Blind, individually-fair routing for heterogeneous servers in the Halfin-Whitt Regime

Rami Atar* Yair Y. Shaki† Adam Shwartz‡

Department of Electrical Engineering
Technion–Israel Institute of Technology
Haifa 32000, Israel

May 7, 2009

1 Introduction

This paper is a continuation of [3]. In this paper, we consider a queueing system with many servers. Customers arrive into the system according to a Poisson process, and are either queued in a buffer with infinite room or routed to one of the servers according to a routing policy. Customers from the buffer are routed to servers according to a first-come-first-served rule. We consider work conserving routing policy, so that no server may be idle when at least one customer is in the buffer. Each customer leaves the system when its service requirement is fully processed. For servers, service times are independent exponentially distributed.

We propose the following routing policy: each customer that arrives when least both idle servers is routed to the server with the longest accumulated idle time. We show that this is fair in the heavy traffic limit.

*Research supported in part by the Israel Science Foundation (Grant 1349/08) and the fund for promotion of research at the Technion

†Research supported in part by the Viterbi Postdoctoral Fellowship and the ISF (Grant 1349/08)

‡The Julius M. and Bernice Neiman Chair in Engineering. Research supported in part by the fund for promotion of research at the Technion.

2 Setting, notation and main result

Customers arrive to a system according to a renewal process denoted by $A(t)$. Each arrival has a single noninterruptible service requirement. Arriving customers are routed to one of N servers, according to a routing policy, provided a free server is available: if not, they are queued in a buffer with infinite room. Customers from the buffer are routed to servers according to a first-come-first-served rule. We consider work conserving routing policy, so that no server may be idle when at least one customer is in the buffer. Each customer leaves the system when its service requirement is fully processed.

There are N servers, the servers are labeled $1, \dots, N$. We write K for $\{1, \dots, N\}$. Server k serves according to an exponential service time distribution with rate μ_k.

All processes are assumed to have right-continuous sample paths.

For $k \in K$ and $t \geq 0$, let $I_k(t)$ take the value 1 if server k is idle at time t, and let it be 0 otherwise. Set $Z_k = 1 - I_k$. Let

$$J_k(t) = \int_0^t I_k(s)ds, \qquad k \in K, t \geq 0.$$

Denote by $I(t)$ the number of idle servers at time t, (see (1)) and define $Z(t)$ in a similar manner. Then both I and Z are stochastic processes taking values in $[0, N]$, and

$$I = \sum_{k \in K} I_k = N - Z. \tag{1}$$

The modeling of service completions will require usage of standard Poisson processes $S_k, k \in K$ with rate 1. Arrivals are according to a renewal process with finite second moment for the interarrival time. The number of service completions by all servers until time t is denoted by $D(t)$, and it can be represented as

$$D(t) = S(T(t)) = \sum_{k \in K} S_k(T_k(t)),$$

where T_k, T are defined as

$$T_k(t) = \mu_k \int_0^t Z_k(s)ds \qquad T(t) = \sum_{k \in K} T_k(t), \quad t \geq 0. \tag{2}$$

The number-in-system and the number-in-buffer processes are denoted by X and Q, respectively. The initial configuration, namely,

$$(\{I_k(0), k \in K\},\, Q(0)),$$

and the processes A and S_k, $k \in K$, are assumed to be mutually independent entities.

Let

$$J = \sum_{k \in K} J_k.$$

Then $J(t)$ represents the overall idleness time accumulated by servers until time t.

We will say that a routing policy is *work conserving* if for all $t \geq 0$,

$$Q(t) = (X(t) - N)^+, \quad \text{or equivalently} \quad I(t) = (N - X(t))^+.$$

Note that this imposes an assumption on the initial configuration as well as on the policy.

The routing policy we propose prioritizes servers according to their accumulated idle time: more precisely

- The policy is *work conserving*; in particular, when a server becomes available and there is a customer in the queue, a customer is routed to the server. When a customer arrives to find some available servers, it is routed to one of them.
- If a customer is to be routed at time t to an available server, and $AV(t) \subset K$ denotes the set of available servers at this time, it is routed to any one of the available servers $j \in AV(t)$ for which

$$J_j(t) \geq J_i(t) \quad \text{for all } i \in AV(t). \tag{3}$$

To formulate the notion of a Halfin Whitt regime, we consider a sequence of systems, parameterized by n, where the number of servers in the nth system is $N^n = n$. All processes receive a superscript n (which do and which don't; which parameters do and which don't depend on n). There's no need however to parameterize the standard Poisson processes S_k.

