



Dynamic Routing in Large-Scale Service Systems with Heterogeneous Servers

MOR ARMONY

marmony@stern.nyu.edu

Stern School of Business, New York University, 44 West Fourth Street #8-62, New York, NY 10012

Received 1 October 2004; Revised 11 March 2005

Abstract. Motivated by modern call centers, we consider large-scale service systems with multiple server pools and a single customer class. For such systems, we propose a simple routing rule which asymptotically minimizes the steady-state queue length and virtual waiting time. The proposed routing scheme is FSF which assigns customers to the Fastest Servers First. The asymptotic regime considered is the Halfin-Whitt many-server heavy-traffic regime, which we refer to as the Quality and Efficiency Driven (QED) regime; it achieves high levels of both service quality and system efficiency by carefully balancing between the two. Additionally, expressions are provided for system limiting performance measures based on diffusion approximations. Our analysis shows that in the QED regime this heterogeneous server system outperforms its homogeneous server counterpart.

Keywords: call centers, heavy-traffic, routing, control of queueing systems, heterogeneous servers, Halfin Whitt regime, QED regime, asymptotic analysis

AMS subject classification: 60K25, 68M20, 90B22

1. Introduction

Motivated by call centers we study large scale service systems with many heterogeneous servers. Heterogeneity in server population is typical in large service systems due to their size. Moreover, service systems such as call-centers often operate from multiple locations, which tends to increase the level of server heterogeneity even more. We study such systems via a stylized model with a single customer class and multiple server pools/types. Servers of each pool are characterized by the speed in which they serve customers. We denote this model as the \wedge model (which we name the ‘inverted-V model’). It is depicted in figure 1. A special application of the \wedge model is the two types of servers example with a pool of trainees and a pool of experienced servers. More generally, it can capture natural heterogeneity within the server population. With respect to the \wedge model we ask the following routing question: How to match between customers and servers of the different pools so as to optimize system performance?

An intuitive solution to the routing question would be to assign customers to the Faster Servers First (FSF). That is, upon a customer arrival or a service completion send the customer at the head of the line to the fastest server available. Though intuitive, it turns out that this solution is not optimal. As was shown for the two server

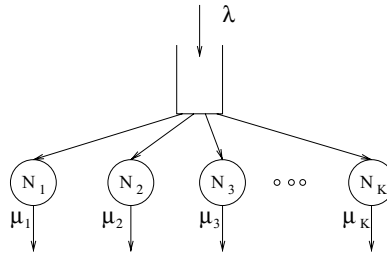


Figure 1. The inverted-V model—a single customer class and multiple server types.

case (e.g. [18,39]) it is sometimes necessary to keep customers waiting even when the slower server is idle in order not to starve the faster server. Specifically, optimality is obtained by assigning customers to the slower server only if the faster server is busy *and* the number of customers in queue exceeds a certain threshold. The value of the threshold depends on the system parameters. This puts a computational burden on those who wish to implement this threshold policy. Matters simplify when “conventional” heavy-traffic (traffic intensity converges to 1, but number of servers stays fixed) is considered. For this case [55] showed that a state-independent threshold policy is asymptotically optimal as traffic intensity approaches 1. In particular, in the limit, the threshold level approaches infinity. According to [19], the problem of finding an optimal policy for the general multi-heterogeneous server case (considered in [41,48,49]) is still open. But what about the many-server heavy-traffic regime that typifies large call centers? Is the FSF policy optimal in that regime? Is a threshold needed in that case?

The asymptotic framework considered in this paper is the many-server heavy-traffic regime, first appearing in Erlang [20], and formally introduced by Halfin and Whitt [32]. We refer to this regime as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy a rare combination of high efficiencies together with high quality of service. In the context of this limiting regime the routing problem becomes: How to match between customers and servers of the different pools so as to *asymptotically* optimize system performance, as the arrival rate and the number of servers in each pool increase to infinity, according to the QED regime?

