A Jackson Network under general regime

Yair Y. Shaki

Department of Industrial Engineering and Management,
The Jerusalem College of Technology,
Jerusalem, Israel.

May 9, 2018

Abstract

We consider a Jackson network in a general heavy traffic diffusion regime with the *α-parametrization*. We also assume that each customer may abandon the system while waiting. We show that in this regime the queue-length process converges to a multi-dimensional regulated Ornstein-Uhlenbeck process.

AMS subject classifications: 60K25, 60J60, 60F17, 90B22, 68M20.

Keywords: Jackson network; diffusion limits; many-server queue; heavy traffic; conventional diffusion regime; Halfin-Whitt regime.

BJPS - Accepted Manuscript

1 Introduction

The literature on diffusion approximation of queueing systems in heavy traffic focuses on two regimes. In the first one, named the *conventional regime* (see Chen and Yao [7]), the arrival and the service rates grow in the same proportion, while the number of the servers does not change. In contrast, the regime introduced by Halfin and Whitt (in short the *HW regime*) considers systems with a large number of servers, while the individual service rates do not change.([11])

Whitt [21, Theorem 2.2] and Mandelbaum [16] considered an M/M/N queue with an arrival rate that scales as N and with all the individual service rates being equal, while maintaining

a critically loaded system. They showed that the scaled queue-length process,

$$N^{-1}Q_N(Nt), \tag{1}$$

converges to a reflected Brownian motion (RBM) as $N \to \infty$. Gurvich [10, Proposition 5.1.1] extended that result to a more general scaled queue-length process.

Atar [1] argues that these limit theorems correspond to a diffusion regime that is different from the conventional and the HW regimes. In particular, he generalized these results to a system with multiple, heterogeneous servers working in parallel and obtained limit processes that do not appear in the other two regimes.

Atar [1] introduces the *α-parametrization*. To describe it, he discussed a single-class queueing model M/M/N with parameters λ^n, μ^n and N^n depending on some scaling parameter n. Let the external arrival rate increase as $O(n)$. Given some $\alpha \in [0, 1]$, assume that the number of servers is $O(n^\alpha)$, while each of the individual service rates scales as $n^{1-\alpha}$. In addition, let a suitable critical load condition hold. Then the extremal cases $\alpha = 0$ and $\alpha = 1$ correspond to the conventional and HW regimes, respectively. Atar [1] considered a wide range of work-conserving policies. The model also allowed for abandonment of customers waiting to be served.

The importance of the *α-parametrization* (see [1]) is surprisingly supported by a new study on optimal service capacity allocation in a loss system [13]. The research considers a loss system with a fixed budget for servers. The system owner decides the optimal number of the servers in order to maximize his profits. In the heavy traffic case, the optimal regime differs from the two common regimes (i.e. the HW regime [11] and the conventional regime) and belongs to an *α-parametrization* with $\alpha = 2/3$.

In both of the common heavy traffic regimes ($\alpha = 0$ and $\alpha = 1$) there have been studies on control problems. Atar and Solomon [2] studied the control problem in an NDS regime ($\alpha = 1/2$) for the first time and considered minimizing the 'running cost' random variables under policies that allow for service interruption. They construct a sequence of policies that asymptotically achieve this goal.

In recent years, much research has been devoted to considering the Jackson network in two

BJPS - Accepted Manuscript

heavy traffic regimes. Reiman ([19]) analyzed the Jackson network in the conventional heavy traffic regime, see also Chen and Yao ([7]). In this case, the limit of the queue-length process is a regulated multidimensional Brownian motion. A version for a model with abandonment is presented by Reed and Ward [18], where the limiting process is shown to be a regulated multidimensional Ornstein-Uhlenbeck process. Mandelbaum, Massey and Reiman [15] proved various results on many-server limits, including extensions about approximations for Markovian service networks in the framework of HW regime. Their work covers the Jackson network case as well.

The current paper applies for the first time a Jackson network under the α-parametrization with any $\alpha \in [0, 1)$. To describe it, we discussed a Jackson network with J interconnected queues where the external arrival process A_i to station i is a renewal process with rate λ_i^n. Let the external arrival rate λ_i^n increase as $O(n)$. Given some $\alpha \in [0, 1)$, assume that the number of servers N_i^n in the station i is $O(n^\alpha)$, while each of the individual service rates scales as $n^{1-\alpha}$. In addition, let a suitable critical load condition hold. Our analysis discusses a general class of work-conserving policies and allows also for customer abandonment. We prove that the multi-dimensional queue-length process converges to a regulated Ornstein-Uhlenbeck process by using a general multidimensional Skorohod equation [8].

This paper is organized as follows. Section 2 contains the setting and notations as well as the result concerning the limiting queue-length process in a general regime for any parameter $\alpha \in [0, 1)$. Section 3 presents the proof of this result. Section 4 provides some directions for future research.

2 Main Result

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, i.e., a probability space such that every subset of a set of probability 0 is measurable. This probability space will support all random variables and stochastic processes defined below. For each positive integer d, let $\mathcal{D}^d$ be the space of all right-continuous functions with left limits (RCLL) $x : [0, \infty) \to \mathbb{R}^d$. We endow $\mathcal{D}^d$ with the usual Skorohod J_1-topology ([4]). Let $\Rightarrow$ denote convergence in distribution.

BJPS -Accepted Manuscript

2.1 Multi-dimensional Queue-length Process

We consider a Jackson network with J interconnected queues where the external arrival process A_i to station i is a renewal process with rate λ_i. We assume that λ_i is strictly positive for at least one $1 \leq i \leq J$. Station i has several servers, and each arrival to station i is routed to a particular server. The service discipline at all queues is the first-in-first-out rule (FIFO). We suppose a work-conserving routing policy, so that no server may be idle when at least one customer is in its buffer. A customer completing service at station i will either move to some new queue j with probability p_{ij} or leave the system with probability $1 - \sum_{j=1}^{J} p_{ij}$, which is non-zero for some subset of the queues. We denote by $R^i(m)$ the vector with components $R^i_j(m)$ that denote the number of the first m jobs completing service at station i that are routed to station j. The routing matrix $P = (p_{ij})$ thus has spectral radius less than one so that each customer leaves the network after serving a finite number of stations. In other words, this network is open.