The rate of arrival λ^n is assumed to satisfy $\lambda^n/n \to \lambda \in (0,\infty)$ and moreover,

$$\widehat{\lambda}_n =: \frac{\lambda^n - \lambda n}{\sqrt{n}} \to \widehat{\lambda} \in \mathbb{R}. \tag{4}$$

The parameters μ_k^n are assumed to satisfy

$$\underline{\mu} \le \mu_k^n \le \overline{\mu}, \qquad k \in K,\, n \in \mathbb{N}, \tag{5}$$

where $0 < \underline{\mu} < \overline{\mu} < \infty$ are constants. In addition, it is assumed that the limits

$$\bar{\mu}^n := \frac{1}{n} \sum_{k \in K^n} \mu_k^n \to \mu \in [\underline{\mu}, \overline{\mu}], \tag{6}$$

and

$$\widehat{\mu}^n := \frac{1}{\sqrt{n}} \sum_{k \in K^n} (\mu_k^n - \mu) \to \widehat{\mu} \in \mathbb{R}, \tag{7}$$

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

$$\lambda = \mu. \tag{8}$$

We will also denote $\beta_n = \widehat{\lambda}^n - \widehat{\mu}^n$ and assume

$$\widehat{\beta} := \widehat{\lambda} - \widehat{\mu} < 0. \tag{9}$$

Note that the random variable $X^n(0)$ is given by $Q^n(0) + Z^n(0)$. The 'second order asymptotics' of $X^n(0)$ is assumed to satisfy

$$n^{-\frac{1}{2}}(X^n(0) - n) \text{ is a tight sequence of random variables.} \tag{10}$$

Given $\varepsilon > 0$ let $\gamma^n(\varepsilon) := \inf\{t : \sqrt{n}\widetilde{J}^p(t) \ge \varepsilon\}$. Our main result states that under a our policy, given any level of precision, equalization of the cumulative idleness processes is achieved soon after $\gamma^n(\varepsilon)$, in the large n limit, with large probability.

Theorem 2.1 *Let $\widetilde{J}^p$ be the cumulative idle time of server which is the closest to the pth percentile. For every $\varepsilon > 0$, $T > 0$ and $0 < p < 1$*

$$\lim_{n\to\infty} P\Big\{ \sup_{s\in[0,T]} \Big|\widetilde{J}^{1-p}(s) - \widetilde{J}^p(s)\Big| \ge \varepsilon \Big\} = 0.$$

Moreover, for any $t \geq 0$, the random variables $\gamma^n = \gamma^n(\varepsilon)$ are tight, and one has

$$\liminf_{n\to\infty} P\Big\{\gamma^n < \infty \quad \textit{and} \quad \Big|\frac{\widetilde{J}^{1-p}(\gamma^n + t)}{\widetilde{J}^p(\gamma^n + t)} - 1\Big| \leq \varepsilon\Big\} \geq 1 - \varepsilon.$$

Note that measuring fairness in terms of ratios is meaningful only when $J^n > 0$. This is why the formulation of the last assertion above involves γ^n. As will be clear from the proof of the result, in case that the random variables $\widehat{X}^n(0)$ (10) are further assumed to be bounded above by some $-\delta < 0$, the random times $\gamma^n(\varepsilon)$ will be small with probability tending to 1 (as $n \to \infty$), provided that ε is sufficiently small. In this case, the above result asserts that equalization is attained soon after time zero.

The following is almost an immediate consequence of the above result. Its purpose is to emphasizes that equalization is in fact achieved (with high probability) after sufficiently large time.

Theorem 2.2 *Under the hypotheses of Theorem 2.1, for every $\varepsilon > 0$ there exists T such that for every $T_1 \in [T, \infty)$,*

$$\liminf_{n\to\infty} P\Big\{J^n(T) > 0 \quad \textit{and} \quad \sup_{s\in[T,T_1]} \Big|\frac{\widetilde{J}^{1-p}(s)}{\widetilde{J}^p(s)} - 1\Big| \leq \varepsilon\Big\} \geq 1 - \varepsilon.$$