The main contributions of this paper are:

1. *Asymptotic optimality of FSF*: We show that the FSF policy is asymptotically optimal in the QED regime. In particular, no thresholds are needed.
2. *Asymptotic coupling*: Methodologically, we prove the asymptotic optimality of FSF using a non-trivial coupling argument between FSF and a related preemptive FSF policy:
 - (a) *Preemptive optimality*: The preemptive FSF policy, which keeps the faster servers busy whenever possible (even at the account of handing off customers from

- a slower server to a faster one) is optimal in the sense that it stochastically minimizes the steady-state queue length and waiting time.
- (b) *State-space collapse*: In the limit, the multi-dimensional process which describes the system state under FSF is equivalent to a one-dimensional process.
 - (c) *Non-preemptive asymptotic optimality*: Due to the state-space collapse, the FSF policy and its preemptive counterpart have the same *limiting* performance. Therefore, FSF is asymptotically optimal.
3. *Limiting performance analysis*: We provide expressions for system performance measures based on a diffusion limit and the state-space collapse result.
 4. *Comparison to the homogeneous server system*: We show that in the QED regime, the heterogeneous server system outperforms its homogeneous server counterpart. This is true, provided that (a) the mean service time and the total service capacity are the same for both systems, and (b) FSF is used for routing decisions in the heterogeneous server system.

The remainder of the paper is organized as follows: As part of the introduction, we proceed with a more concrete summary of our results. We then conclude the introduction by reviewing the relevant literature. In Section 2, we detail the (single-customer-class multiple-server-types) \wedge model, and the asymptotic framework used in our analysis. In Section 3, we describe the preemptive FSF policy and establish its optimality. In Section 4, the FSF routing policy is studied and its asymptotic optimality is established. Most proofs appear in the appendix. Conclusions and extensions are finally discussed in Section 5.

1.1. Summary of the results

The QED asymptotic regime considered in this paper can be formally defined as follows: A system with large call volume (demand) and many servers is operating in the QED regime if (a) the delay probability is less than 1 (but not 0), or (b) its total service *capacity* is equal to the demand plus a safety capacity which is of the same order of magnitude as the square root of the demand. (The familiar reader may notice that this is a slightly different characterization of the QED regime in terms of the total service capacity instead of the total number of servers. This characterization is suitable for the \wedge -design studied in this paper.) It is easily seen that, under (b), the system operates in heavy-traffic, and hence the high server efficiencies. The quality aspect of the QED regime is seen from characterization (a). This high performance, which is typically impossible to achieve for systems in heavy traffic, is obtained here due to the economies of scale associated with the large number of servers. These two characterizations of the QED regime are shown to be equivalent in various settings (see Section 1.2), including the one considered in this paper.

The asymptotically optimal routing policy we propose is the policy Faster Server First (FSF) that simply assigns newly arriving or waiting customers to the fastest server

available. FSF is shown to be asymptotically optimal among all the non-anticipating non-preemptive FIFO policies (Theorem 4.1). The asymptotic optimality is in terms of the steady-state queue length and waiting time distributions in the QED regime. More specifically, consider a sequence of systems indexed by the arrival rate λ , where $\lambda \uparrow \infty$. For any fixed value of λ , let N_k^λ represent the number of servers of type k , $k = 1, \dots, K$. Also, let $\vec{N}^\lambda = (N_1^\lambda, N_2^\lambda, \dots, N_K^\lambda)$ be the staffing vector, and $N^\lambda = N_1^\lambda + N_2^\lambda + \dots + N_K^\lambda$ be the total number of servers. Suppose that the service rates: μ_1, \dots, μ_K are fixed independently of λ . To be consistent with the QED regime assume that the total service capacity, $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + \dots + \mu_K N_K^\lambda$, is equal to the arrival rate plus a square root safety capacity. Formally, suppose that

$$\sum_{k=1}^K N_k^\lambda \mu_k = \lambda + \delta \sqrt{\lambda} + o(\sqrt{\lambda}), \quad (1.1)$$

for some positive number δ . Let Q^λ and W^λ be the queue length and the virtual waiting time processes, respectively. For asymptotic purposes let $\tilde{Q}^\lambda = Q^\lambda / \sqrt{N^\lambda}$ and $\tilde{W}^\lambda = \sqrt{N^\lambda} W^\lambda$ be the *scaled* queue length and waiting time processes, respectively, and let $\tilde{Q}^\lambda(\infty)$ and $\tilde{W}^\lambda(\infty)$ be the corresponding steady-state distributions. The asymptotic optimality of the FSF policy is in terms of stochastic minimization of the limiting distributions of $\tilde{Q}^\lambda(\infty)$ and $\tilde{W}^\lambda(\infty)$ as $\lambda \rightarrow \infty$ (see Theorem 4 for further details).