There are N servers, arranged in J stations, so that the number of servers in station i is N_i, for $1 \leq i \leq J$. The servers are labeled $1, \ldots, N$, and the set of k's for which server k is in station i is denoted by K_i. We write $K = \{1, \ldots, N\} = \cup_i K_i$, and $|K_i| = N_i$, $1 \leq i \leq J$.

For $1 \leq q \leq N$ and $t \geq 0$, let $I_{(q)}(t)$ take the value 1 if server q is idle at time t, and let it be 0 otherwise.

Denote by $I_i(t)$ the number of idle servers from station i at time t, for $1 \leq i \leq J$ (see (2)). Then I_i is stochastic process taking values in $[0, N_i]$, and

$$I_i = \sum_{q \in K_i} I_{(q)}, \qquad 1 \leq i \leq J. \tag{2}$$

The modeling of service completions will require usage of standard Poisson processes $S_{(q)}$ with rate $\mu_{(q)}$ per server, $1 \leq q \leq N$. The number of service completions by server q until time t is $S_{(q)}(T_{(q)}(t))$, where $T_{(q)}$ is defined as

$$T_{(q)}(t) = \mu_{(q)} \int_0^t (1 - I_{(q)}(s))ds, \qquad 1 \leq i \leq J, t \geq 0. \tag{3}$$

We consider a sequence of systems, with counter n, where the number of servers in the nth system is N_i^n. We assume that $N_i^n \to \infty$ as $n \to \infty$, and that $\inf_n N_i^n \geq 1$. The counter n of

BJPS - Accepted Manuscript

almost all processes and system parameters is denoted as a superscript. However, we do not need to count the standard Poisson processes as they do not depend on the system number, and they will be denoted by $S_{(q)}$.

Recall that the process of arrivals to a queue j is modeled as a renewal process. Let $\lambda_j^n \geq 0$, $n \in \mathbb{N}$, be parameters and consider sequences of positive i.i.d. random variables $\{IA_j(l), l \in \mathbb{N}\}$ ('inter-arrival times'), with mean $\mathbb{E}[IA_j(1)] = 1$ and variance $a_j = \mathrm{Var}(IA_j(1)) \in [0, \infty)$. For $\lambda_j^n > 0$, the number of arrivals up to time t for the nth system is given by

$$A_j^n(t) = \sup\Big\{l \geq 0 : \sum_{l=1}^{l} IA_j(l) \leq \lambda_j^n t\Big\}, \quad t \geq 0.$$

If $\lambda_j^n \equiv 0$ then $A_j^n(t) = 0$ for all $t > 0$.

The arrival rate λ_i^n is assumed to satisfy

$$\lambda_i^n / n \to \bar{\lambda}_i \in (0, \infty). \tag{4}$$

The parameters are assumed to satisfy

$$\min_{q \in K^n} \mu_{(q)}^n \to \infty. \tag{5}$$

In addition, it is assumed that the following limits exist:

$$\bar{\mu}_j^n := \frac{1}{n}\mu_j^n = \frac{1}{n} \sum_{q \in K_j^n} \mu_{(q)}^n \to \bar{\mu}_j > 0. \tag{6}$$

Recall that $N_i^n \to \infty$ as $n \to \infty$. From equations (4)-(6), there exists $\alpha \in [0, 1)$ such that the number of servers is $O(n^\alpha)$, while each of the individual service rates scales as $n^{1-\alpha}$. Hence, our model is under the α-parametrization.

The heavy traffic assumption makes the system critically loaded by relating the arrival and service rates as

$$\frac{1}{\sqrt{n}}[\lambda^n - (I - P^T)\mu^n] \to \beta \tag{7}$$

where $\lambda^n, \mu^n, \beta \in \mathbb{R}^J$ and P^T is the transpose matrix of the routing matrix.

We include the concept of customer impatience in the Jackson network. The abandonment rate per unit time, per customer waiting in the queue, is given by constants $\gamma_j^n \geq 0$, assumed

BJPS-Accepted Manuscript

to satisfy

$$\gamma_j^n \to \gamma_j \in [0, \infty), \quad 1 \leq j \leq J. \tag{8}$$

Let $\widetilde{S}_j$ be standard Poisson processes and let $Q_j^n(s)$ represents the number of customers in the queue j. The number of customers abandoning a queue j, until time t, will be given by

$$\widetilde{S}_j^n(t) := \widetilde{S}_j(\widetilde{T}_j^n(t)), \tag{9}$$

where

$$\widetilde{T}_j^n(t) = \gamma_j^n \int_0^t (Q_j^n - N_j^n)^+(s)ds, \tag{10}$$

and $\phi^+(s) = \max(\phi(s), 0)$. Note that if $\gamma_j^n = 0$ for all n, then there is no possibility of abandonment in the model.

The processes A_j^n, $S_{(q)}$, R^j, $I_{(q)}^n$ and Q_j^n are all assumed to have RCLL sample paths. Furthermore, we assume that the primitive processes A_j^n, $S_{(q)}$, R^j and $\widetilde{S}_j$ and the initial values

$$(\{I_{(q)}(0), q \in K\}, \{Q_j(0), 1 \leq j \leq J\}),$$

are mutually independent.

The following equation expresses the above verbal description:

$$Q_j^n(t) = Q_j^n(0) + A_j^n(t) + \sum_{k=1}^{J} R_j^k(S_k(T_k^n(t))) - S_j(T_j^n(t)) - \widetilde{S}_j^n(\widetilde{T}_j^n(t)).$$

2.2 Convergence of Queue-length Process

We suppose that the routing policies do not use information from the future. For this, we assume identical assumptions to those of Atar[1].