3 Proof

The main result will be proved by diffusion scale analysis. To this end, we define processes at diffusion scale, as follows. We denote centered, normalized versions of the processes at diffusion scale, for $k \in K$ and $t \geq 0$, by

$$\widehat{A}(t) = \frac{A(nt) - \lambda^n t}{\sqrt{n}}, \qquad \widehat{Q}(t) = \frac{Q^n(t)}{\sqrt{n}}, \qquad \widehat{J}(t) = \frac{J^n(t)}{\sqrt{n}}, \tag{11}$$

$$\widehat{S}_k(t) = \frac{S_k(nt) - nt}{\sqrt{n}}, \qquad \widehat{X}(t) = \frac{X^n(t) - n}{\sqrt{n}}, \qquad \widehat{I}(t) = \frac{I^n(t)}{\sqrt{n}} \tag{12}$$

We will also use The fluid-scale process

$$\bar{T}_k(t) = \frac{1}{n} T_k^n(t) \ . \tag{13}$$

Lemma 3.1 *Define*

$$F^n(t) = \sum_{k\in K} \mu_k \int_0^t \widehat{I}_k(s)ds, \tag{14}$$

$$W^n(t) = \widehat{A}^n(t) - \sum_{k\in K} \widehat{S}_k(\bar{T}_k(t)) \tag{15}$$

$$\widehat{\beta}^n = \frac{1}{\sqrt{n}}\Big(\lambda^n - \sum_{k\in K} \mu_k\Big). \tag{16}$$

Then

$$\widehat{X}^n(t) - \widehat{X}^n(0) = W^n(t) + \widehat{\beta}^n t + F^n(t), \tag{17}$$

Proof: See [3].

Throughout, we let $|f|_t^* = \sup_{0\le s\le t} |f(s)|$. Denote the modulus of continuity of a function f by

$$w_\theta(f,\delta) := \sup_{0\le s\le t\le (s+\delta)\wedge\theta} |f(t) - f(s)|, \qquad f:[0,\theta]\to\mathbb{R},\ \delta>0.$$

A sequence of processes defined on $[0,\theta]$, 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 $\{X^n\}$, implies tightness of $|X^n|_\theta^*$ and $\lim_{\delta\to 0}\limsup_n w_\theta(X^n,\delta)\to 0$, (see [4, Section 7]). These facts will be used in the sequel in conjunction with the application of the following lemma.

Lemma 3.2 *Given any $\theta\in(0,\infty)$, the sequence of random variables*

$$\{|\widehat{A}^n|_\theta^* \vee |\widehat{S}(\bar{T})|_\theta^* \vee |\widehat{I}|_\theta^*\}_{n=1}^\infty$$

is tight. In fact, $\{\widehat{A}^n\}_{n\in\mathbb{N}}$ and, $\{\widehat{S}^n\circ\bar{T}^n\}_{n\in\mathbb{N}}$, are C-tight. Furthermore, given any $\varepsilon_1,\varepsilon_2>0$ there exists t_1 such that

$$\limsup_{n\to\infty} P(|R^n|_t^* \ge \varepsilon_1 t) \le \varepsilon_2, \qquad t\ge t_1, \tag{18}$$

where R^n is any one of the processes $\widehat{A}^n$ or $\widehat{S}^n\circ\bar{T}^n$.

Proof: See [3].

Lemma 3.3 *The processes* $\{\widehat{X}^n\}_{n\in\mathbb{N}}$ *and* $\{\widehat{I}\}_{n\in\mathbb{N}}$ *are C-tight.*

Proof: Using the equation (17),

$$\widehat{X}^n(t) - \widehat{X}^n(0) = \widetilde{W}^n(t) + F^n(t), \tag{19}$$

where $\widetilde{W}^n(t) = W^n(t) + \widehat{\beta}^n t$

It is easy to see for any s, t such that $0 \le s \le t \le (s+\delta) \wedge \theta$,

$$0 \le F^n(t) - F^n(s) \le \bar{\mu}(t-s)|\widehat{I}|^*_\theta$$

Therefore

$$w_\theta(\widehat{X}^n, \delta) \le w_\theta(\widetilde{W}^n, \delta) + w_\theta(F^n, \delta) \le w_\theta(\widetilde{W}^n, \delta) + \bar{\mu}\delta|\widehat{I}|^*_\theta$$

By Lemma 3.2, $\widetilde{W}^n$ are C-tight. Since $|\widehat{I}|^*_\theta$ is tight (See (10)), the convergence in probability $w_\theta(\widehat{X}^n, \delta) \to 0$ as $\delta \to 0$, By [4, Section 7] it follows that $\{\widehat{X}^n\}_{n\in\mathbb{N}}$ are C-tight. From $I(t) = X^-(t)$, we obtain $\{\widehat{I}^n\}_{n\in\mathbb{N}}$ is also C-tight.