Although FSF is a very intuitive policy, its optimality is not easy to establish. Particularly, the question as to what is the optimal policy for the general multiple heterogeneous servers case is still open [19]. However, in the QED regime, we are able to show that, although the FSF is probably not optimal, it is indeed *asymptotically* optimal. To establish the asymptotic optimality of FSF we first introduce a related *preemptive* policy, FSF_p . This policy keeps the faster servers busy whenever possible, even at the cost of handing-off customers from slower servers to faster ones. The policy FSF_p is shown (in Proposition 3.1) to stochastically minimize the steady-state queue length and waiting time, for any fixed system in the sequence (associated with a fixed value of λ). Consequently, we show that, in the limit as $\lambda \rightarrow \infty$, both policies give rise to the same performance measures (Propositions 4.5 and 4.6, and Lemma 4.1). That is, in the limit, they both have the same distributions for $\tilde{Q}^\lambda(\infty)$ and $\tilde{W}^\lambda(\infty)$. In particular, the limiting *waiting probability* in steady-state is also minimized by FSF.

Provided that the system operates under the FSF routing policy, our analysis gives explicit expressions for the diffusion processes which are obtained as the limits of the (centered and scaled) total number of customers in the system, total queue length and waiting time processes (Propositions 4.2 and 4.3, and Remark 4.7). A state-space collapse result then shows that, in the limit, the number of busy servers of each server pool can be expressed as a function of the total number of busy servers (Proposition 4.1 and Remark 4.5). Additionally, we show that the steady-state distribution of all these processes converges to the respective steady-state distribution of the corresponding limiting diffusion process (Propositions 4.5 and 4.6). This implies that one can approximate steady-state performance measures associated with either waiting times, queue length or server

idleness by using the steady-state expressions of their diffusion counterparts; moreover, these approximations become exact as the system approaches the QED regime.

Surprisingly, we find that, in the QED regime, the heterogeneous server system considered in this paper outperforms its homogeneous server counterpart, provided that the former operates under the FSF routing policy (Corollary 1 and Remark 4.9). Specifically, consider a system with statistically identical servers, each operates with a service rate μ which satisfies $\mu = q_1\mu_1 + q_2\mu_2 + \dots + q_K\mu_K$. For $k = 1, \dots, K$, the weights q_k are the limiting fractions of N_k , the number of servers in pool k , out of the total number of servers, N . Suppose that the total service capacity of the homogeneous server system is equal to the total server capacity in the heterogeneous server systems (which, in the case that the fractions N_k/N are fixed, is equivalent to assuming that the number of servers is the same in both systems). Then, in the QED limit, the heterogeneous system is stochastically better than the homogeneous server one in terms of the queue length and the virtual waiting time. This result is intuitive and not difficult to prove in a preemptive setting using sample-path arguments. Specifically, for any arbitrary number of customers the total service rate devoted to these customers in the heterogeneous server system (operating under FSF $_{\rho}$) is no less than it is in the homogeneous server system. However, in the non-preemptive setting things are more intricate. In particular, examples can be found in which the homogeneous servers perform better than the heterogeneous ones. Nevertheless, in the QED limit, since the FSF routing policy is asymptotically as good as its preemptive version, the heterogeneous system is indeed better.

