbf{X}})$ in $(\mathcal{D}_{l,r}, SM_1)$ as $n \to \infty$, where*

$$\hat{\mathbf{X}} := \left(W_1^0 \circ F_1 - \sigma_1 \mu_1^{3/2} W_1, (\sigma_1 \mu_1^{3/2} W_1 - \sigma_2 \mu_2^{3/2} W_2)\right)^T,$$

$\bar{\mathbf{X}} = (F_1 - \mu_1 e, (\mu_1 - \mu_2)e)^T$ *and $e : \mathbb{R} \to \mathbb{R}$ is the identity map.*

The proof follows that of Theorem 6 and is omitted. Note that the compact support assumption is required, due to the fact that we prove weak convergence over compact intervals of time (see Section 7.2 of Honnappa et al. (2014) for a discussion on this point).

6. High-intensity Analysis of Tandem Networks

We illustrate the utility of the afore-developed approximations in bottleneck analysis of transitory tandem networks. Bottleneck detection in queueing networks has received significant interest in the literature over the years. Almost all of the analysis in the literature has focused on the characterization and detection of bottlenecks in stationary queueing networks. Of particular relevance to our results in this paper is the heavy-traffic bottleneck phenomenon Suresh and Whitt (1990), Whitt (2001b). To recall, the heavy-traffic bottleneck phenomenon corresponds to the state space collapse that is observed when the traffic intensity at a single queue approaches 1, while the traffic intensity at other queues remains below 1. In this case, the well known heavy-traffic approximations in Iglehart and Whitt (1970), Reiman (1984), Chen and Mandelbaum (1991c) indicate that the network workload process will collapse to a single dimension determined by the bottleneck node. In other words, the non-bottleneck nodes behave like switches where the service time is effectively zero.

In general, exact bottleneck analysis is very difficult (if not impossible), and there have been several approximations been proposed in the literature, particularly the parametric-decomposition approach Whitt (1983), Buzacott and Shanthikumar (1992), the stationary-interval method Whitt (1984), and Reiman's individual (IBD) and sequential bottleneck decomposition (SBD) algorithms Reiman (1990). Bottleneck analysis, however, has largely been ignored in non-stationary environments, and in transitory networks in particular. The key difference (and difficulty) in the transitory setting is that, for general arrival time distribution F, the bottleneck queue is time dependent. The situation is considerably simpler when F is uniform, and we focus on this case first to illustrate the main ideas.

Consider a series network of K queues. Let the service rate at queues 1 through $K-1$ be μ_1 and μ_K at queue K. Without loss of generality we assume that $\mu_K < 1 \leq \mu_1$. Assume that the traffic arrival epochs are randomly scattered per a uniform distribution function, over the interval $[0,1]$. Then, in the fluid population acceleration limit as observed in Theorem 3, it can be observed that each of the queues $1,\ldots,K-1$ behave like instantaneous switches and $O(n)$ fluid accumulates at the final queue. Extending the analysis in Corollary 3 to a K-node tandem network it is straightforward to compute that $\bar{\mathbf{X}} = (\bar{X}_1,\ldots,\bar{X}_{K-1},\bar{X}_K)$, where $\bar{X}_1(t) = F_1(t) - \mu_1 t = (1-\mu_1)t \leq 0$ and $\bar{X}_k = 0$ for all $k = 2,\ldots,K-1$, and $\bar{X}_K(t) = (\mu_1 - \mu_K)t > 0$. Since the routing matrix is

$$\mathbf{P} = \begin{pmatrix} 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ & & \vdots & & \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix}$$

a simple (if tedious) calculation shows that

$$\bar{\mathbf{Q}}(t) = \begin{cases} (0,\ldots,0,(1-\mu_K)t) & t \in [0,1/\mu_K], \\ (0,\ldots,0) & t > 1/\mu_K. \end{cases}$$

Now, the fluid workload process in this network can established as a corollary to Theorem 3:

COROLLARY 6 **(Workload Approximation)**. *Recall that* $\mathbf{M}$ *is a diagonal matrix defined as*

$$\mathbf{M} := diag(1/\mu_1, \ldots, 1/\mu_1, 1/\mu_K).$$

Then the fluid workload process $\bar{\mathbf{Z}} = \mathbf{M}\bar{\mathbf{Q}}$*, and the diffusion workload process is* $\hat{\mathbf{Z}} = \mathbf{M}\hat{\mathbf{Q}}$.