Assumption 2.1 *For each n there exists a filtration $\mathbb{F}_n = \{\mathcal{F}_n(t), t \geq 0\}$ that is right-continuous and $\mathbb{P}$-complete, such that the following holds:*

1. *The processes $A_i^n, Q_i^n, I_{(q)}^n, T_{(q)}^n, S_{(q)}^n(T_{(q)}^n), \widetilde{S}_j^n(\widetilde{T}_j^n)$ are adapted to the filtration.*
2. *For each $q \in K^n$, $j \in J$*

$$S_{(q)}(T_{(q)}^n(t)) - T_{(q)}^n(t)\ ,\ R_j^i(S_i(T_i^n(t))) - p_{ij}S_i(T_i^n(t)) \quad \text{are } \mathbb{F}_n\text{-martingales.}$$

BJPS -Accepted Manuscript

3. *Given any a.s-finite $\mathbb{F}_n$-stopping time τ, the conditional joint law of the N^n processes*

$$\{S_{(q)}(T^n_{(q)}(\tau)+s)-S_{(q)}(T^n_{(q)}(\tau)), s\geq 0, q\in K^n\}$$

conditional on $\mathcal{F}_n(t)$, is that of N^n i.i.d standard Poisson processes.

4. *For any $t\geq 0$ and any event $E_n\in\mathcal{F}_n(t)$, the N^n-dimensional process*

$$\{S_{(q)}(T^n_{(q)}(t)+s)-S_{(q)}(T^n_{(q)}(t)), s\geq 0, q\in K_n\},$$

the processes

$$\{A^n_j(\sigma^n_j(t)+s)-A^n_j(\sigma^n_j(t)), s\geq 0\},$$

and the event E_n are mutually independent where $\sigma^n_j(t)$ is a first jump time after a time t.

5. *The routing policy is work conserving, in the sense that for all $t\geq 0$,*

$$I^n_j(t)=(N^n_j-Q^n_j(t))^+. \tag{11}$$

We define processes at diffusion scale, as follows. We denote centered, normalized versions of the processes, for $j\in J$ and $t\geq 0$, by

$$\widehat{R}^i_j(t)=\frac{R^i_j(\lfloor nt\rfloor)-np_{ij}t}{\sqrt{n}},\qquad \widehat{I}^n(t)=\frac{I^n(t)}{\sqrt{n}},\qquad \widehat{Q}^n_j(t)=\frac{Q^n_j(t)-N^n_j}{\sqrt{n}}, \tag{12}$$

$$\widehat{A}^n_j(t)=\frac{A^n_j(t)-\lambda^n_j t}{\sqrt{n}},\qquad \widehat{S}^n_j(t)=\frac{S^n_j(nt)-nt}{\sqrt{n}}. \tag{13}$$

In addition, we denote

$$L^n_j(t)=n^{-\frac{1}{2}}\sum_{q\in K_j}\mu_{(q)}\int_0^t(I_{(q)}(s))ds. \tag{14}$$

The initial number of customers in the system is assumed to satisfy

$$\widehat{Q}^n(0)\Rightarrow\widehat{Q}(0),\quad \text{as } n\to\infty, \tag{15}$$

where $\widehat{Q}(0)$ is a J-dimensional random variable satisfying $\widehat{Q}(0)\geq 0$ with probability one.

Let w be a driftless J-dimensional Brownian motion having covariance matrix C with (k,l)th element

$$\bar{\lambda}_k a_k\delta_{kl}+\sum_{j=1}^J\bar{\mu}_j p_{jk}(\delta_{kl}-p_{jl})+\bar{\mu}_j(p_{jk}-\delta_{jk})(p_{jl}-\delta_{jl}),$$

BJPS - Accepted Manuscript

independent of $\widehat{Q}(0)$, and let $\mathcal{F}_t$ be the $\mathbb{P}$-completion of the smallest σ-field with respect to which $w(s), 0 \leq s \leq t$ and $\widehat{Q}(0)$ are measurable, and Γ is a $J \times J$ diagonal matrix with $\Gamma_{ii} = \gamma_i$. A pair $(\widehat{Q}, l)$ in $\mathcal{D}^{2J}$ will be said to be a solution to the J-dimensional Skorohod equation

$$\widehat{Q}(t) = \widehat{Q}(0) + \beta t + w(t) - \int_0^t \Gamma \widehat{Q}(s) ds + (I - P^T) l(t) \tag{16}$$

with data $(\widehat{Q}(0), w)$, if $\widehat{Q}$ and l are RCLL functions, $\{\mathcal{F}_t\}$-adapted processes satisfying the following conditions $\mathbb{P}$-a.s.:

- equation (16) holds;
- $\widehat{Q}(t) \geq 0$, $t \geq 0$;
- l_i is non-decreasing, for all $i \in J$;
- $\int_{[0,\infty)} \widehat{Q}_i(t) dl_i(t) = 0$ a.s.

Dupuis and Ishii [8, Theorem 3.3] show that there exists a unique solution to the general J-dimensional Skorohod equation, which includes cases as equation (16) ([8, Section 5.1]).

The following theorem argues that the multi-dimensional queue-length process converges to a multi-dimensional regulated Ornstein-Uhlenbeck process.

Theorem 2.2 *Let $\{A^n, Q^n, I^n, T^n\}$ be any sequence of J-dimensional processes satisfying all assumptions stated above. Then, $(\widehat{Q}^n, L^n, \widehat{I}^n)$ converges in distribution, uniformly on compacts, to $(\widehat{Q}, l, 0)$ where $(\widehat{Q}, l)$ denotes the unique solution to the Skorohod equation* (16) *with data $(\widehat{Q}(0), w)$.*

3 Proof

3.1 Setting and notation

Fix $u \in [1, \infty)$, $\varrho \in (0, 1/2)$. Let

$$\tau_n = \min_j \inf\{t \geq 0 : L_j^n(t) \geq n^\varrho\} \wedge u. \tag{17}$$

BJPS-Accepted Manuscript

In addition, we denote

$$V_j^n(t) = n^{-1/2}(S_j^n(T_j^n(t)) - T_j^n(t)), \tag{18}$$

$$\widetilde{V}_j^n(t) = \sum_{i=1}^{J} \widehat{R}_j^{i,n}(\bar{S}_i^n(T_i^n(t))). \tag{19}$$

In the sequel, we will use in fact that

$$\gamma_j \int_0^t \widehat{Q}_j^n(s)^+ ds + F_j^n(t) = n^{-1/2}\widetilde{S}_j\Big(\gamma_j^n n^{1/2} \int_0^t \widehat{Q}_j^n(s)^+ ds\Big) = n^{-1/2}\widetilde{S}_j^n(t) \tag{20}$$

is nondecreasing in t.