Lemma 3.4 *Let* i *be any server and let* $\Delta^n(t) = \sqrt{n}(J_i(t) - \widetilde{J}^p(t))^+$. *Then, for any* θ, $|\Delta^n|^*_\theta \to 0$ *in probability as* $n \to \infty$. *In fact, the convergence of* $|\Delta^n|^*_\theta$ *is uniform on* K, *i.e,*

$\forall \gamma, \delta > 0 \quad \exists n_0$ *such that* $\forall i$ *and* $\forall n > n_0$, $P(|\Delta^n|^*_\theta > \delta) < \gamma$

Proof: We start by representing result about a scenario where no jobs are routed to any servers set (pool) and the empty queue within a given interval. More precisely, fix $n \in \mathbb{N}$, fix $\theta > 0$, and let η and ζ be $[0, \theta]$-valued random variables such that $\eta \le \zeta$. The collection of servers that have accumulated strictly less idle time than $\widetilde{J}^p$ at time η is denoted $L(\eta)$. Let H be any event under which

- $Q = 0$ within the interval $[\eta, \zeta]$; and
- no jobs are routed to L within the same interval.

Then, with the notation $Y[a,b] = Y(b) - Y(a)$, under H (see [3]),

$$\sum_{j\in L^c} \widehat{I}_j[\eta,\zeta] + \widehat{A}[\eta,\zeta] - \sum_{j\in L^c} \widehat{S}_j \circ \bar{T}_j[\eta,\zeta] - \frac{1}{\sqrt{n}} \sum_{j\in L^c} \mu_j \int_\eta^\zeta Z_j(s)ds + \frac{\lambda^n}{\sqrt{n}}(\zeta - \eta) = 0. \qquad (20)$$

In what follows, fix $\varepsilon > 0$. Note that $\Delta^n(0) = 0$, and let

$$\tau_i^n = \inf\{t \mid \Delta^n(t) = \varepsilon\}.$$

and $E^n = \{\tau^n \le \theta\}$. To prove the lemma, it suffices to show that $P(E^n) \to 0$. Let us define on the event E^n

$$\sigma_i^n = \sup\{t \mid t < \tau_i^n, \Delta^n(t) = \varepsilon/2\}, \qquad \kappa_{\sigma,i}^n = \inf\{t \ge \sigma^n \mid Z_i^n(t) = 1\}.$$

Note that on E^n we always have $\sigma_i^n \in [0, \tau_i^n]$. The random variable κ_i^n represents the first time between σ_i^n and τ_i^n when server i is occupied. If this never happens within $[\sigma_i^n, \tau_i^n]$, we have, by definition, $\kappa_i^n = \infty$.

For simplicity, we usually will omit the symbols n and i from the notation of $\tau_i^n, \sigma_i^n, \kappa_{\sigma,i}^n$.

For $B \in \mathcal{F}$, we write $P_E(B) = P(E^n \cap B)$. The proof proceeds in four steps.

Step 1: We will show that for every $\delta > 0$,

$$P_E(\sqrt{n}(\kappa_\sigma \wedge \tau - \sigma) > \delta) \to 0 \quad \text{as} \quad n \to \infty \quad > \text{uniformly in } i \in K. \qquad (21)$$

Fixing δ, we will use the foregoing analysis concerning the event H, with $H = E \cap \{\kappa \wedge \tau - \sigma > \delta\}$. We take $\eta = \sigma$, $\zeta = \kappa \wedge \tau$ and $L = L(\eta)$. Under H, by the definition of κ, at any time within $[\sigma, \kappa \wedge \tau)$ server i is idle, and so by work conservation, no customer is in the queue.

Also, under the event E^n, on the interval $[\sigma, \kappa \wedge \tau)$, server i continues to accumulate idle time at rate 1. Since the servers in L cannot accumulate idle time at a faster rate, by definition of L and of our policy, L receives no jobs at any time during this interval. This shows that H satisfies both bullet conditions from the first part of the proof. Consequently (20) is valid.

Let R^n denote the sum of the first three terms on the l.h.s. of (20).