The proof of this corollary follows by analogous arguments to Proposition 4 in Honnappa et al. (2014). Straightforward algebra shows that

$$\bar{\mathbf{Z}}(t) = \begin{cases} (0, \ldots, 0, (\mu_K^{-1} - 1)t) & t \in [0, 1/\mu_K], \\ (0, \ldots, 0) & t > 1/\mu_K. \end{cases}$$

Thus, in the fluid limit, we find that the tandem queueing network "collapses" to a single queue in the fluid limit (this is an example of a *state space collapse* as defined in Reiman (1984)), and the sojourn time through the network, in the fluid scale and large population limit, is determined entirely by the delay at node K. The fluid analysis

On the other hand, as the diffusion limit in Theorem 6 shows, there is non-zero variability in the queue length at each node in the network. Indeed, Theorem 6 and Corollary 6 imply that the diffusion limit of the workload vector in a tandem network is $\hat{\mathbf{Z}} = \mathbf{M}\Psi(\hat{\mathbf{X}})$, where

$$\hat{\mathbf{X}}(t) = \left((W_1^0(t) - \sigma\mu_1^{3/2}W_1(t)), (\sigma_1\mu_1^{3/2}W_1(t) - \sigma_1\mu_1^{3/2}W_2(t), \ldots, \sigma_1\mu_1^{3/2}W_{K-1}(t) - \sigma_K\mu_K^{3/2}W_K(t))\right).$$

Now, if $\mu_1 > 1$, then $Z_k \stackrel{D}{=} 0$ for $k = 1, \ldots, K-1$ and $Z_K(t) \stackrel{D}{=} \mu_K^{-1}(\hat{X}_K(t) + \sup_{0 \le s \le t}(-X_K(s)))$ with $X_K = \sigma_1\mu_1^{3/2}W_{K-1} - \sigma_K\mu_K^{3/2}W_K$. That is, in the population acceleration scaling the distribution of the sojourn time through the network is asymptotically equal to the delay distribution in the last queue.

On the other hand, if $\mu_1 = 1$, then $Z_1 = \mu_1^{-1}(\hat{X}_1(t) + \sup_{0 \le s \le t}(-\hat{X}_1(s)))$ with $\hat{X}_1 = W_1^0 - \sigma\mu_1^{3/2}W_1$, $Z_k \stackrel{D}{=} 0$ for $k = 2, \ldots, K-1$ and

$$Z_K = \begin{cases} \mu_K^{-1}(\sigma_1\mu_1^{3/2}W_{K-1} - \sigma_K\mu_K^{3/2}W_K) & \forall t \in [0, 1] \\ \mu_K^{-1}(-\sigma_K\mu_K^{3/2}W_K) & \forall t \in (1, 1/\mu_K] \\ 0 & \forall t > 1/\mu_K. \end{cases}$$

This indicates that there are two bottlenecks at queues 1 and K. Thus, there is a state space collapse to a two-dimensional vector $\hat{\mathbf{Z}} = (Z_1, Z_K)$, and the sojourn time through the network is asymptotically equal in distribution to the sum of the delays in these two queues.

Now, suppose F_1 is not uniform, but unimodal with support on $[0,1]$ and consider the two queue tandem network aluded to in Corollary 5. The uni-modality of the arrival epoch distribution implies that up to time $\tau := \arg\max\{F'(t) : t \in [0,1]\}$ the distribution function is convex increasing, while after τ it is concave decreasing. As a consequence, the bottleneck behavior of the network is quite similar to the uniform arrival epoch distribution case above. For simplicity, we assume that the distribution function is symmetric around τ and that the service rates are the same in the two networks. The fluid netput process is $\bar{\mathbf{X}}(t) = (F_1(t) - \mu t, 0)$ and the fluid workload process, as a consequence of Corollary 6, is

$$\bar{\mathbf{Z}}(t) = \begin{cases} (0,0) & t \in [0, \tau_1] \\ \mathbf{M}\left(F_1(t) - F_1(\tau_1) - \mu(t-\tau_1), 0\right) & t \in (\tau_1, 1] \\ (0,0) & t > 1. \end{cases}$$

That is, the only bottleneck in the network is the first queue in the time horizon $(\tau_1, 1]$.

7. Concluding Statements

In this paper we developed asymptotic 'population acceleration' approximations of the queue length and (implicitly) the workload processes in a network of transitory queues. These results complement and add to the body of research studying single class generalized Jackson networks. In particular, our fluid limit results accomodate rather general traffic and service models. On the other hand, we can only establish point-wise diffusion approximations in the most general case, owing to the difficulties in the existence of the so-called directional derivative oblique reflection map. Nonetheless, we establish functional central limit theorems in the special case of a tandem network and we also present direct consequences of these developments on bottleneck analysis.