For $x \in \mathcal{D}^d$ and $u > 0$, let

$$||x||_u \equiv \sup_{0\le t\le u} \max_{j\in\{1,\dots,d\}} |x_j(t)|,$$

and

$$|x_j|_u^* \equiv \sup_{0\le t\le u} |x_j(t)|.$$

Finally, an operator $f : \mathcal{D}^d \to \mathcal{D}^d$ is called Lipschitz continuous if for any $u > 0$, there exists a constant κ_u such that for $x_1, x_2 \in \mathcal{D}^d$

$$||f(x_1) - f(x_2)||_u \le \kappa_u ||x_1 - x_2||_u.$$

A random variable X is tight if for each $\epsilon > 0$ there exists a compact set K such that

$$P(X \notin K) < \epsilon.$$

We denote (see [4] p. 80)

$$\bar{w}_u(x, \delta) = \sup_{s,t\in[0,u];\ |s-t|\le\delta} |x(t) - x(s)|,$$

for $x : [0, u] \to \mathbb{R}$, $\delta > 0$.

A sequence of processes defined on $[0, u]$, with sample paths in the Skorohod space, is said to be *C-tight* if it is tight, and every subsequential limit has continuous sample paths with probability one. C-tightness of, say $\{U_n\}$, implies the convergence in probability of $\bar{w}_u(U_n, \delta) \to 0$, for every δ.

BJPS -Accepted Manuscript

In the sequel, we will use the Burkholder-Davis-Gundy inequality (see [17] p. 58 and p. 175). For any local martingale X and $p \geq 1$,

$$\mathbb{E}\{(|X|_t^*)^p\} \leq c_p \mathbb{E}\{[X,X]_t^{p/2}\}, \quad t \in [0,\infty), \tag{21}$$

where the constant c_p depends only on p, and $[X,X]$ is defined by $[X,X] = X^2 - 2\int X_- dX$. By Theorem 22(ii) in [17], if X has piecewise smooth sample paths, null at zero, then $[X,X]_t$ is given by $\sum_{s \leq t} \Delta X(s)^2$.

3.2 Some Lemmas and Proof of Main Result

First, we note that the number of service completions by station-i servers until time t is

$$D_i^n(t) = \sum_{q \in K_i} S_{(q)}(T_{(q)}(t)). \tag{22}$$

We denote

$$T_i^n(t) = \sum_{q \in K_i^n} \mu_{(q)}^n \int_0^t (1 - I_{(q)}^n(s))ds, \quad 1 \leq i \leq J, t \geq 0. \tag{23}$$

The following proposition states that a station which servers are ruled by Poisson processes is ruled itself by a Poisson process

Proposition 3.1 *Fix n. Then, there are independent standard Poisson processes $\{S_1^n, ..., S_J^n\}$ such that*

$$D_i^n(t) = S_i^n(T_i^n(t)).$$

Proof: By assumption 2.1,

$$D_i^n(t) - T_i^n(t) \quad \textit{is an } \mathbb{F}_n\textit{-martingale.} \tag{24}$$

Theorem T9 in Bremaud[5] shows

$$\sum_{q \in K_i^n} \mu_{(q)}^n I_{(q)}^n(t) \quad \textit{is the} \quad \mathbb{F}_n\textit{-intensity.} \tag{25}$$

Since our system is critically loaded,

$$D_i^n(\infty) = \infty.$$

BJPS -Accepted Manuscript

For all t, define the $\mathbb{F}_n$-stopping time $\tau^n(t)$ by $\tau^n(t) = \inf\{a|T_i^n(a) = t\}$.

By Theorem T16 in [5], $S_i^n(u) := D_i^n(\tau^n(u))$ is a Poisson process with rate 1. To show that $D_i^n(t) = S_i^n(T_i^n(t))$, put $u = T_i^n(t)$ and then we obtain, $S_i^n(T_i^n(t)) = D_i^n(t')$ where $t' = \inf\{a|T_i^n(a) = T_i^n(t)\}$, (22) yields $S_i^n(T_i^n(t)) = D_i^n(t)$.

Now, we prove that $S_i^n(u)$ are independent random variables $(1 \le i \le J)$. By the definition,

$$S_i^n(u) = D_i^n(\tau^n(u)) = \sum_{q \in K_i} S_{(q)}(T_{(q)}(\tau^n(u))),$$

where $\sum_{q \in K_i^n} T_{(q)}(\tau^n(u)) = T_i^n(\tau^n(u)) = u$. Assumption 2.1 completes the proof.

□

To prove Theorem 2.2, we need the following Lemmas.

Lemma 3.2 *Define*

$$W^n(t) = \widehat{A}^n(t) + \sum_{k=1}^{J} \widehat{R}^{k,n}(\bar{S}_k^n(T_k^n(t))) - (I - P^T)\widehat{S}^n(\bar{T}^n(t)), \tag{26}$$

$$\widetilde{W}^n(t) = W^n(t) + \frac{1}{\sqrt{n}}[\lambda^n - (I - P^T)\mu^n]t, \tag{27}$$

$$\widetilde{S}_j^{*,n}(t) = n^{-1/2}\widetilde{S}_j(n^{1/2}t) - t, \quad t \ge 0, \tag{28}$$

$$F_j^n(t) = \widetilde{S}^{*,n}\Big(\gamma_j^n \int_0^t \widehat{Q}_j^n(s)^+ ds\Big) + (\gamma_j^n - \gamma_j) \int_0^t \widehat{Q}_j^n(s)^+ ds.$$

One has

$$\widehat{Q}^n(t) = \widehat{Q}^n(0) + \widetilde{W}^n(t) - F^n(t) + (I - P^T)L^n(t) - \int_0^t \Gamma \widehat{Q}^n(s)^+ ds. \tag{29}$$

Proof:

