Efficient routing in heavy traffic under partial sampling of service times

Rami Atar and Adam Shwartz
Department of Electrical Engineering
Technion–Israel Institute of Technology
{atar}{adam}@ee.technion.ac.il
http://www.ee.technion.ac.il/people/{adam}{atar}

March 14, 2008

Abstract

We consider a queue with renewal arrivals and n exponential servers in the Halfin-Whitt heavy traffic regime, where n and the arrival rate increase without bound, so that a critical loading condition holds. Server k serves at rate μ_k, and the empirical distribution of $\{\mu_k\}_{k=1,\ldots,n}$ is assumed to converge weakly. We show that very little information on the service rates is required for a routing mechanism to perform well. More precisely, we construct a routing mechanism that has access to a *single sample* from the service time distribution of each of $n^{\frac{1}{2}+\varepsilon}$ randomly selected servers ($\varepsilon > 0$), but not to the actual values of the service rates, the performance of which is asymptotically as good as the best among mechanisms that have the complete information $\{\mu_k\}_{k=1,\ldots,n}$.

Keywords: Halfin-Whitt regime; routing policies; service time sampling
MSC2000: Primary: 60F17. Secondary: 68M20, 90B15, 90B22, 60K30, 60K25.

1 Introduction

In the many-server parametric regime of Halfin and Whitt [10], a critically loaded diffusively scaled system has the property that the fraction of time when queues are empty is neither close to 0 nor 1, a situation that is often observed in applications. Particularly, it has been suggested that this regime is suitable for modeling large call centers [8], and various models motivated by this application have been studied, where a many-server system operates in this regime (see [16] for a review). In models that involve heterogenous servers, a principal problem is to find an efficient routing policy [1, 2, 4, 5, 6, 9, 14, 15]. In all previous works on routing control in this regime, the proposed routing mechanisms are assumed to have complete information about the service rates of each server (where by 'rate' we refer to the parameter of the exponential service time distribution, assumed by most authors; however see [14] for more general service times). Since often in applications the routing control mechanism has little knowledge of the performance of each individual server, it is natural to ask whether it can perform near optimality with less information on these parameters. Our goal in this note is to argue that sufficient information for this purpose is a single sample of service time from a negligible fraction of the servers.

The pioneering work of Halfin and Whitt [10] considers a queue with renewal arrivals and identical exponential servers, where the number of servers and the rate of arrivals are scaled up so that the queue remains critically loaded. The second order asymptotics of the process representing the number of customers in the system is shown to converge to a diffusion. When the servers are heterogenous, it was shown in [1] that, in presence of customers of a single class, the policy which routes jobs to the fastest server among those that are free at the time of routing (and does not allow interruption of service) is asymptotically optimal in terms of the queue length as well as the delay of an arriving customer. Analogous results are available for the case of random, i.i.d. service rates [3] and, under appropriate assumptions, for hyperexponential service times [14].

We mention that works that characterize the fluid and diffusion scaling limits are available for homogeneous servers with general service time distributions [11, 13]. The question that we address here is also very natural in this wider context. Note, however, that for heterogenous servers with general service times, an asymptotically optimal routing policy is not known even when the routing mechanism has access to all service time distributions (with the exception of [14]). For this reason we confine our treatment to the exponential case.

As mentioned above, we assume that the routing mechanism has access only to *samples* from the service time distribution of some of the servers. We show that, perhaps counter-intuitively, very little sampling is required for asymptotically optimal performance: It suffices to collect a *single* sample from each server in a set of r randomly selected servers, where r is as small as $n^{\frac{1}{2}+\varepsilon}$ ($\varepsilon > 0$). The proposed policy always routes jobs to non-sampled servers if such are available, and otherwise, routes to the server for which the sampled service time is smallest among the (sampled) servers that are available at the time. It is shown to be asymptotically optimal in the sense that the diffusion limit of the process representing the total number of customers in the system (characterized in Theorem 2.1) is stochastically dominated by any subsequential limit under any (work conserving, nonanticipating) policy (see Theorem 2.2). This includes policies that have access to the complete information on service rates. A similar statement holds for the queue length processes (simply by (13)).

A clear practical advantage of our approach is that it is not necessary to invest in measuring various characteristics precisely, or collect accurate information on the performance of the servers. In addition, the policy proposed has a desired robustness property in that its performance is nearly optimal regardless of the values of system parameters, as long as the basic assumptions hold. These assumptions on the empirical measure of the rates and its first and second-order limits (1)–(3) are quite general, and so are the assumptions on the limiting distribution.

An intuitive explanation of the result is as follows. As known from [1] and [3], asymptotically optimal performance is achieved by policies under which, at every moment where some servers are idle, most of these servers are the slowest (i.e., ones that have rate very close to the quantity μ_*, defined in the second paragraph of Section 2). Note that because the service is non-interruptible, it is not immediate that this property is attained by routing mechanisms that prioritize servers according to their service rates; however, this claim is proved in the above citations. Because the scaling is diffusive and the system is critically loaded, it can be shown that the maximum number of servers that are idle over a given finite time interval is of the order of magnitude of $n^{\frac{1}{2}}$. Thus, roughly speaking, for a routing policy to perform near optimality, it suffices that it has access to the service rates of $n^{\frac{1}{2}+\varepsilon}$ servers with rate within $(\mu_* - \delta, \mu_* + \delta)$, so that it can assign them the lowest priority, and achieve asymptotic optimality just as if it had access to all service rates. As we will show, a certain tail property of a collection of $r = n^{\frac{1}{2}+\varepsilon}$ independent exponential random

variables implies that ordering these r quantities according to (the reciprocal to) a single sample from each, rather than their rate, results in a negligible error in determining which of these servers have low rates. As a consequence, relying on samples rather than exact rates does not degrade the performance (see Lemma 3.1 for a precise statement).

As an example, one might compare the proposed policy with one in which the r selected servers are ranked highest by the routing mechanism. The result would be that most of the time the selected servers would be busy, while the $O(n^{\frac{1}{2}})$ servers that are idle would have rates with arbitrary values, and the system would operate far from optimality.

The situation described here is reminiscent of a phenomenon discovered over the last decade, sometimes called 'the power of choice', which refers to the following setting. Parallel stations are available to serve a stream of customers, and each customer is given the possibility to choose between two (or a larger fixed number of) randomly selected stations, where to be queued and eventually served. The choice results in a dramatic improvement of load balancing with respect to that achieved under random routing (see [12] for a review of various related results). The problem studied in the current paper is, of course, different in many respects (a single queue, heterogenous servers, centralized routing, heavy traffic, diffusion scale, and more), but it is interesting to notice that it does share with the phenomenon alluded to above the property that information about a small random subset of servers can gain much in performance. In fact, in our case it suffices to obtain optimality.