Given set of servers K_0, exist a Poisson process S_{K_0} with rate 1, such that

$$\sum_{j\in K_0} S_j \circ T_j(t) = S_{K_0}(\sum_{j\in K_0} T_j(t)).$$

By simple calculations (see also Lemma 3.2 in [3]),

$$\sum_{j\in K_0}\widehat{S}_j\circ\bar{T}_j(t)=\widehat{S}_{K_0}(\sum_{j\in K_0}\bar{T}_j(t))\le\sup_{0\le t\le\theta}\widehat{S}_{K_0}(\sum_{j\in K_0}\bar{T}_j(t))\le|\widehat{S}_{K_0}|^*_{\bar{\mu}\theta}$$

Therefore,

$$\sum_{j\in L^c}\widehat{S}_j\circ\bar{T}_j(t)\le\sup_{K_0}\sum_{j\in K_0}\widehat{S}_j\circ\bar{T}_j(t)\le\sup_{K_0}|\widehat{S}_{K_0}|^*_{\bar{\mu}\theta}=|\widehat{S}|^*_{\bar{\mu}\theta}$$

since S_{K_0} is a Poisson process with rate 1, $|\widehat{S}_{K_0}|^*_{\bar{\mu}\theta}$ doesn't depend on K_0. Hence

$$R^n\le 2\sup_{0\le t\le\theta}(\sum_{j\in L^c}\widehat{I}_j(t)+\widehat{A}(t)-\sum_{j\in L^c}\widehat{S}_j\circ\bar{T}_j(t))\le 2|\widehat{I}|^*_\theta+2|\widehat{A}|^*_\theta+2|\widehat{S}_{K_0}|^*_{\bar{\mu}\theta}$$

By Lemma 3.2, the random variables $R^n\mathbf{1}_{H^n}$ are tight.

Using (20) and the inequality $Z_j\le 1$, we have on H^n_i,

$$R^n+\Big(\frac{\lambda^n}{\sqrt{n}}-\frac{1}{\sqrt{n}}\sum_{j\in L^c}\mu_j\Big)(\zeta-\eta)\le 0.$$

Using the notation $\widehat{\beta}^n$ (16), and the inequality $\sum_{j\in L}\mu_j\ge\underline{\mu}|L|\ge\underline{\mu}pn$, on H^n_i

$$R^n+(\widehat{\beta}^n+\sqrt{n}p\underline{\mu})(\zeta-\eta)\le 0. \tag{22}$$

Since $\widehat{\beta}^n$ converge (cf. (4), (7)), R^n is tight and $p\underline{\mu}>0$, and these terms does not depend on i, it follows that $P_E((\zeta-\eta)>\delta)\to 0$ uniformly on K. By Appendix $R^n\to 0$ as $n\to\infty$ and by additional using of (22), it follows that $P_E(\sqrt{n}(\zeta-\eta)>\delta)\to 0$ uniformly on K, establishing (21).

Step 2: $P_E(\sqrt{n}(J_i(\tau)-J_i(\sigma))>\delta)\to 0$ as $n\to\infty$ uniformly on K.

On the event E^n, denote $\sigma_0=\sigma$, $\sigma_m=\inf\{s>\kappa_{\sigma_{m-1}}\mid I_i(s)=1\}$ where κ_{σ_k} is defined as above, so that $\sigma_1<\kappa_{\sigma_1}<\cdots<\sigma_l<\kappa_{\sigma_l}\wedge\tau\in[\kappa_\sigma,\tau]$.

Note that l is bounded by a random variable with Poisson distribution with rate $\theta\overline{\mu}$ that does not depend on i or on n.

Since $I_i(s) = 0$ for $s \in [\kappa_{\sigma_m}, \sigma_{m+1})$ and and $I_i(s) = 1$ for $s \in [\sigma_m, \kappa_{\sigma_m})$, we obtain

$$\begin{aligned} P(\sqrt{n}(J_i(\tau) - J_i(\sigma)) > \delta) &= P\left(\sum_{m=0}^{l} \sqrt{n} \int_{\sigma_m}^{\kappa_{\sigma_m}} I_i(s)ds > \delta\right) \\ &= P\left(\sum_{m=0}^{l} \sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta\right) . \end{aligned}$$

Now given $\gamma > 0$ fix ℓ_γ so that $P(\ell \geq \ell_\gamma) \leq \gamma/2$. Then

$$\begin{aligned} & P\left(\sum_{m=0}^{l} \sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta\right) \\ &= P\left(\sum_{m=0}^{l} \sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta,\ \ell \geq \ell_\gamma\right) + P\left(\sum_{m=0}^{l} \sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta,\ \ell < \ell_\gamma\right) \\ &\leq P\left(\ell \geq \ell_\gamma\right) + P\left(\sum_{m=0}^{l_\gamma} \sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta\right) \\ &\leq \gamma/2 + \sum_{m=0}^{l_\gamma} P\left(\sqrt{n}(\kappa_{\sigma_m} - \sigma_m) > \delta/l_\gamma\right) . \end{aligned}$$

From step 1 applied to each of the terms of the r.h.s sum (with $L = L(\sigma_m)$) and the arbitrariness of γ it follows that $P_E(\sqrt{n}(J_i(\tau) - J_i(\sigma)) > \delta) \to 0$ as $n \to \infty$ uniformly on K.