Let $Q_j^n(t)$ be the queue-length process at station j, then

$$Q_j^n(t) = Q_j^n(0) + A_j^n(t) + \sum_{k=1}^{J} R_j^k(S_k^n(T_k^n(t))) - S_j^n(T_j^n(t)) - \widetilde{S}_j(\widetilde{T}_j^n(t))$$

$$Q_j^n(t) - N_j^n = Q_j^n(0) - N_j^n + (A_j^n(t) - \lambda_j^n t) + \lambda_j^n t + \sum_{k=1}^{J}[R_j^k(S_k^n(T_k^n(t))) - p_{kj}S_k^n(T_k^n(t))]$$
$$+ \sum_{k=1}^{J} p_{kj}S_k^n(T_k^n(t)) - S_j^n(T_j^n(t)) - [\widetilde{S}_j(\widetilde{T}_j^n(t)) - \widetilde{T}_j^n(t)] - \widetilde{T}_j^n(t)$$

BJPS -Accepted Manuscript

$$Q_j^n(t) - N_j^n = Q_j^n(0) - N_j^n + (A_j^n(t) - \lambda_j^n t) + \lambda_j^n t + \sum_{k=1}^{J}[R_j^k(S_k^n(T_k^n(t))) - p_{kj}S_k^n(T_k^n(t))]$$

$$+\sum_{k=1}^{J} p_{kj}[S_k^n(T_k^n(t)) - T_k^n(t)] + \sum_{k=1}^{J} p_{kj}T_k^n(t) - [S_j^n(T_j^n(t)) - T_j^n(t)] - T_j^n(t) - [\widetilde{S}_j(\widetilde{T}_j^n(t)) - \widetilde{T}_j^n(t)] - \widetilde{T}_j^n(t)$$

and dividing by $\sqrt{n}$

$$\widehat{Q}_j^n(t) = \widehat{Q}_j^n(0) + \widehat{A}_j^n(t) + \sum_{k=1}^{J} \widehat{R}_j^k(\bar{S}_k^n(T_k^n(t))) + \sum_{k=1}^{J} p_{kj}\widehat{S}_k^n(\bar{T}_k^n(t)) - \widehat{S}_j^n(\bar{T}_j^n(t))$$
$$+ \frac{1}{\sqrt{n}}(\sum_{k=1}^{J} p_{kj}T_k^n(t) - T_j^n(t) + \lambda_j^n t) - \gamma_j \int_0^t \widehat{Q}_j^n(s)^+ ds - F_j^n(t).$$

By (23)

$$T_j^n(t) = \sum_{q\in K_j} \mu_{(q)} \int_0^t (1 - I_{(q)}(s))ds = \mu_j^n t - \sum_{q\in K_j} \mu_{(q)} \int_0^t (I_{(q)}(s))ds$$

$$\widehat{Q}_j^n(t) = \widehat{Q}_j^n(0) + \widehat{A}_j^n(t) + \sum_{k=1}^{J} \widehat{R}_j^k(\bar{S}_k^n(T_k^n(t))) + \sum_{k=1}^{J} p_{kj}\widehat{S}_k^n(\bar{T}_k^n(t)) - \widehat{S}_j^n(\bar{T}_j^n(t))$$
$$+ \frac{1}{\sqrt{n}}[\lambda_j^n + \sum_{k=1}^{J} p_{kj}\mu_k^n - \mu_j^n]t - \gamma_j \int_0^t \widehat{Q}_j^n(s)^+ ds - F_j^n(t)$$
$$- n^{-\frac{1}{2}} \sum_{i=1}^{J} p_{ij} \sum_{q\in K_i} \mu_{(q)} \int_0^t (I_{(q)}(s))ds + n^{-\frac{1}{2}} \sum_{q\in K_j} \mu_{(q)} \int_0^t (I_{(q)}(s))ds.$$

□

Now, we prove some elementary estimates of expressions have been defined above.

Lemma 3.3 *With* $\bar{T}_j^n(t) := \frac{1}{n}T_j^n(t)$, $\bar{T}_j(t) := \bar{\mu}_j t$, $t \in [0, u]$, *one has*

$$\bar{T}_j^n \to \bar{T}_j \quad \textit{in probability, uniformly on } [0, u], \tag{30}$$

$$\sup_n \mathbb{E}[(||\widetilde{V}^n||_u)^2] < \infty, \tag{31}$$

$$\sup_n \mathbb{E}[(||V^n||_u)^2] < \infty, \tag{32}$$

$$\textit{the random variables} \quad ||\widehat{Q}^{n,+}||_u \quad \textit{are tight}, \tag{33}$$

$$F^n \to 0 \textit{ in probability, uniformly on } [0, u], \tag{34}$$

BJPS -Accepted Manuscript

and

$$\mathbb{P}(\tau_n < u) \to 0, \quad as\ n \to \infty. \tag{35}$$

Proof: First, the proof of the following limit is similar to Lemma A.2 in Atar[1].

$$\sup\{|\bar{T}_j^n(t) - \bar{\mu}_j t| : t \le \tau_n\} \to 0 \quad \text{in probability, as } n \to \infty. \tag{36}$$

We continue to show (31) by using the Burkholder-Davis-Gundy inequality (see (21)). From the assumption 2.1 it follows that $\widetilde{V}_n$ is an $\mathbb{F}_n$-martingale. Since each of its jumps are of size $n^{-1/2}$ and the number of jumps until time t is $\sum_{i=1}^{J} R_j^{i,n}(S_i^n(T_i^n(t)))$, it follows that

$$\begin{aligned}
\mathbb{E}[(|\widetilde{V}_j^n|_u^*)^2] &\le \frac{c_2}{n}\mathbb{E}[\sum_{i=1}^{J} R_j^{i,n}(S_i^n(T_i^n(u)))] = \sum_{i=1}^{J}\frac{c_2}{n}\mathbb{E}[p_{ij}(S_i^n(T_i^n(u)))] \\
= \sum_{i=1}^{J}\frac{c_2 p_{ij}}{n}\mathbb{E}[(T_i^n(u))] &= \sum_{i=1}^{J}\frac{c_2 p_{ij}}{n}\mathbb{E}\Big[\sum_{q\in K_j}\mu_{(q)}\int_0^u (1 - I_{(q)}^n(s))ds\Big] \\
&\le \sum_{i=1}^{J} c_2 p_{ij} u\, \bar{\mu}_j^n,
\end{aligned}$$

and by (6) follows (31).