The proof of the main result is based on an estimate on the number of errors in ordering the servers according to their sampled data (Lemma 3.1), an estimate on the total idle time encountered by servers that have relatively high priority (Lemma 3.2), and the technique developed in [3] (proof of Theorem 2.1).

In the next section we describe the model and the proposed policy, and state the main results. The proofs appear in Section 3.

2 Model and main results

We fix some notation. Denote by $\mathbb{D}$ the space of functions from $\mathbb{R}_+$ to $\mathbb{R}$ that are right continuous on $\mathbb{R}_+$ and have finite left limits on $(0,\infty)$ (RCLL), endowed with the usual Skorohod topology [7]. If X^n, $n \in \mathbb{N}$ and X are processes with sample paths in $\mathbb{D}$ (respectively, real-valued random variables) we write $X^n \Rightarrow X$ to denote weak convergence of the measures induced by X^n on $\mathbb{D}$ (respectively, on $\mathbb{R}$) to the measure induced by X, as $n \to \infty$. For $X \in \mathbb{D}$ we write $|X|_{*,t} := \sup_{0\le s\le t}|X(s)|$. For $x \in \mathbb{R}$, write $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$.

A complete probability space $(\Omega, \mathcal{F}, P)$ is given, supporting all random variables and stochastic processes defined below. Expectation w.r.t. P is denoted by E. We consider a single queue fed by renewal arrivals, with parallel exponential servers. The model is parameterized by $n \in \mathbb{N}$, where n also represents the number of servers. The n servers are labeled as $1, \dots, n$, and, for the nth system, deterministic parameters $\mu_k^n \in [\underline{\mu}, \bar{\mu}]$ are given, where μ_k^n represents service rate of server k, and $0 < \underline{\mu} \le \bar{\mu} < \infty$ are constants independent of n. We assume weak convergence of the empirical measure of $\{\mu_k^n\}$,

$$L^n = n^{-1} \sum_k \delta_{\mu_k^n} \to m, \tag{1}$$

where m is a probability measure on $\mathbb{R}$ (supported on $[\underline{\mu}, \bar{\mu}]$). The mean is denoted by $\mu = \int x dm$.

A second order type approximation is further assumed on the rate parameters, namely that the limit

$$\lim_n n^{-\frac{1}{2}} \sum_{k=1}^{n} (\mu_k^n - \mu) := \widehat{\mu} \tag{2}$$

exists as a finite number. Denoting $\mu_* = \operatorname{ess\,inf} m$, we finally assume

$$\lim_{n\to\infty} \#\{k : \mu_k^n < \mu_* - \varepsilon\} n^{-\frac{1}{2}} = 0, \quad \varepsilon > 0. \tag{3}$$

Example 2.1 A special case of assumptions (1), (2) and (3) is when there is a fixed number of pools of servers with $a_i n + O(1)$ servers at pool i, and where each server at pool i serves at rate $b_i + c_i n^{-\frac{1}{2}}$ (for constant a_i, b_i, c_i; $a_i > 0$), a setting that is common (for example [1] in a single class setting, and [15] in a multiclass setting).

Example 2.2 We point out that there is more flexibility in the choice of the parameters. For example, if we have two pools of size $0.2n + n^{\frac{3}{4}}$ and $0.8n + n^{\frac{4}{5}}$ with rates $1 + 4n^{-\frac{1}{6}} + n^{-\frac{1}{2}}$ and, respectively, $2 - n^{-\frac{1}{6}}$, then our assumptions still hold. A more general case is as follows. We have a fixed number of pools of sizes $a_i n + f_i(n)$, with respective rates $b_i + c_i n^{-\frac{1}{2}} + g_i(n)$. Then assumptions (1)–(3) hold provided that $f_i(n) = o(n)$, $g_i(n) = o(1)$ and that the limit

$$\lim_{n\to\infty} n^{\frac{1}{2}} \sum_i a_i g_i(n)$$

exists. This is verified by a straightforward, if lengthy calculation using $\mu = \sum_i a_i b_i$.

Example 2.3 It is sometimes very natural to regard the rates $\{\mu_k\}$ as random variables, and thus to consider the queueing process, as well as its scaling limit, as processes in random environment. The case where the service rates are i.i.d. random variables, drawn from a common distribution m, was considered in [3]. In this case, the law of large numbers implies that (1) and (3) hold with probability one, and the central limit theorem implies a variation of (2), in which $\widehat{\mu}$ is a normal random variable. Although we assume throughout that the service rates are deterministic, we would like to comment that all our results can be formulated for an i.i.d. random environment, with basically the same proofs.

The initial configuration is now described. Let Q_0^n be a $\mathbb{Z}_+$-valued random variable, representing the initial number of customers in the buffer. Let $B_{k,0}^n$, $k = 1, \ldots, n$ be $\{0,1\}$-valued random variables representing the initial state of each server, where $B_{k,0}^n = 1$ if and only if server k initially serves a customer. We restrict to non-idling policies, so that in particular $Q_0^n > 0$ only if $B_{k,0}^n = 1$ for all $k = 1, \ldots, n$. The total number of customers initially in the system is denoted by $X_0^n = Q_0^n + \sum_{k=1}^n B_{k,0}^n$. Note that, by assumption, we have the relation $Q_0^n = (X_0^n - n)^+$. We assume

$$\widehat{X}_0^n := n^{-\frac{1}{2}}(X_0^n - n) \Rightarrow \xi_0, \tag{4}$$

where ξ_0 is a random variable.

To define the arrival process, we are given parameters $\lambda^n > 0$, $n \in \mathbb{N}$ satisfying $\lim_n \lambda^n/n = \lambda > 0$, and a sequence of strictly positive i.i.d. random variables $\{\check{U}(l), l \in \mathbb{N}\}$, with mean $E\check{U}(1) = 1$ and variance $\check{C}^2 = \mathrm{Var}(\check{U}(1)) \in [0, \infty)$. With $\sum_1^0 = 0$, the number of arrivals up to

time t for the nth system is given by $A^n(t) = \sup\{l \geq 0 : \sum_{i=1}^{l} \check{U}(i)/\lambda^n \leq t\}$. The arrival rates are further assumed to satisfy the second order relation

$$\lim_n n^{-\frac{1}{2}}(\lambda^n - n\lambda) = \widehat{\lambda}, \tag{5}$$

for some $\widehat{\lambda} \in \mathbb{R}$. The 'heavy traffic' condition on the first order parameters is assumed, namely

$$\lambda = \mu, \tag{6}$$

indicating that the system is critically loaded. For each $k = 1, \ldots, n$, we let B_k^n be a stochastic process taking values in $\{0, 1\}$, representing the status of server k: when $B_k^n(t) = 1$ [resp., 0] we say that server k is busy [resp., idle]. Let $I_k^n(t) = 1 - B_k^n(t)$ for $k = 1, \ldots, n$, and $t \geq 0$. For $k = 1, \ldots, n$, let R_k^n [resp., D_k^n] be a $\mathbb{Z}_+$-valued process with nondecreasing right-continuous sample paths, representing the number of routings of customers to server k within $[0, t]$ [resp., the number of jobs completed by server k by time t]. Thus

$$B_k^n(t) = B_{k,0}^n + R_k^n(t) - D_k^n(t), \quad k = 1, \ldots, n, \ t \geq 0. \tag{7}$$