1.2. Literature review

The QED regime: Asymptotic theory of many-server queues

The QED regime has been given much attention in the last few years, especially in the “ I^k ”-model, which corresponds to multiple independent queues, each with its own devoted server pool (no overlap in skills). For a formal description, consider a sequence of multiple server queues, indexed by the arrival rate λ , with the number of servers N^λ growing to ∞ as $\lambda \uparrow \infty$. Define the offered load by $R^\lambda = \frac{\lambda}{\mu}$, where μ is the service-rate, which is held fixed (independent of λ). The QED regime is achieved by letting $\sqrt{N^\lambda}(1 - \rho^\lambda) \rightarrow \beta$, as $\lambda \uparrow \infty$, for some finite β . Here $\rho^\lambda = R^\lambda/N^\lambda$ is the servers’ long-run utilization. Equivalently, the staffing level is approximately given by

$$N^\lambda \approx R^\lambda + \beta\sqrt{R^\lambda}, \quad -\infty < \beta < \infty. \tag{1.2}$$

Yet another equivalent characterization is a non-trivial limit (within $(0, 1)$) of the fraction of *delayed* customers. The latter equivalence was established for *GI/M/N* [32], *GI/D/N* [36] and *M/M/N* with exponential patience [27]. More recently, this equivalence was also shown to be true in the multiple customer classes case [30], as well as in the setting considered in this paper [4].

The QED regime was explicitly recognized already in Erlang’s 1923 paper (that appeared in [20]) which addresses both Erlang-B (*M/M/N/N*) and Erlang-C (*M/M/N*)

models. Later on, extensive related work took place in various telecom companies but little has been openly documented, as in Sze [54] (who was actually motivated by AT&T call centers operating in the QED regime). A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given in Jagerman [35]; see also [57], and then [44] for the analysis of finite buffers. But the operational significance of the QED regime, in particular its balancing of “service and economy” via a non-trivial delay probability, was first discovered and formalized by Halfin and Whitt [32]: Within the GIM/N framework, they analyzed the scaled number of customers, both in steady-state and as a stochastic process. Recent generalizations are [59,60]. Convergence of the scaled queueing process, in the more general $GI/PH/N$ setting, was established in [46]. Application of QED queues to modelling and staffing of telephone call centers and communication networks, taking into account customers’ impatience, can be found in [27] and [23], respectively. The optimality of the QED regime, under revenue maximization or constraint satisfaction, is discussed in [2,3,12,42]. Readers are referred to Sections 4 and 5.1.4 of [25] for a survey of the QED regime, both practically and academically.

It is important to note that the QED regime differs in significant ways from the conventional (or “classical”) heavy traffic regime. Indeed, QED combines light and heavy traffic characteristics. For example, in conventional heavy traffic, the theory of which has been well established [16,58], essentially *all* customers are delayed prior to service. In the QED regime, on the other hand, a non-trivial fraction is served immediately upon arrival. Also, conventional heavy traffic can be achieved by setting $N \approx R + \beta$, for some constant β , rather than the square-root form in (1.2). For more details, readers are referred to [25].

Skill-based routing

The routing problem considered in this paper is a special case of skill-based routing where multiple customer classes are matched with multiple server pools. There is an extensive literature on skill-based routing. Examples include: *Exact (non-asymptotic) analysis*: Lin and Kumar [39], Kella and Yechiali [37], Federgruen and Groenvelt [22], Brandt and Brandt [14], Gans and Zhou [26], Armony and Bambos [1], Rykov [48], Luh and Viniotis [41], Bhulai and Koole [11], de Véricourt and Zhou [18], Shumsky [52], and Rykov and Efrosinin [49], *Asymptotic analysis—“conventional” heavy traffic*: Foschini [24], Reiman [47], Kelly and Laws [38], Harrison [33], Bell and Williams [10], Glazebrook and Niño-Mora [28], Teh and Ward [55], Mandelbaum and Stolyar [43] and Stolyar [53], and *Asymptotic analysis—many server systems*: Armony and Maglaras [2,3], Harrison and Zeevi [34], Atar et al. [7], Atar [5,6], Gurvich [30], Gurvich, Armony and Mandelbaum [31], Wallace and Whitt [56] and Bassamboo, Harrison and Zeevi [8,9].

2. Model formulation

Consider a service system with a single customer class and K server types (each type in its own server pool), all are capable of fully handling customers’ service requirements.

Service times are assumed to be exponential, with service rate that depends on the pool (type) of the particular server. Specifically, the average service time of a customer that is served by a server of type k ($k = 1, 2, \dots, K$) is $1/\mu_k$. We assume that the service rates are ordered as follows: $\mu_1 < \mu_2 < \dots < \mu_K$. Customers arrive to the system according to a Poisson process with rate λ . Delayed customers wait in an infinite buffer, and are served according to a FCFS discipline. All interarrival times and service times are assumed to be statistically independent.