(32) is immediately obtained in the same way.

We now prove

$$\text{the random variables} \quad ||\widehat{Q}^{n,+}||_{\tau_n} \quad \text{are tight.} \tag{37}$$

Let $r > 0$ be given. Consider the event $\{||\widehat{Q}^{n,+}||_{\tau_n} > 2r\}$. On this event there exists $j \in J$, $0 \le \alpha_n \le \theta_n \le \tau_n$ such that $\widehat{Q}_j^n(\alpha_n) \le r$, $\widehat{Q}_j^n(\theta_n) \ge 2r$, while $\widehat{Q}_j^n(t) > 0$ for $t \in [\alpha_n, \theta_n]$. Hence by (11), (12) and (14), $L_j^n(\theta_n) = L_j^n(\alpha_n)$. Recall that, $\sum_{i=1}^{J} p_{ij} L_i^n(t)$ and the process (20) are nondecreasing in t. Using this in the jth component of Lemma 3.2, shows

$$\mathbb{P}(||\widehat{Q}^{n,+}||_{\tau_n} > 2r) \le \mathbb{P}(\text{there exist } 0 \le \alpha^n \le \theta^n \le \tau^n \text{ such that } \widetilde{W}_j^n(\theta^n) - \widetilde{W}_j^n(\alpha^n) \ge r), \tag{38}$$

and since $\widehat{A}^n$ converge and (7), (32) and (31) hold, it follows that $||\widetilde{W}^n||_u$ are tight random variables. This shows that the r.h.s. of (38) tends to zero as $r \to \infty$, whence follows (37).

It is easy to see that $\widetilde{S}^{*,n}$ converges to zero in distribution, uniformly on compacts. By using (37), $\gamma_j^n \to \gamma_j$,

$$|F^n|_{\tau_n}^* \to 0, \quad \text{in probability.} \tag{39}$$

BJPS - Accepted Manuscript

By the assumptions on the matrix P, there is a diagonal matrix Λ having positive diagonal elements such that the matrix $P^* = \Lambda P \Lambda^{-1}$ has maximal row sum less than one (Veinott, see Harrison and Rieman[12] page 304). Hence (see [12]), the equation in Lemma 3.2 may be replaced by the equation

$$\Lambda \widehat{Q}^n(t) = \Lambda(\widehat{Q}^n(0) + \widetilde{W}^n(t) - F^n(t) - \int_0^t \Gamma \widehat{Q}^n(s)^+ ds) + (I - P^{*T})\Lambda L^n(t). \tag{40}$$

We observe the sum of all jth components of (40). Let $\{\tau_n < u\}$ be an event. On this event there exists k such that $L_k^n(\tau^n) = n^\varrho$. By the property of P^*,

$$\sum_{j=1}^J \{\lambda_j L_j^n(t) - \sum_{i=1}^J p_{ij}^* \lambda_i L_i^n(t)\} \quad \textit{is positive} \tag{41}$$

and

$$|\sum_{j=1}^J \{\lambda_j L_j^n(t) - \sum_{i=1}^J p_{ij}^* \lambda_i L_i^n(t)\}|_{\tau_n}^* \geq (1 - \sum_{i=1}^J p_{ki}^*)\lambda_k n^\varrho. \tag{42}$$

Using (41) and (42), show

$$\begin{aligned}
\mathbb{P}(\tau_n < u) \leq \quad & \mathbb{P}\big(|[\sum_{j=1}^J \lambda_j \widehat{Q}_j^n(\cdot) - \lambda_j \widehat{Q}_j^n(0)]^+|_{\tau_n}^* \geq (1 - \sum_{i=1}^J p_{ki}^*)\lambda_k n^\varrho/4\big) \\
& + \mathbb{P}(|\sum_{j=1}^J \lambda_j \widetilde{W}_j^n|_{\tau_n}^* \geq (1 - \sum_{i=1}^J p_{ki}^*)\lambda_k n^\varrho/4) \\
& + \mathbb{P}\Big(\sum_{j=1}^J \gamma_j \int_0^{\tau_n} \lambda_j \widehat{Q}_j^n(s)^+ ds \geq (1 - \sum_{i=1}^J p_{ki}^*)\lambda_k n^\varrho/4\Big).
\end{aligned}$$

We showed that the random variables $||\widetilde{W}_n||_{\tau_n}$ are tight. The tightness of $||\widehat{Q}^n(0)||$, $||\widehat{Q}^{n,+}||_{\tau_n}$ (see (15) and (37)) implies (35).

Finally, in view of (35), (36) implies (30), (37) implies (33) and (39) implies (34). □

Lemma 3.4 *Let $\widetilde{W}^n$ be defined as in (27). Then $\widetilde{W}^n \Rightarrow \widetilde{w}$ where $\widetilde{w}$ is a J-dimensional Brownian motion with drift β having covariance matrix C with (k,l)th element*

$$\bar{\lambda}_k a_k \delta_{kl} + \sum_{j=1}^J \bar{\mu}_j p_{jk}(\delta_{kl} - p_{jl}) + \bar{\mu}_j (p_{jk} - \delta_{jk})(p_{jl} - \delta_{jl}).$$

BJPS-Accepted Manuscript

Proof: We note that the J-dimensional processes

$$\widehat{A}^n, \widehat{S}^n, \widehat{R}^n \quad \textit{converge in distribution to} \quad w_A, w_S, w_R, \tag{43}$$

respectively (see [7, Theorem 5.11]), where w_A, w_S, w_R are independent J-dimensional drift-less Brownian motions having covariance matrices $C^A, C^S, C^{R,i}$, such that (k,l)th element is defined as ([7, Section 7.5])

$$C^A_{kl} = \bar{\lambda}_k a_k \delta_{kl}, \quad C^S_{kl} = \delta_{kl}, \quad C^{R,i}_{kl} = p_{ik}(\delta_{kl} - p_{il}).$$

We use (43), the random time change theorem([4, section 14]). By (30)

$$\widehat{S}^n(\bar{T}^n(t)) \Rightarrow w_S(\bar{\mu}t).$$

By [7, Theorem 5.10], $\frac{1}{n}S^n_j(n\bar{T}^n_j) \to \bar{\mu}_j t$ a.s . Hence,

$$\widehat{R}^{j,n}(\bar{S}^n_j(T^n_j(t))) \Rightarrow w_{R,j}(\bar{\mu}_j t).$$

Using (7), completes the proof.