To describe the processes D_k^n, let $\{S_k, k \in \mathbb{N}\}$ be i.i.d. rate-1 Poisson processes, each having right-continuous sample paths. The processes D_k^n are assumed to satisfy

$$D_k^n(t) = S_k(T_k^n(t)), \quad k = 1, \ldots, n, \tag{8}$$

where

$$T_k^n(t) = \mu_k^n \int_0^t B_k^n(s)ds, \quad k = 1, \ldots, n. \tag{9}$$

Let X^n, Q^n and I^n be defined as

$$X^n(t) = X_0^n + A^n(t) - \sum_{k=1}^{n} D_k^n(t), \quad Q^n(t) = Q_0^n + A^n(t) - \sum_{k=1}^{n} R_k^n(t), \quad I^n(t) = \sum_{k=1}^{n} I_k^n(t). \tag{10}$$

These processes represent the number of customers in the system, the number of customers in the buffer and, respectively, the number of servers that are idle.

The routing policy, that will be described below, does not have access to the service rates μ_k^n, but it has access to samples from the service time of r of the servers, selected at random, and no information at all on service rates of the others. More precisely, let $r = r_n \in \mathbb{N}$, $r \leq n$ be given, and let $\Sigma = \Sigma^n$ be a random variable uniformly distributed over the set of all subsets of $\{1, \ldots, n\}$ that have cardinality r. We denote $\Sigma^c = \{1, \ldots, n\} \setminus \Sigma$. For each $k \in \Sigma$, let $\sigma_k = \sigma_k^n$ be an independent copy drawn from the service time distribution of server k. That is, σ_k is an exponential random variable with parameter μ_k^n and, conditioned on Σ, $\{\sigma_k\}_{k\in\Sigma}$ are independent. We choose

$$r_n = [n^{\beta_0}], \tag{11}$$

where $\beta_0 \in (\frac{1}{2}, 1]$. Denote $\widehat{\mu}_k = 1/\sigma_k$, $k \in \Sigma$.

The four stochastic primitives introduced, as listed below, are assumed to be mutually independent, for each n:

$$\left(X_0^n, \{B_{k,0}^n\}_{k=1,\ldots,n}\right), \quad \{S_k\}_{k\in\mathbb{N}}, \quad A^n, \quad \left(\Sigma^n, \{\sigma_k^n\}_{k\in\Sigma^n}\right). \tag{12}$$

Routing is based on an ordering of the servers according to whether they are in Σ and, within Σ, according to the value of $\widehat{\mu}_k$. A permutation $\text{Rank} = \text{Rank}_n$ of $\{1,\ldots,n\}$ is defined as follows. On the probability-one event that the $\widehat{\mu}_k$ are all distinct, the set Σ is mapped by Rank onto $\{1,\ldots,r\}$ (and Σ^c onto $\{r+1,\ldots,n\}$). For $k,l \in \Sigma$, $\text{Rank}(k) < \text{Rank}(l)$ if and only if $\widehat{\mu}_k < \widehat{\mu}_l$. For $k,l \in \Sigma^c$, $\text{Rank}(k) < \text{Rank}(l)$ if and only if $k < l$.

The routing policy favors servers ranked higher (namely those that have high value under the map Rank). That is, when a customer arrives to the system to find more than one idle server, the customer is routed to the server with highest rank among those servers. Since it is assumed that the routing policy is work conserving (non-idling), when the queue is nonempty and a server has just finished serving, a customer (from the head of the line) is routed to this server, and when a customer arrives to the system to find exactly one server that is idle, it is instantaneously routed to that server. As a result,

$$Q^n(t) = (X^n(t) - n)^+, \quad I^n(t) = (X^n(t) - n)^- \tag{13}$$

holds for all t. Also, service is non-interruptible, in the sense that a customer completes service at the server it is first assigned.

This completes the description of the process

$$\Pi_0^n := (\{B_k^n\}, \{R_k^n\}, \{D_k^n\}, X^n, Q^n, I^n).$$

It can be seen that this description uniquely determines Π_0^n. We sometimes refer to this process as *policy* Π_0^n. Later we use some of the symbols above (such as X^n) to denote quantities that have the same meaning (such as the number of customers in the nth system) under a different routing policy Π^n. To avoid confusion, we therefore make specific reference to policy Π_0^n when necessary.

Finally, we make a simplifying assumption about the initial occupation of servers, namely that only servers that are ranked low may initially be idle:

$$B_{k,0}^n = 1_{\{\text{Rank}(k) > I_0^n\}}, \tag{14}$$

where

$$I_0^n = (X_0^n - n)^- \tag{15}$$

is the initial number of idle servers.

Let $\widehat{X}^n$ be a centered, normalized version of the process X^n, defined by

$$\widehat{X}^n = n^{-\frac{1}{2}}(X^n - n). \tag{16}$$

Our main result is the following.

Theorem 2.1 *Under policy Π_0^n, the processes $\widehat{X}^n$ converge weakly to the unique solution ξ of*

$$\xi(t) = \xi_0 + \sigma w(t) + (\widehat{\lambda} - \widehat{\mu})\, t + \mu_* \int_0^t \xi(s)^- ds, \quad t \geq 0, \tag{17}$$

where $\sigma^2 = \mu(\check{C}^2 + 1)$ and w is a standard Brownian motion, independent of ξ_0.

The result above is to be compared with Proposition 4.2 of [1] and Theorem 2.2 of [3] (for the case of a finite number of server pools and, respectively, random environment). In these references, equation (17) arises in the limit under a policy defined similarly to Π_0, but where the servers are ordered according to the actual values of μ_k, $k = 1, \ldots, n$. In [1] it is further shown that this policy asymptotically achieves the best performance in a large class of routing policies. Because our setting is different from [1], we will state and prove an analogous result, so as to exhibit that Π_0 is asymptotically optimal.