There are several directions in which this research will be expanded in the future. The extension of these results to general polling queueing networks will be interesting, exploiting some recently

observed connections between acceleration scalings and polling networks in Kavitha (????). Second, the arrival counts in non-overlapping intervals under the $\Delta_{(i)}$ traffic model have strong negative association. How soon will this correlation be 'forgotten' as traffic passes through multiple stages of service? This requires a study of the possible sample paths of the workload process. We believe this question has deep connections with directed percolation models; this is not a novel observation: Glynn and Whitt Glynn and Whitt (1991) identify this connection when there are no traffic dynamics. In on-going work we are working towards extending their analysis to transitory networks. A further interesting question is how the last passage percolation time scales with the population size in a non-stationary setting (as opposed to the classical setting where the percolation model is only studied in the stationary setting). The connection between percolation time and the sojourn time through the network affords yet another bottleneck/performance analysis measure in networks of transitory queues that will be highly relevant in the context of manufacturing lines. We will consider these questions in future papers.

8. Proofs of Theorems

8.1. Proof of Theorem 1

The following lemma establishes a fluid to the arrival process $\mathbf{A}_n$.

LEMMA 7. *The multivariate traffic process* $\mathbf{A}_n = (A_1, \ldots, A_J) := \sum_{m=1}^{n} \mathbf{a}_m$ *satisfies a functional strong law of large numbers where*

$$n^{-1}\mathbf{A}_n \to \mathbf{F} \text{ in } (C^J, U) \text{ a.s.}$$

as $n \to \infty$, *where* $\mathbf{F} = (F_1, \ldots, F_J)$ *and* $F_j(t) = \mathbb{E}[\mathbf{1}_{\{T_j \leq t\}}]$ *for all* $t \in [0, T]$.

Proof: First, for each $j \in \mathcal{E}$, the classical Glivenko-Cantelli theorem implies that

$$n^{-1}A_j \to F_j \text{ in } (\mathcal{C}, U) \text{ a.s.} \tag{17}$$

as $n \to \infty$. By the multivariate strong law of large numbers it is straightforward to argue that for a fixed $t \in [0, T]$

$$\mathbf{A}_n(t) \to \mathbf{F}(t) \text{ a.s.} \tag{18}$$

as $n \to \infty$. The functional limit follows as a consequence of (17). ∎

This proves part (i) of Theorem 1. The next lemma establishes part (ii).

LEMMA 8. *The multivariate traffic process* $\mathbf{A}_n$ *satisfies a functional central limit theorem where*

$$\sqrt{n}\left(n^{-1}\mathbf{A}_n - \mathbf{F}\right) \Rightarrow W^0 \circ \mathbf{F} \text{ in } (\mathcal{C}^J, U),$$

where $\mathbf{W}^0 \circ \mathbf{F}$ *is a* J*-dimensional Brownian bridge process as defined in Definition 1, with covariance function* $(R(t),\ t \geq 0) = ([F_{i,j}(t) - F_i(t)F_j(t)],\ t \geq 0)$.

Proof: Once again, Donsker's theorem for empirical processes implies that

$$\hat{A}_j := \sqrt{n}\left(n^{-1}A_j - F_j\right) \Rightarrow W_j^0 \circ F_j \text{ in } (\mathcal{C}, U) \tag{19}$$

as $n \to \infty$ for every $j \in \mathcal{K}$. This implies that the marginal arrival processes are tight. (Whitt 2001b, Theorem 11.6.7) implies that the multivariate process $\mathbf{A}_n$ is also tight.The multivariate central limit theorem (Whitt 2001b, Theorem 4.3.4) implies that the scaled process $\hat{\mathbf{A}}_n(t) = (\hat{A}_1(t), \ldots, \hat{A}_J(t))$ (for fixed $t \in [0,T]$) satisfies

$$\hat{\mathbf{A}}_n(t) = \sqrt{n}\left(\frac{\mathbf{A}_n(t)}{n} - \mathbf{F}(t)\right) \Rightarrow \mathcal{N}(0, R(t)),$$

where $\mathcal{N}(0,R(t))$ is a mean zero J-dimensional Gaussian random vector with covariance matrix $R(t) = [F_{i,j}(t) - F_i(t)F_j(t)]$. The Cramér-Wold device together with this result implies that the finite-dimensional distributions of $\mathbf{A}_n$ converge weakly to a tuple of Gaussian random vectors. The tightness of the processes $\{\mathbf{A}_n\}$, the continuity of the limit process and Prokhorov's theorem implies that $\hat{\mathbf{A}}_n$ converges weakly to the multivariate Gaussian stochastic process $\mathbf{W}^0 \circ \mathbf{F}$ with mean zero and covariance function $(R(t), t \geq 0)$ in $(\mathcal{C}^J, U)$. ∎

References

Billingsley, P. 1968. *Convergence of Probability Measures*. Wiley & Sons.