□

Proof of Theorem 2.2. We observe again the sum of all jth components of (40). Recall that (41) holds. Hence, the tightness of the random variables $|\sum_{j=1}^{J} \widehat{Q}^{n,+}_j|^*_u$ and $|\sum_{j=1}^{J} \widetilde{W}^n_j|^*_u$ shows that (41), and, in turn, $|L^n_i|^*_u$ are tight random variables for all i.

Now, the tightness of $||L^n||_u$, $|\widehat{Q}^{n,+}_j|^*_u$ and $|\widetilde{W}^n_j|^*_u$ implies that $|\widehat{Q}^n_j|^*_u$, and, in turn, $|\widehat{I}^n_j|^*_u$ are tight random variables.

We show that

$$||\widehat{I}^n||_u \to 0 \quad \text{in probability.} \tag{44}$$

Given $\varepsilon > 0$, consider the event

$$\Omega^\varepsilon_n = \{\widehat{I}^n_j(0) \geq \varepsilon,\ |\widehat{I}^n_j|^*_u > 3\varepsilon\}.$$

On this event, there exists $0 \leq \alpha_n < \theta_n \leq u$ such that $\widehat{I}^n_j(\alpha_n) \leq 2\varepsilon$, $\widehat{I}^n_j(\theta_n) \geq 3\varepsilon$, and $\widehat{I}^n_j(t) \geq \varepsilon$ for $t \in [\alpha_n, \theta_n]$ (by tightness of $\widehat{I}^n_j$, the jumps of $\widehat{I}^n_j$ are a.s. bounded by $cn^{-1/2}$). Hence by (5) and (14),

$$L^n_j(\theta_n) - L^n_j(\alpha_n) \geq \mu^n_{min}\varepsilon(\theta_n - \alpha_n).$$

BJPS - Accepted Manuscript

Since $\widehat{Q}_j^n$ is negative on this event for $t \in [\alpha_n, \theta_n]$, the last term on the r.h.s. of Lemma 3.2 does not vary between the times α_n and θ_n. Also, the process L_j^n is nondecreasing, and so

$$-\varepsilon \geq \widehat{Q}_j^n(\theta_n) - \widehat{Q}_j^n(\alpha_n) \geq \widetilde{W}_j^n(\theta_n) - \widetilde{W}_j^n(\alpha_n) - \sum_{i=1}^J p_{ij} L_i^n(\theta_n) + \sum_{i=1}^J p_{ij} L_i^n(\alpha_n) := WL^n(\theta_n) - WL^n(\alpha_n).$$

We obtain, for each ε and n,

$$\mathbb{P}(\Omega_n^\varepsilon) \leq \mathbb{P}(\text{there exists } \delta > 0 \text{ such that } \bar{w}_u(WL^n, \delta) \geq \varepsilon,\ \varepsilon\delta\mu_{min}^n \leq |L_j^n|_u^*).$$

By tightness of $|L_j^n|_u^*$, there is a function g such that $\lim_{r\to\infty} g(r) = 0$, and, for every $r > 0$,

$$\mathbb{P}(\Omega_n^\varepsilon) \leq g(r) + \mathbb{P}\Big(\bar{w}_u\Big(WL^n, \frac{r}{\varepsilon\mu_n^{\min}}\Big) \geq \varepsilon\Big). \tag{45}$$

We assumed that $\mu_{min}^n \to \infty$ (5). Since L_i is continuous and $\widetilde{W}^n$ converges to a Brownian motion, $\sum_{i=1}^J p_{ij} L_i^n$, $\widetilde{W}^n$ are C-tight. Hence, the last term on (45) converges to zero as $n \to \infty$. In fact, r is arbitrary, so that $\lim_n \mathbb{P}(\Omega_n^\varepsilon) = 0$. Finally, since the weak limit $\widehat{I}_j(0)$ of $\widehat{I}_j^n(0)$ is nonnegative (15), it follows that $\lim_n \mathbb{P}(|\widehat{I}_j^n|_u^* > 3\varepsilon) = 0$. This shows (44).

By Lemma 3.2,

$$\widehat{Q}^n(t)^+ = \zeta^n(t) + (I - P^T)L^n(t), \tag{46}$$

where

$$\zeta^n(t) := \widehat{Q}^n(0) + \widetilde{W}^n(t) - \int_0^t \Gamma\widehat{Q}^n(s)^+ ds + e_n(t) \quad \textit{and} \quad e_n := \widehat{I}^n(t) - F^n(t) \tag{47}$$

- $\widehat{Q}^n(t)^+ \geq 0$, $\forall t \geq 0$;
- $L_i^n(0) = 0$, L_i^n is non-decreasing, for all $i \in J$;
- $\int_{[0,\infty)} \widehat{Q}_i^n(t)^+ dL_i^n(t) = 0$ a.s.

Dupuis and Ishii [8, Theorem 3.3] show that there exists a unique map $\widehat{Q}^{n,+} = \Phi(\zeta^n)$ from $\mathcal{D}^J$ to $\mathcal{D}^J$ which is Lipschitz continuous under uniform norm (see [8, Section 5.1]), such that $(\widehat{Q}^{n,+}, \zeta^n, \widehat{Q}^{n,+} - \zeta^n)$ solves the Skorohod equation (46).

By the tightness of $\widehat{Q}^n(0)$, $\widehat{Q}^{n,+}$, $\widetilde{W}_n$ and $F_n \to 0$ ((15), (33), (31), (32), (7) and (34)), it is sufficient to show that all subsequential limits of $(\widehat{Q}^{n,+}, \zeta^n, \widehat{Q}^{n,+} - \zeta^n)$ are equal to the solution of the Skorohod equation (16).