Toward this end, let us first comment on an alternative representation of the departure process. By (8), this process is given as $\sum_{k=1}^n D_k^n(t) = \sum_{k=1}^n S_k(T_k^n(t))$, where S_k are independent rate-1 Poisson processes. In fact, the departure process can also be represented as

$$\sum_{k=1}^n D_k^n(t) = S^n\Big(\sum_{k=1}^n T_k^n(t)\Big), \tag{18}$$

where, for every n, S^n is a rate-1 Poisson process, independent of the remaining primitive data, that is, of the first, third and fourth items of (12). This statement (along with a variation of it, stated in Section 3) is due to a standard superposition argument for Poisson processes, for which the reader is referred to Proposition 3.1 of [3].

We now define a class of policies by keeping the description of this section but abandoning the specifics of the routing mechanism. More precisely, we write $\Pi^n \in \mathcal{P}^n$ for any process

$$\Pi^n = (\{B_k^n\}, \{R_k^n\}, \{D_k^n\}, X^n, Q^n, I^n)$$

satisfying all relations stated throughout this section, from its beginning to the statement of Theorem 2.1, save the two paragraphs following display (12), and satisfying, in addition, work conservation (13) and the representation (18), for some rate-1 Poisson processes S^n, independent of the remaining primitive data. Note, in particular, that the routing mechanism may have access to $\{\mu_k\}$. See Remark 3.2 about the role played by the work conservation condition (13). We refer to any element of $\mathcal{P}^n$ as a *policy*.

Theorem 2.2 *For $n \in \mathbb{N}$ and any policy $\Pi^n \in \mathcal{P}^n$, let $\widehat{X}^n$ be the normalized version* (16) *of the corresponding process X^n. Then there exist processes Ξ^n that converge weakly, as $n \to \infty$, to the solution ξ to* (17) *and*

$$\widehat{X}^n(t) \geq \Xi^n(t), \quad t \geq 0, \ P\text{-a.s.}, \ n \in \mathbb{N}.$$

Since by Theorem 2.1, ξ is obtained as the limit under Π_0^n, the result above demonstrates that Π_0^n asymptotically optimal.

3 Proofs

We begin with the following.

Lemma 3.1 *Let $0 < \phi < \psi < \infty$, $\beta > \frac{1}{2}$, $c_1 > 0$ and $c_2 > 0$ be given constants. For $n \in \mathbb{N}$ denote $\ell^1 = \ell_n^1 = [c_1 n^\beta]$, $\ell^2 = \ell_n^2 = [c_2 n^\beta]$, and let $\phi_1^n, \ldots, \phi_\ell^n$ and $\psi_1^n, \ldots, \psi_\ell^n$ be positive real numbers with*

$$\sup_{n,i} \phi_i^n \leq \phi < \psi \leq \inf_{n,i} \psi_i^n.$$

For $n \in \mathbb{N}$ and $i \in \{1, \ldots, \ell\}$ let Φ_i^n [resp., Ψ_i^n] be an exponential random variable with parameter ϕ_i^n [resp., ψ_i^n]. For each n assume that $\{\Phi_i^n\}$ are mutually independent and that so are $\{\Psi_i^n\}$. Then there exist $\gamma > 0$ and $\kappa > 0$ such that, with $\theta_n = \gamma \log n$, one has

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{\ell_n^1} \mathbf{1}_{\{\Phi_i^n \geq \theta_n\}} \leq n^{\frac{1}{2}+\kappa}\right) = 0, \tag{19}$$

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{\ell_n^2} \mathbf{1}_{\{\Psi_i^n \geq \theta_n\}} \geq n^{\frac{1}{2}-\kappa}\right) = 0. \tag{20}$$

Proof. By stochastic monotonicity of the exponential random variable with respect to its parameter, it clearly suffices to prove the claim for the case $\phi_i^n = \phi$, $\psi_i^n = \psi$, all n and i. To prove the claim under this assumption, let κ and γ be strictly positive constants satisfying

$$\phi\gamma < \beta - \frac{1}{2} - \kappa < \beta - \frac{1}{2} + \kappa < \psi\gamma. \tag{21}$$

Write Φ_i [resp. Ψ_i] for Φ_i^n [resp., Ψ_i^n]. Then for any $\alpha > 0$ and $\{\theta_n\}$,

$$\begin{aligned}
P\left(\sum_{i=1}^{\ell_n^1} \mathbf{1}_{\{\Phi_i \geq \theta_n\}} \leq n^{\frac{1}{2}+\kappa}\right) &\leq e^{\alpha n^{\frac{1}{2}+\kappa}} E \exp\left[-\alpha \sum_{i=1}^{\ell_n^1} \mathbf{1}_{\{\Phi_i \geq \theta_n\}}\right] \\
&= e^{\alpha n^{\frac{1}{2}+\kappa}} \left(P(\Phi_1 < \theta_n) + e^{-\alpha} P(\Phi_1 \geq \theta_n)\right)^{\ell_n^1} \\
&= e^{\alpha n^{\frac{1}{2}+\kappa}} \left(1 - e^{-\phi\theta_n} + e^{-\alpha} e^{-\phi\theta_n}\right)^{\ell_n^1} \\
&= e^{\alpha n^{\frac{1}{2}+\kappa}} \exp\left\{\ell_n^1 \log\left(1 + e^{-\phi\theta_n}(e^{-\alpha} - 1)\right)\right\} \\
&\leq \exp\left\{\alpha n^{\frac{1}{2}+\kappa} - \ell_n^1 \left(e^{-\phi\theta_n}(1 - e^{-\alpha})\right\}\right) .
\end{aligned}$$

For the last expression to converge to zero, we need $K_n := \alpha n^{\frac{1}{2}+\kappa} - \ell_n^1 \left(e^{-\phi\theta_n}(1 - e^{-\alpha})\right) \to -\infty$. Fix $\nu > 0$ and set $\theta_n = \gamma \log n$, $\alpha = \alpha_n = \nu \log n$. Then

$$K_n \leq \nu \log n \cdot n^{\frac{1}{2}+\kappa} - (c_1 n^\beta - 1)\left(n^{-\phi\gamma}(1 - n^{-\nu})\right).$$

Since by (21) $\frac{1}{2} + \kappa < \beta - \phi\gamma$, we have $K_n \to -\infty$, as desired, and thus (19) holds.