Step 3: Since on $[\sigma, \tau]$ we have $J_i(t) > \widetilde{J}^p(t)$ we get

$$\begin{aligned} P(E) &= P_E(\Delta^n(\tau) - \Delta^n(\sigma) = \epsilon/2) \\ &= P_E(\sqrt{n}[J_i(\tau) - \widetilde{J}^p(\tau)] - \sqrt{n}[J_i(\sigma) - \widetilde{J}^p(\sigma)] = \frac{\epsilon}{2}) \\ &= P_E(\sqrt{n}[J_i(\tau) - J_i(\sigma)] + \sqrt{n}[\widetilde{J}^p(\sigma) - \widetilde{J}^p(\tau)] = \frac{\epsilon}{2}) \end{aligned}$$

Since $\forall \delta > 0$, $P_E(\sqrt{n}(J_i(\tau) - J_i(\sigma)) > \delta) \to 0$ as $n \to \infty$ uniformly on K and $\sqrt{n}(\widetilde{J}^p(\sigma) - \widetilde{J}^p(\tau)) \leq 0$ because $\widetilde{J}^p(t)$ is a increasing process, it follows that $|\Delta^n|^*_\theta \to 0$ in probability as $n \to \infty$ uniformly on K.

Corollary 3.5 *Let $\Delta^n_2(t) = \sqrt{n}|\widetilde{J}^{1-p}(t) - \widetilde{J}^p(t)|$. Then $|\Delta^n_2|^*_\theta \to 0$ in probability as $n \to \infty$.*

Proof: Given $0 < p < \frac{1}{2}$, it easy to show

$$B_i(t) := \sqrt{n}\{|(J_i(t) - \widetilde{J}^p(t))^+ - (J_i(t) - \widetilde{J}^{1-p}(t))^+|\}$$
$$= \sqrt{n} \begin{cases} |\widetilde{J}^{1-p}(t) - \widetilde{J}^p(t)|, & \widetilde{J}^{1-p}(t) < J_i(t) \\ J_i(t) - \widetilde{J}^p(t), & \widetilde{J}^p(t) < J_i(t) < \widetilde{J}^{1-p}(t) \\ 0, & \widetilde{J}^p(t) > J_i(t) \end{cases} \quad (23)$$

For $t \le \theta$, $B_i(t) \le |\Delta^n(p)|^*_\theta + |\Delta^n(1-p)|^*_\theta \le 2|\Delta^n(p)|^*_\theta$ where $\Delta^n(p) = \sqrt{n}(J_i(t) - \widetilde{J}^p(t))^+$.

Using the equality (23) and the inequality $B_i(t) \le 2|\Delta^n(p)|^*_\theta$,

$$\sqrt{n}|\widetilde{J}^{1-p}(t) - \widetilde{J}^p(t)| = \frac{[\sum_{i=1}^n B_i(t)] - \sum_{i=1}^n \sqrt{n}(J_i(t) - \widetilde{J}^p(t))1_{\{\widetilde{J}^p(t) < J_i(t) < \widetilde{J}^{1-p}(t)\}}}{p \cdot n}$$
$$\le \frac{\sum_{i=1}^n B_i(t)}{p \cdot n} \le \frac{\sum_{i=1}^n 2|\Delta^n(p)|^*_\theta}{p \cdot n}.$$

Let us $\gamma, \delta_0 > 0$, by Lemma 3.4 for all sufficiently large n, with probability of at least $1 - \gamma$ and $\delta = p\delta_0/2$, one has

$$|\Delta^n_2|^*_\theta \le \frac{\sum_{i=1}^n 2|\Delta^n(p)|^*_\theta}{p \cdot n} \le \frac{\sum_{i=1}^n 2\delta}{p \cdot n} \le \frac{2\delta}{p} = \delta_0.$$

Lemma 3.6 *For every $\eta, \varepsilon > 0$, there is $t_1 > 0$ such that for all $t > t_1$,*

$$\liminf_n P(\widehat{J}^n(t) \ge \eta) \ge 1 - \varepsilon.$$

Proof: See [3].