BJPS - Accepted Manuscript

A Lemma 3.4, (44), (34), the continuity of Φ and a continuous mapping theorem complete the proof.

□

4 Discussion and future research

The research on networks of queues has important applications in computer science, telecommunications, and large manufacturing systems. Since exact analysis proves impossible in most cases, a large part of the research has focused on approximate models. When the service rates are roughly balanced with the arrival rates, one can approximate such systems by suitable diffusion processes.

This paper presents a wide range of possible approximations of such systems. As we mentioned in the introduction, the research of the general α-parameterization has advanced considerably in recent years, thus giving rise to a considerably larger collection of possible regime models to provide better approximations. It is reasonable to suppose that optimization problems concerning heavy traffic, with the decision variable being the number of servers, can be solved by an α-parameterization model with $\alpha \neq 0, 1$ (for example, see [13]).

Now, we present the discussion and some possibilities for future studies.

1. In this paper, we considered a Jackson network under heavy traffic with any parameter $\alpha \in [0, 1)$. We showed that the sequence of normalized queue length processes of the Jackson network converge weakly to a multi-dimensional regulated Ornstein-Uhlenbeck process in the orthant, (or regulated Brownian motion, if we omit from the model the abandonment of customers) as the traffic intensity approaches unity. However, it is not known whether the stationary distribution of regulated Ornstein-Uhlenbeck process provides a valid approximation for the steady-state of the original network. This problem solved by [9] under conventional heavy traffic regime

2. Kleinrock [14] found the optimal vector of service rates $\mu = (\mu_1, \ldots, \mu_k)$ in a Jackson network (without heavy traffic), in order to minimize the sojourn time per customer subject to the budget constraint $D = \sum d_i \mu_i$, where d_i is the unit cost of capacity at

BJPS - Accepted Manuscript

station k and D is the total available budget. Wein [20] generalized this result to general arrival and service time distributions.

The question arises whether it is possible to extend this result to a Jackson network in any general heavy traffic model. For example, given a Jackson network in a heavy traffic regime with a fixed parameter $\alpha \in [0,1)$, so that the vectors λ^n, μ^n satisfy the assumptions (5)-(7), one can search the optimal vector of service rates $\mu^n = (\mu_1^n, \ldots, \mu_k^n)$, in order to minimize the sojourn time per customer subject to the budget constraint $D^n = \sum \mu_i^n$ where D^n is the total available budget. Thereafter, one can search for the parameter α which gives the absolutely minimal sojourn time.

3. In Jackson networks many research has been done in regard to control problems. For example, Azaron and Ghomi [3] considered optimal control of service and arrival rates in a Jackson network. They studied the total waiting times and the total operating costs per period in Jackson networks. The question arises whether it is possible to extend this result to a Jackson network in any general heavy traffic model.

4. It is well-documented that in a considerable class of the aforementioned research applications, the service times are not exponentially distributed [6]. Thus arises naturally the question to know whether similar results can be obtained by dropping the assumption of exponentially distributed serves times, whereby the main problem lies in the fact it is not known how to determine in general the distribution of the sum of non-Poissonian server processes.

Acknowledgments: I thank Rami Atar for his inspiring discussions and valuable comments.

References

[1] Atar, R. (2012). A diffusion regime with non-degenerate slowdown. Operations Research **2**, 490-500.

BJPS -Accepted Manuscript

[2] Atar, R. and Solomon, N. (2011). Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown. Queueing Systems **3** 217-235.

[3] Azaron, A. and Fatemi Ghomi, S.M.T. (2003). Optimal control of service rates and arrivals in Jackson networks. European Journal of Operational Research **147**, 17-31.

[4] Billingsley, P. (1999). Convergence of Probability Measures. John Wiley and Sons, Inc.

[5] Bremaud, P. (1981). Point processes and queues, martingale dynamics. Springer-Verlag, New York.

[6] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao. L. (2005). Statistical analysis of a telephone call center. Journal of the American statistical association **469** 36-50.

[7] Chen, H. and Yao, D. D. (2001). Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization. Springer-Verlag, New York.

[8] Dupuis, P. and Ishii, H. (1991) On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. Stochastics Stochastics Reports **1**, 31-62.

[9] Gamarnik, D. and Zeevi, A. (2006). Validity of heavy traffic steady-state approximations in generalized Jackson networks. The Annals of Applied Probability **1**, 56-90.

[10] Gurvich. I. (2004). Design and control of the M/M/N queue with multi-class customers and many servers. M.Sc. Thesis, Technion.

[11] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. Operations Research **3**, 567-588.

[12] Harrison, J. M. and Reiman, M. I. (1981). Reflected Brownian motion on an orthant. Annals of Probability **9**, 302-308.

[13] Hassin, R., Shaki, Y.Y. and Yovel, U. (2015). Optimal service capacity allocation in a loss system. Naval Research Logistics. **62**, 81-97.

BJPS -Accepted Manuscript

[14] Kleinrock R. L. (1964). Communication Nets: Stochastic message flow and delay. Dover publications, Inc, New York.

[15] Mandelbaum, A., Massey, W.A. and Reiman, M. (1998). Strong approximations for Markovian service networks. Queueing Systems, **30**, 149-201.

[16] Mandelbaum. A. (2003). QED Q's. Notes from a lecture delivered at the Workshop on Heavy Traffic Analysis and Process Limits of Stochastic Networks, EURANDOM. http://ie.technion.ac.il/serveng/References/references.html

[17] Protter. P. E. (2004). Stochastic integration and differential equations. Second edition. Springer-Verlag, Berlin.

[18] Reed, J. E. and Ward A. R. A (2004). Diffusion Approximation for a Generalized Jackson Network with Reneging. Proceedings of the 42nd Annual Conference on Communication, Control, and Computing. Sept. 29-Oct. 1.

[19] Reiman, M. I. (1984). Open queueing networks in heavy traffic. Mathematics of Operations Research **9**, 441-458.

[20] Wein L.M. (1989). Capacity allocation in generalized Jackson networks. Operations Research Letters **8**, 143-146.

[21] Whitt, W. (2003). How multiserver queues scale with growing congestion-dependent demand. Operations Research **4**, 531-542.

BJPS -Accepted Manuscript