Fix $\eta > 0$. Since $(1 + x)^k \leq e^{kx}$ for $x > -1$, we have

$$\begin{aligned}
P\left(\sum_{i=1}^{\ell_n^2} \mathbf{1}_{\{\Psi_i \geq \theta_n\}} \geq n^{\frac{1}{2}-\kappa}\right) &\leq e^{-\eta n^{\frac{1}{2}-\kappa}} E \exp\left[\eta \sum_{i=1}^{\ell_n^2} \mathbf{1}_{\{\Psi_i \geq \theta_n\}}\right] \\
&= e^{-\eta n^{\frac{1}{2}-\kappa}} \left(1 - e^{-\psi\theta_n} + e^{-\eta} e^{-\psi\theta_n}\right)^{\ell_n^2} \\
&\leq e^{-\eta n^{\frac{1}{2}-\kappa}} \exp\left\{\ell_n^2 \left(e^{-\psi\theta_n}(e^{-\eta} - 1)\right)\right\} \\
&= \exp\left\{-\eta n^{\frac{1}{2}-\kappa} + c_2 n^\beta n^{-\psi\gamma}(e^{-\eta} - 1)\right\},
\end{aligned}$$

where on the last line above we substituted $\theta_n = \gamma \log n$. The expression on the last line converges to zero because $\frac{1}{2} - \kappa > \beta - \psi\gamma$ by (21), and (20) follows. ■

Remark 3.1 **(a)** The convergence in (19), (20) is at a geometric rate, as the proof shows. Thus, by the Borel-Cantelli lemma, both events occur for only a finite number of n, with probability one.
(b) As can be seen in the proof, κ and γ depend only on β, ϕ and ψ (cf. (21)), and not on $\{c_i\}$.

Recall that $\mu_* = \operatorname{ess\,inf} m$. Fix $\varepsilon > 0$ and let $\alpha \in (\mu_*, \mu_* + \varepsilon)$ be a continuity point of $x \mapsto m([0, x])$. In what follows, the symbols n and ε are omitted from the notation of all random variables and stochastic processes, and from the parameters μ_k^n. Let $M_0 = [\underline{\mu}, \mu_* - \varepsilon)$, $M_1 = [\mu_* - \varepsilon, \alpha)$ and $M_2 = [\alpha, \bar{\mu}]$ (where $[a, b)$ and $[a, b]$ are interpreted as the empty set if $a > b$), and set

$$K_i = \{k \in \{1, \ldots, n\} : \mu_k \in M_i\}, \quad i = 0, 1, 2.$$

Denote

$$I^{(i)}(t) = \sum_{k \in K_i} I_k(t), \quad T^{(i)}(t) = \sum_{k \in K_i} T_k(t), \quad i = 0, 1, 2. \tag{22}$$

Let also $\widehat{I}^{(i)} = n^{-\frac{1}{2}} I^{(i)}$. By (8) and (10),

$$\widehat{X}(t) = \widehat{X}_0 + n^{-\frac{1}{2}} A(t) - n^{-\frac{1}{2}} \sum_{k=1}^{n} S_k(T_k(t)). \tag{23}$$

By a superposition argument for Poisson processes (cf. Proposition 3.1 of [3]),

$$\widehat{X}(t) = \widehat{X}_0 + n^{-\frac{1}{2}} A(t) - n^{-\frac{1}{2}} \sum_{i=0}^{2} S^{(i)}(T^{(i)}(t)), \tag{24}$$

where $S^{(i)}$, $i = 0, 1, 2$ are rate-1 Poisson processes, mutually independent, and independent of the first, third and fourth items of (12). In particular,

$$D^{(i)}(t) := \sum_{k \in K_i} D_k(t) = S^{(i)}(T^{(i)}(t)), \quad i = 0, 1, 2. \tag{25}$$

The calculation that follows shows

$$\widehat{X}(t) = \widehat{X}_0 + W(t) + bt + F(t), \tag{26}$$

where we recall that all quantities depend on n and ε, and where

$$W(t) = \widehat{A}(t) - \sum_{i=0}^{2} W^{(i)}(t), \tag{27}$$

$$\widehat{A}(t) = n^{-\frac{1}{2}}(A(t) - \lambda^n t), \tag{28}$$

$$W^{(i)}(t) = n^{-\frac{1}{2}}(S^{(i)}(T^{(i)}(t)) - T^{(i)}(t)), \quad i = 0, 1, 2, \tag{29}$$

$$b = n^{-\frac{1}{2}}(\lambda^n - n\lambda) - n^{-\frac{1}{2}} \sum_{k=1}^{n} (\mu_k - \mu), \tag{30}$$

$$F(t) = n^{-\frac{1}{2}} \int_0^t \sum_{k=1}^{n} \mu_k I_k(s) ds. \tag{31}$$

Indeed, by (24), (28)–(29),

$$\widehat{X}(t) = \widehat{X}_0 + \widehat{A}(t) - \sum_{i=0}^{2} W^{(i)}(t) + n^{-1/2}\Big[\lambda^n t - \sum_{i=0}^{2} T^{(i)}(t)\Big]$$
$$= \widehat{X}_0 + W(t) + n^{-\frac{1}{2}}\Big[\lambda^n t - \sum_{k=1}^{n} \mu_k \int_0^t B_k(s)ds\Big],$$

where (27), (9) and (22) are used in the second equality. Since $B_k = 1 - I_k$,

$$\widehat{X}(t) = \widehat{X}_0 + W(t) + n^{-\frac{1}{2}}\Big[\lambda^n - \sum_{k=1}^{n} \mu_k\Big]t + n^{-\frac{1}{2}} \sum_{k} \mu_k \int_0^t I_k(s)ds.$$

By (6), $\mu = \lambda$, hence by (30) the penultimate term above is equal to bt. This shows (26).

Lemma 3.2 *Under Π_0^n, given $\bar{t} > 0$ and $\varepsilon > 0$,*

$$|\widehat{I}^{(2)}|_{*,\bar{t}} \to 0 \text{ in probability, as } n \to \infty. \tag{32}$$

Proof. Step 1: We will show here that there is a (deterministic) sequence a_n increasing to infinity, so that $a_n n^{\frac{1}{2}} \le r_n$, and such that, out of the $a_n n^{\frac{1}{2}}$ servers ranked lowest, the number of those that are in K_2 is $o(n^{\frac{1}{2}})$, in the following sense:

$$\frac{\#\{k \in K_2 : \text{Rank}(k) \le a_n n^{\frac{1}{2}}\}}{n^{\frac{1}{2}}} \Rightarrow 0. \tag{33}$$