Lemma 3.7 *For every $\eta, \varepsilon > 0$, there is $t_1 > 0$ such that for all $t > t_1$,*

$$\liminf_n P(\sqrt{n}\widetilde{J}^p(t) \ge \eta) \ge 1 - \varepsilon.$$

Proof: Clearly $\widehat{J}^n(t) = \frac{1}{n}\sum_{i=1}^n \sqrt{n}J_i(t) = \frac{1}{n}\sum_{i=1}^n \sqrt{n}(J_i(t) - \widetilde{J}^p(t)) + \sqrt{n}\widetilde{J}^p(t)$.

Let us $\eta_0, \varepsilon > 0$ be given. By Lemma 3.6, for every η, there is $t_1 > 0$ such that for all $t > t_1$, with probability at least $1 - \varepsilon/2$, for all sufficiently large n,

$$\eta \le \widehat{J}^n(t) \le \frac{1}{n}\sum_{i=1}^n \sqrt{n}(J_i(t) - \widetilde{J}^p(t))^+ + \sqrt{n}\widetilde{J}^p(t).$$

Let θ be so large that $t_1 < t < \theta$, by Lemma 3.4 with probability at least $1 - \varepsilon/2$, for all sufficiently large n,

$$|\Delta^n|^*_\theta \leq \eta/2$$

Hence for all sufficiently large n, with probability of at least $1 - \varepsilon$, one has

$$\eta/2 < \eta - \frac{1}{n}\sum_{i=1}^{n} |\Delta^n|^*_\theta \leq \sqrt{n}\widetilde{J}^p(t).$$

Lemma 3.7 follows upon setting $\eta = 2\eta_0$.

Proof of Theorem 2.1 Let $\varepsilon > 0$ and $t \geq 0$ be given. Tightness of $\gamma^n(\varepsilon)$ follows from Lemma 3.7. Let θ be so large that $\gamma^n = \gamma^n(\varepsilon) < \theta - t$ with probability at least $1-\varepsilon/2$, for all sufficiently large n. Fix $\theta_1 \geq \theta$. Let $\delta > 0$ be given, and consider the event $|\sqrt{n}(\widetilde{J}^{1-p}(t) - \widetilde{J}^p(t))|^*_{\theta_1} \leq \delta$. By Corollary 3.5, this event has probability at least $1 - \varepsilon/2$, for all sufficiently large n. Hence for all sufficiently large n, with probability of at least $1 - \varepsilon$, one has

$$\gamma^n + t < \theta \quad \text{and} \quad \Big|\frac{\widetilde{J}^{1-p}(\gamma^n + t)}{\widetilde{J}^p(\gamma^n + t)} - 1\Big| \leq \frac{\delta}{\sqrt{n}\widetilde{J}^p(\gamma^n + t)} \leq \frac{\delta}{\varepsilon},$$

where we used the fact $\sqrt{n}X_p(\gamma) = \varepsilon$ and that the process $\widetilde{J}^p$ is nondecreasing. Theorem 2.1 follows upon setting $\delta = \varepsilon^2$.

To prove Theorem 2.2, we take $t = 0$ in the above argument, and set $T = \theta$. Then we fix some $T_1 \geq T$ and set $\theta_1 = T_1$. On the event analyzed in the previous paragraph, for all sufficiently large n, with probability at least $1 - \varepsilon$, we obtain

$$T \geq \gamma^n \text{ (hence } \widetilde{J}^p(T) \geq \varepsilon) \quad \text{and} \quad \sup_{s\in[T,T_1]} \Big|\frac{\widetilde{J}^{1-p}(s)}{\widetilde{J}^p(s)} - 1\Big| \leq \frac{\delta}{\widetilde{J}^p(T)} \leq \frac{\delta}{\varepsilon} = \varepsilon.$$

■

4 Appendix

In this appendix, we complete some proofs about the convergence of difference of $\widehat{A}$, $\widehat{S}$ and $\widehat{I}$ that contain some random elements.