Buzacott, John A, J George Shanthikumar. 1992. Design of manufacturing systems using queueing models. *Queueing Systems* **12**(1-2) 135–213.

Chen, H., A. Mandelbaum. 1991a. Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.* **16**(2) 408–446.

Chen, H., A. Mandelbaum. 1991b. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann.Probab.* 1463–1519.

Chen, H., W. Whitt. 1993. Diffusion approximations for open queueing networks with service interruptions. *Queueing Syst.* **13**(4) 335–359.

Chen, H., D.D. Yao. 2001. *Fundamentals of Queueing Networks: Performance, asymptotics, and optimization*. Springer.

Chen, Hong, Avi Mandelbaum. 1991c. Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *The Annals of Probability* 1463–1519.

Duffield, N.G., W.A. Massey, W. Whitt. 2001. A nonstationary offered-load model for packet networks. *Telecomm. Syst.* **16**(3-4) 271–296. doi:10.1023/A:1016654625257. URL `http://dx.doi.org/10.1023/A%3A1016654625257`.

Durrett, R. 2010. *Probability: Theory and Examples*. 4th ed. Cambridge University Press.

Glynn, Peter W, Ward Whitt. 1991. Departures from many queues in series. *The Annals of Applied Probability* 546–572.

Harrison, J. M., M. I. Reiman. 1981. Reflected brownian motion on an orthant. *Ann. Probab.* 302–308.

Honnappa, H., R. Jain, A.R. Ward. 2013. On Transitory Queueing. Submitted.

Honnappa, H., R. Jain, A.R. Ward. 2014. A Queueing Model with Independent Arrivals, and its Fluid and Diffusion Limits. *Queueing Syst.* .

Iglehart, Donald L, Ward Whitt. 1970. Multiple channel queues in heavy traffic. ii: Sequences, networks, and batches. *Adv. Appl. Probab.* **2**(2) 355–369.

Jackson, J. R. 1957. Networks of waiting lines. *Oper. Res.* **5**(4) 518–521.

Kavitha, Veeraruna. ???? Personal Communication.

Keller, J. B. 1982. Time-dependent queues. *SIAM Rev.* 401–412.

Liu, Y., W. Whitt. 2011. A network of time-varying many-server fluid queues with customer abandonment. *Operations research* **59**(4) 835–846.

Liu, Y., W. Whitt. 2014. Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* **24**(1) 378–421.

Mandelbaum, A., W.A. Massey. 1995. Strong approximations for time-dependent queues. *Math. Oper. Res.* **20**(1).

Mandelbaum, A., K. Ramanan. 2010. Directional derivatives of oblique reflection maps. *Math. Oper. Res.* **35**(3) 527.

Mandelbaum, Avi, William A Massey, Martin I Reiman. 1998. Strong approximations for markovian service networks. *Queueing Syst.* **30**(1-2) 149–201.

Massey, W.A. 1985. Asymptotic analysis of the time dependent M/M/1 queue. *Math. Oper. Res.* 305–327.

Reiman, M. I. 1984. Open queueing networks in heavy traffic. *Math. Oper. Res.* **9**(3) 441–458.

Reiman, Martin I. 1990. Asymptotically exact decomposition approximations for open queueing networks. *Operations research letters* **9**(6) 363–370.

Suresh, S, W Whitt. 1990. The heavy-traffic bottleneck phenomenon in open queueing networks. *Operations Research Letters* **9**(6) 355–362.

Whitt, W. 2001a. *Internet Supplement To Stochastic Process Limits*. Springer.

Whitt, W. 2001b. *Stochastic Process Limits*. Springer.

Whitt, Ward. 1983. The queueing network analyzer. *Bell System Technical Journal* **62**(9) 2779–2815.

Whitt, Ward. 1984. Approximations for departure processes and queues in series. *Naval Research Logistics Quarterly* **31**(4) 499–521.