We will apply Lemma 3.1. To this end let $\phi \in (\mu_*, \alpha)$ be a continuity point of $x \mapsto m([\underline{\mu}, x])$. Let $\psi = \alpha$. Let $\widetilde{K} = \{k \in \{1, \dots, n\} : \mu_k \le \phi\}$. Since $m([\mu_*, \phi]) > 0$, it follows from (1) that, for some constant $c > 0$ and with probability increasing to 1, the cardinality of $\widetilde{K}$ is at least cn. Since the subset Σ is uniformly distributed and the number of samples satisfies (11), it follows that, on some events Ω^n satisfying $P(\Omega^n) \to 1$,

$$\#\widetilde{K} \cap \Sigma \ge c_1 n^{\beta_0}, \quad \#K_2 \cap \Sigma \le \#\Sigma = n^{\beta_0},$$

for a constant $c_1 > 0$. Recall $\widehat{\mu}_k = 1/\sigma_k$, the reciprocal to the sampled service time. We apply Lemma 3.1 with Φ_i being the samples σ_k with index set $\widetilde{K} \cap \Sigma$, and Ψ_i those with index set $K_2 \cap \Sigma$. We obtain, that on an event $\Omega_1^n \subset \Omega^n$, which also satisfies $\lim_{n\to\infty} P(\Omega_1^n) = 1$,

$$\#\{k \in \widetilde{K} \cap \Sigma : \widehat{\mu}_k \le 1/\theta_n\} > n^{\frac{1}{2}+\kappa}, \tag{34}$$
$$\#\{k \in K_2 \cap \Sigma : \widehat{\mu}_k \le 1/\theta_n\} < n^{\frac{1}{2}-\kappa}, \tag{35}$$

where, without loss of generality, $\frac{1}{2} < \frac{1}{2} + \kappa < \beta_0$. Now, (34) and the way the map Rank is defined, imply that all servers k with $\text{Rank}(k) \le n^{\frac{1}{2}+\kappa}$ have $\widehat{\mu}_k \le 1/\theta_n$ and are in Σ. As a result,

$$\begin{aligned}\#\{k \in K_2 : \text{Rank}(k) \le n^{\frac{1}{2}+\kappa}\} &= \#\{k \in K_2 \cap \Sigma : \text{Rank}(k) \le n^{\frac{1}{2}+\kappa}\} \\ &\le \#\{k \in K_2 \cap \Sigma : \widehat{\mu}_k \le 1/\theta_n\} \\ &\le n^{\frac{1}{2}-\kappa},\end{aligned}$$

by (35). This proves (33) with $a_n = n^\kappa$.

Step 2: Denote

$$K' = \{k \in \{1, \dots, n\} : \text{Rank}(k) > a_n n^{\frac{1}{2}}\},$$

and $I' = \sum_{k\in K'} I_k$, $T' = \sum_{k\in K'} T_k$, $D' = \sum_{k\in K'} D_k$. An argument as the one following equation (23) shows that $S'(T'(t)) = D'(t)$, $t \geq 0$, where S' is a rate-1 Poisson process. Set $\widehat{S}'(t) = n^{-\frac{1}{2}}(S'(nt) - nt)$ and $\widehat{I}' = n^{-\frac{1}{2}} I'$. We shall show that

$$|\widehat{I}'|_{*,\bar{t}} \to 0 \text{ in probability, as } n \to \infty. \tag{36}$$

Note first that the probability of the event $\eta_1 := \{I'(0) = 0\}$ converges to one as $n \to \infty$. Indeed, by (14), $B_{k,0} = 1$ for all k with $\text{Rank}(k) > I_0$. By (4) and (15), $I_0 < a_n n^{\frac{1}{2}}$ with probability converging to 1 as $n \to \infty$. Thus, with probability converging to 1, all servers $k \in K'$ are initially busy, namely $P(\eta_1) \to 1$ as $n \to \infty$. Let

$$\widehat{S}(t) = n^{-\frac{1}{2}}(S(nt) - nt), \quad t \geq 0, \tag{37}$$

where S is a rate-1 Poisson process. It is well known (cf. Lemmas 2 and 4(i) of [4]) that both $\widehat{A}$ (of (28)) and $\widehat{S}$ converge weakly to a zero mean Brownian motion with diffusion coefficient $\lambda^{\frac{1}{2}}\check{C}$, and respectively, 1.

Given $\gamma > 0$, consider the event $\eta := \{|I'|_{*,\bar{t}} > 2\gamma n^{\frac{1}{2}}\}$. On the event $\eta \cap \eta_1$ one can find $0 \leq s < t \leq \bar{t}$ such that $I'(y) > 0$ for all $y \in [s, t]$, and $I'(t) - I'(s) > \gamma n^{\frac{1}{2}}$. Since the servers in K' are all ranked higher than those in the complement set, the routing policy assigns all arrivals within $[s, t]$ to K' servers. Hence by (7), (8) and using $B_k = 1 - I_k$, we have

$$\gamma n^{\frac{1}{2}} < I'(t) - I'(s) = D'(t) - D'(s) - A(t) + A(s),$$

and therefore

$$\gamma < \widehat{S}'(n^{-1}T'(t)) - \widehat{S}'(n^{-1}T'(s)) - \widehat{A}(t) + \widehat{A}(s) + \sum_{k\in K'} \mu_k \int_s^t \widehat{B}_k(y)dy - \lambda n^{\frac{1}{2}}(t - s) - n^{-\frac{1}{2}}(\lambda^n - \lambda)(t - s).$$

We have by (9) and (22) that $n^{-1}T^{(2)}(t) \leq \bar{\mu}\bar{t} =: \tau$. Also, by (5), the last term above is bounded by $c(t - s)$ for some constant c independent of n and ε. Let

$$\bar{w}_\tau(x, z) = \sup_{|s-t|\leq z; s,t\in[0,\tau]} |x(s) - x(t)|, \quad z > 0,$$

denote the modulus of continuity for $x : [0, \tau] \to \mathbb{R}$. Define $C(n, \varepsilon) = n^{-1} \sum_{k\in K'} \mu_k - \lambda$. Then on the event $\eta \cap \eta_1$, with $\delta = t - s$, we have

$$\gamma < \bar{w}_\tau(\widehat{S}', 2\bar{\mu}\delta) + \bar{w}_{\bar{t}}(\widehat{A}, \delta) + n^{\frac{1}{2}} C(n, \varepsilon)\delta + c\delta. \tag{38}$$

By (1), (2), (6) and the definition of K',

$$n^{\frac{1}{2}} C(n, \varepsilon) \leq c_1 - n^{-\frac{1}{2}} \sum_{k:\text{Rank}(k)\leq n^{\frac{1}{2}} a_n} \mu_k \leq c_1 - \underline{\mu} a_n \leq -c_2 a_n,$$

for constants c_1, $c_2 > 0$ and sufficiently large n. Hence

$$P(|\widehat{I}'|_{*,\bar{t}} > 2\gamma) = P(\eta) \le p_1(n,\varepsilon,\gamma) + p_2(n,\varepsilon,\gamma) + P(\eta_1^c),$$

where

$$p_1(n,\varepsilon,\gamma) = P(\text{there exists } \delta \in (0, a_n^{-\frac{1}{2}}) \text{ such that (38) holds}),$$

$$p_2(n,\varepsilon,\gamma) = P(\text{there exists } \delta \in [a_n^{-\frac{1}{2}}, \bar{t}] \text{ such that (38) holds}).$$

Note that

$$p_1(n,\varepsilon,\gamma) \le P(\bar{w}_\tau(\widehat{S}', 2\bar{\mu}a_n^{-\frac{1}{2}}) + \bar{w}_{\bar{t}}(\widehat{A}, a_n^{-\frac{1}{2}}) \ge \gamma/2),$$

$$p_2(n,\varepsilon,\gamma) \le P(\bar{w}_\tau(\widehat{S}', 2\bar{\mu}\bar{t}) + \bar{w}_{\bar{t}}(\widehat{A}, \bar{t}) \ge -c\bar{t} + c_2 a_n^{\frac{1}{2}}).$$

Since $\widehat{S}'$ and $\widehat{A}$ converge to processes with continuous sample paths, both expressions converge to zero as $n \to \infty$. Since $\lim_n P(\eta_1^c) = 0$ and $\gamma > 0$ is arbitrary, (36) follows.