Lemma 4.1 *$\widehat{A}[\eta, \zeta] \to 0$ as $n \to \infty$.*

Proof: By Lemma 3.2, $\widehat{A}$ is C-tight so, $\forall\varepsilon, \exists\delta_0$ such that for all sufficiently large n, $w_\theta(\widehat{A}, \delta_0) < \varepsilon$. Since $\zeta - \eta \to 0$, for all sufficiently large n $\zeta - \eta < \delta_0$. Hence, $|\widehat{A}[\eta, \zeta]| \leq w_\theta(\widehat{A}, \zeta - \eta) < w_\theta(\widehat{A}, \delta_0) < \varepsilon$.

Lemma 4.2 $\sum_{j \in L^c} \widehat{S}_j \circ \bar{T}_j[\eta, \zeta] \to 0$ *as* $n \to \infty$.

Proof: For all $K_0 \subseteq K$, exist a Poisson process with rate 1, S_{K_0} such that

$$\sum_{j \in K_0} \widehat{S}_j \circ \bar{T}_j(t) = \widehat{S}_{K_0} \circ \sum_{j \in K_0} \bar{T}_j(t) \tag{24}$$

We prove that $\sum_{j \in K_0} \widehat{S}_j \circ \bar{T}_j[\eta, \zeta]$ uniformity converges to 0 as $n \to \infty$ in regard to K_0 i.e. $\forall\varepsilon > 0$ and $\exists n_0$ such that for all $n > n_0$ and for all set $K_0 \subseteq K^n$, $\widehat{S}_{K_0} \circ \sum_{j \in K_0} \bar{T}_j[\eta, \zeta] < \varepsilon$.

Since $\forall K_0, S_{K_0}$ is a Poisson process with rate 1, $\widehat{S}_{K_0}$ is C-tight and doesn't depend on K_0. Therefore $\forall\varepsilon > 0, \exists\delta_0$ such that for all sufficiently large n, $\forall K_0, w_\theta(\widehat{S}_{K_0}, \delta_0) < \varepsilon$.

By [3] (Lemma 3.2's proof) $\sum_{j \in K_0} \bar{T}_j(t)$ converges in distribution to $M_{K_0} t$ where M_{K_0} is constant and $0 < M_{K_0} < \bar{\mu}$. Since $\zeta - \eta \to 0$ for all sufficiently large n $\forall K_0$, $\sum_{j \in K_0} \bar{T}_j[\eta, \zeta] < \bar{\mu}(\zeta - \eta) < \delta_0$.

Combining the above two results, $\forall\varepsilon > 0$ for all sufficiently large n,

$$\sum_{j \in L^c} \widehat{S}_j \circ \bar{T}_j[\eta, \zeta] \leq \sup_{K_0} \sum_{j \in K_0} \widehat{S}_j \circ \bar{T}_j[\eta, \zeta] = \sup_{K_0} \widehat{S}_{K_0} \circ \sum_{j \in K_0} \bar{T}_j[\eta, \zeta] < \sup_{K_0} w_\theta(\widehat{S}_{K_0}, \delta_0) < \varepsilon$$

Lemma 4.3 $\sum_{j \in L^c} \widehat{I}_j[\eta, \zeta] \to 0$ *as* $n \to \infty$.

Proof: By definition of the processes X, Q and Z_j, we have $X = Q + \sum_{j \in L} Z_j$. Since Q vanishes on the interval $[\eta, \zeta]$, we have

$$X[\eta, \zeta] = \sum_{j \in L^c} Z_j[\eta, \zeta].$$

Using $Z_j + I_j = N_j$, $j \in L^c$, we have

$$X[\eta, \zeta] = -\sum_{j \in L^c} I_j[\eta, \zeta].$$

By Lemma 3.3 $\widehat{X}$ is C-tight so, in similar to Lemma 4.1 $\widehat{X}[\eta,\zeta] \to 0$ as $n \to 0$, and thus so $\sum_{j \in L^c} I_j[\eta,\zeta] \to 0$ as $n \to 0$.

References

[1] Armony, M., and Ward, A.R., (2008) *Fair Dynamic Routing in Large-scale Heterogeneous-Server Systems*, preprint.

[2] Atar, R., *Central limit theorem for a many-server queue with random service rates*, Ann. Appl. Probab., to appear.

[3] Atar, R., Shaki, Y.Y., Shwartz, A., *A blind policy for equalizing cumulative idleness*, submitted.

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

[5] Halfin, S., and Whitt, W., (1981) *Heavy-traffic limits for queues with many exponential servers*, Oper. Res. 29, no. 3, 567-588.

[6] Hale, J.K., (1980) *Ordinary differential equations*, Robert E. Krieger publishing Company, Huntington, New York.