Step 3: Since $K_2 \subset K' \cup ((K')^c \cap K_2)$, we have

$$\widehat{I}^{(2)}(t) = \frac{I^{(2)}(t)}{n^{\frac{1}{2}}} \le \frac{I'(t)}{n^{\frac{1}{2}}} + \frac{[\#(K' \cup ((K')^c \cap K_2))]}{n^{\frac{1}{2}}}\,\bar{t}, \quad t \in [0,\bar{t}].$$

By Step 1 (display (33)), the last term on the above display converges to zero in probability. Thus by Step 2 (display (36)), statement (32) follows. This completes the proof of the lemma. ■

Proof of Theorem 2.1. Based on Lemmas 3.1 and 3.2, the proof is similar to that of Theorem 2.2 of [3] (only slightly simpler). We include it for completeness and because the proof of Theorem 2.2 is based on it. By (26) and (31), one has

$$\widehat{X}(t) = \widehat{X}_0 + W(t) + bt + \mu_* \int_0^t \widehat{X}(s)^- ds + e(t), \tag{39}$$

(where all the above quantities depend on n) and, with $\widehat{I}_k = n^{-\frac{1}{2}} I_k$,

$$e(t) = \sum_{k=1}^{n} (\mu_k - \mu_*) \int_0^t \widehat{I}_k(s) ds. \tag{40}$$

Fix $\bar{t} > 0$. By (2), (5) and (30), $b \to \widehat{\lambda} - \widehat{\mu}$. We show that the random variables $\{|W^{(i)}|_{*,\bar{t}}, i = 0,1,2,\ n \in \mathbb{N}\}$ are tight. By (22) and (9), for $i = 0,1,2$,

$$n^{-1}T^{(i)}(t) = n^{-1} \sum_{k \in K_i} \mu_k t - n^{-1} \sum_{k \in K_i} \mu_k \int_0^t I_k(s) ds. \tag{41}$$

Hence $0 \le n^{-1}T^{(i)}(t) \le \bar{\mu}\bar{t}$ for $t \le \bar{t}$ and all n. Thus by (29), $|W^{(i)}|_{*,\bar{t}} \le |\widehat{S}^{(i)}|_{*,\bar{\mu}\bar{t}}$, where $\widehat{S}^{(i)}(t) = n^{-\frac{1}{2}}(S^{(i)}(nt) - nt)$. Recall from the proof of Lemma 3.2 that $\widehat{S}^{(i)}$ converge to a Brownian motion. Hence $|W^{(i)}|_{*,\bar{t}}$ are tight.

Next, note that $|e(t)| \le \bar{\mu} \int_0^t |\widehat{X}(s)| ds$. Thus the boundedness of b, the tightness of the random variables $\widehat{X}_0$, $|W^{(i)}|_{*,\bar{t}}$ and $|\widehat{A}|_{*,\bar{t}}$, $n \in \mathbb{N}$ (as follows from the convergence of $\widehat{A}$), and an application of Gronwall's lemma on (39), by which $|\widehat{X}|_{*,\bar{t}} \le (|\widehat{X}_0| + |W|_{*,\bar{t}} + |b|\bar{t}) \exp(2\bar{\mu}\bar{t})$, imply that $\{|\widehat{X}|_{*,\bar{t}}, n \in \mathbb{N}\}$ are tight. Since by (13), $\widehat{I} = \widehat{X}^-$, we have that $\{|\widehat{I}|_{*,\bar{t}}, n \in \mathbb{N}\}$ are tight.

The supremum over $t \le \bar{t}$ of the absolute value of the last term in (41) converges to zero in probability, since μ_k are assumed to be bounded and $|\widehat{I}|_{*,\bar{t}}$ are tight. Also, since α is a continuity point of $x \mapsto m([0,x])$, we have that

$$n^{-1} \sum_{k \in K_i} \mu_k \to \int_{M_i} x dm =: \rho_i, \quad i = 0, 1, 2.$$

Note that $\rho_0 = 0$. As a result, $n^{-1}(T^{(0)}, T^{(1)}, T^{(2)}) \to \widetilde{\rho}$ in probability, uniformly on $[0, \bar{t}]$, where $\widetilde{\rho}(t) = (0, \rho_1 t, \rho_2 t)$. Recall that $(\widehat{A}, \widehat{S}^{(0)}, \widehat{S}^{(1)}, \widehat{S}^{(2)})$ are mutually independent, and that $\widehat{S}^{(i)}$ [resp., $\widehat{A}$] converges to a standard Brownian motion [a zero mean Brownian motion with diffusion coefficient $\lambda^{\frac{1}{2}}\check{C}$] (see comment following (37)). Thus (27), (29) and the lemma on random change of time [7, p. 151] show that W converges weakly to σw, in the uniform topology on $[0, \bar{t}]$, where w is a standard Brownian motion and $\sigma^2 = \lambda\check{C}^2 + \rho_1 + \rho_2 = \lambda\check{C}^2 + \mu = \mu(\check{C}^2 + 1)$.

By the Skorohod representation theorem, we can assume without loss of generality that the random variables $\widehat{X}_0$ and ξ_0 and the processes W and w are realized in such a way that, P-a.s.,

$$(\widehat{X}_0, W) \to (\xi_0, \sigma w) \quad \text{as } n \to \infty. \tag{42}$$

Let ξ be the unique strong solution to equation (17). Then by (17), (39), the inequality $|x^- - y^-| \le |x - y|$, and Gronwall's inequality,

$$|\widehat{X} - \xi|_{*,\bar{t}} \le (|\widehat{X}_0 - \xi_0| + |b - (\widehat{\lambda} - \widehat{\mu})| + |W - \sigma w|_{*,\bar{t}} + |e|_{*,\bar{t}}) \exp(\mu_* \bar{t}). \tag{43}$$

Now, by (40), for n sufficiently large,

$$|e|_{*,\bar{t}} \le \varepsilon \bar{t}\, |\widehat{I}|_{*,\bar{t}} + \bar{\mu}\bar{t}\, |\widehat{I}^{(2)}|_{*,\bar{t}} + (\mu_* - \underline{\mu})\bar{t}\, |\widehat{I}^{(0)}|_{*,\bar{t}}. \tag{44}$$

By (3), the last term above converges weakly to 0. Combining (32), (42), (43) and (44),

$$\limsup_n P(|\widehat{X} - \xi|_{*,\bar{t}} > \varepsilon^{\frac{1}{2}}) \le \limsup_n P(c\varepsilon |\widehat{I}|_{*,\bar{t}} > \varepsilon^{\frac{1}{2}}),$$

where $c \in (0, \infty)$ is a constant independent of n and ε. Note that the law of $|\widehat{I}|_{*,\bar{t}}$ does not depend on ε. Hence by tightness of $\{|\widehat{I}|_{*,\bar{t}}, n \in \mathbb{N}\}$ the r.h.s. in the above display converges to zero as $\varepsilon \to 0$. Thus $|\widehat{X} - \xi|_{*,\bar{t}} \to 0$ in probability. Since $\bar{t}$ is arbitrary, we have $\widehat{X} \Rightarrow \xi$. ■

Proof of Theorem 2.2. By (3) there exists a sequence $\delta_n > 0$ tending to zero such that $\zeta_n := \#\{k : \mu_k^n < \mu_* - \delta_n\} n^{-\frac{1}{2}} \to 0$. Note that (39), (40) still hold. Define Ξ_1 as the solution to

$$\Xi_1(t) = \widehat{X}_0 + W(t) + bt + \mu_* \int_0^t \Xi_1(s)^- ds. \tag{45}$$

Then by (39), $\Delta := \widehat{X} - \Xi_1$ is differentiable, and, using the inequality $a^- - b^- \le -(a-b)^+$ for $a, b \in \mathbb{R}$, we have $\Delta(0) = 0$, and

$$\frac{d}{dt}\Delta(t) \ge -\mu_* \Delta(t)^+ + \frac{d}{dt} e(t).$$

Since $\widehat{I}_k \leq n^{-\frac{1}{2}}$ for each k, we have by (40)

$$\frac{d}{dt}e(t) \geq -v_n, \qquad v_n := \delta_n |\widehat{I}|_{*,\bar{t}} + \mu_* \zeta_n,$$

and $\Delta(0) = 0$. By comparison with the ordinary differential equation $du/dt = -\mu_* u^+ - v_n$, $u(0) = 0$, we obtain that $\Delta(t) \geq -v_n t$, $t \leq \bar{t}$. Hence $X(t) \geq \Xi(t)$, where we define $\Xi(t) = \Xi_1(t) - v_n t$, $t \leq \bar{t}$.

It thus remains to show that $\Xi \Rightarrow \xi$. For this let us review the proof of Theorem 2.1. Rather than three processes $D^{(i)}$ (25) and correspondingly $W^{(i)}$, $i = 0, 1, 2$, (29), we now have a single process $D = \sum_k D_k$ given in terms of a single rate-1 Poisson process S^n (cf. (18)). The adaptation of relation (29) to a single process W is obvious. The arguments in the proof of Theorem 2.1 that lead to the tightness of $|\widehat{I}|_{*,\bar{t}}$ and the convergence of W to σw hold with obvious modifications. As in that proof, we deduce that (42) can be assumed without loss of generality. Equations (17), (45) and Gronwall's inequality thus yield

$$|\Xi_1 - \xi|_{*,\bar{t}} \leq (|\widehat{X}_0 - \xi_0| + |b - (\widehat{\lambda} - \widehat{\mu})| + |W - \sigma w|_{*,\bar{t}}) \exp(\mu_* \bar{t}).$$

Hence (42) and the convergence of b to $\widehat{\lambda} - \widehat{\mu}$ imply that Ξ_1 converges in probability to ξ uniformly over $[0, \bar{t}]$. The random variables v_n converge to zero by tightness of $|\widehat{I}|_{*,\bar{t}}$. Since $\bar{t}$ is arbitrary, we thus obtain that $\Xi \Rightarrow \xi$. This completes the proof of the theorem. ■

Remark 3.2 Note that the non-idling property is used in the proof of Theorem 2.1 (on which the above proof is based) for deducing tightness of $|\widehat{I}|_{*,\bar{t}}$ from that of $|\widehat{X}|_{*,\bar{t}}$. As can be easily seen, $|\widehat{I}|_{*,\bar{t}}$ are not in general tight if the restriction to non-idling policies is removed.

Acknowledgements: We thank the AE and referee for very useful comments and for pointing out the relation to reference [12]. Adam Shwartz holds The Julius M. and Bernice Naiman Chair in Engineering. Work of both authors is supported in part by the fund for promotion of research and the fund for promotion of sposored research at the Technion.

References

[1] M. Armony. Dynamic routing in large-scale service systems with heterogeneous servers, *Queueing Systems*, 51(3-4) (2005), 287–329.

[2] R. Atar. Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab.*, 15, No. 4, 2606–2650. (2005)

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

[4] R. Atar, A. Mandelbaum and M. I. Reiman. Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 14 (2004), No. 3, 1084–1134.

[5] R. Atar, A. Mandelbaum and G. Shaikhet. Simplified control problems for multi-class many-server queueing systems. In preparation.

[6] A. Bassamboo, J. M. Harrison, A. Zeevi. Design and control of a large call center: asymptotic analysis of an LP-based method. *Oper. Res.* 54 (2006), No. 3, 419–435.

[7] P. Billingsley. *Convergence of probability measures. Second edition.* Wiley, New York, 1999.

[8] N. Gans, G. Koole and A. Mandelbaum. Telephone call centers: Tutorial, review, and reserach prospects. *Manufacturing and Service Oper. Manag.* 5, 79–141.

[9] I. Gurwich, W. Whitt. Scheduling flexible servers with convex delay costs in many-server service systems. Preprint.

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

[11] H. Kaspi and K. Ramanan. Fluid limits for the GI/GI/N queue. Preprint.

[12] M. Mitzenmacher, A. W. Richa and R. Sitaraman. The power of two random choices: a survey of techniques and results. *Handbook of randomized computing, Vol. I, II, 255–312, Comb. Optim.*, 9, Kluwer Acad. Publ., Dordrecht, 2001.

[13] J. E. Reed. The G/GI/N queue in the Halfin-Whitt regime. Preprint.

[14] T. Tezcan. Asymptotically optimal control of many-server heterogeneous service systems with hyper-exponential service times. Preprint.

[15] T. Tezcan and J. Dai. Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Preprint.

[16] W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. Preprint.