Let N_k be the number of server in pool k . Respectively, let $\vec{N} = (N_1, N_2, \dots, N_K)$ be the staffing vector (here and elsewhere, \vec{x} is used to denote a vector whose elements are x_1, x_2, \dots). Suppose that the following necessary condition for stability is satisfied:

$$N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K > \lambda, \tag{2.1}$$

that is, the total service capacity is larger than the arrival rate. Let Π be the set of all non-preemptive non-anticipating FIFO routing policies. Denote by $\pi := \pi(\lambda, \vec{N}) \in \Pi$, a policy that operates in a system with arrival rate λ and staffing vector \vec{N} (at times we will omit the arguments λ and \vec{N} when it is clear from the context which arguments should be used). Given a policy $\pi \in \Pi$, let $P_\pi(\text{wait} > 0)$ be the steady-state probability that a customer is delayed before his service starts. (If the steady-state distribution does not exist, consider $P_\pi(\text{wait} > 0)$ as the random variable corresponding to the essential limsup of the long term proportion of customers who are delayed before receiving service.) Our goal in this paper is to find a policy in Π which stochastically minimizes the steady-state queue length and virtual waiting time.

A more ambitious goal is to identify staffing levels N_1, \dots, N_K and a routing policy to minimize staffing costs subject to a constraint on system performance such as the probability of waiting. Generally, solving the staffing and control problems concurrently is too difficult. Hence, researchers commonly end up solving one while assuming the solution to the other is fixed. A distinguishing feature of our approach is that we identify a policy which is near-optimal in terms of the steady-state queue-length and virtual waiting time given *any* staffing level, and therefore, we are able to solve the staffing and the control problems concurrently. The staffing problem is studied in [4].

Suppose that the routing policy $\pi \in \Pi$ is used, and let $t \geq 0$ be an arbitrary time point. We denote by $Z_k(t; \pi)$ the number of busy servers of pool k ($k = 1, 2, \dots, K$) at time t , and $Q(t; \pi)$ the queue length at this time. Finally, let $Y(t; \pi)$ be the total number of customers in the system. That is, $Y(t; \pi) = Z_1(t; \pi) + Z_2(t; \pi) + \dots + Z_K(t; \pi) + Q(t; \pi)$. We use $t = \infty$ whenever we refer to the steady-state. At times, we will omit π if it is clear from the context which routing policy is used.

Definition. A policy $\pi \in \Pi$ is called *work conserving* if there are no idle servers whenever there are some delayed customers in the queue. In other words, π is work conserving if $Q(t; \pi) > 0$ implies that $Z_1(t; \pi) + Z_2(t; \pi) + \dots + Z_K(t; \pi) = N$, where

$$N := N_1 + N_2 + \dots + N_K$$

is the total number of servers.

Note that in general a $K + 1$ dimensional vector is required to specify the state of the system, namely, $Q(t; \pi)$ and $Z_1(t; \pi), \dots, Z_K(t; \pi)$. However, for work conserving policies, the state space can be described by the K -dimensional vector $(Z_1(t; \pi) + Q(t; \pi), Z_2(t; \pi), \dots, Z_K(t; \pi))$. In fact, the queue length can be added to the number of busy servers of pool k , for any k , because if π is work conserving then $Q(t; \pi) = [Q(t; \pi) + Z_k(t; \pi) - N_k]^+$ (where $[x]^+ := \max\{x, 0\}$) and $Z_k(t; \pi) = \min\{Q(t; \pi) + Z_k(t; \pi), N_k\}$. Work conserving policies also have the appealing property that the waiting probability can be stated in terms of the total number of busy servers. In particular, if $\pi \in \Pi$ is work conserving, and there exists a steady-state distribution for its underlying processes, then

$$\begin{aligned} P_\pi(\text{wait} > 0) &= P(Z_1(\infty; \pi) + Z_2(\infty; \pi) + \dots + Z_K(\infty; \pi) = N) \\ &= P(Y(\infty; \pi) \geq N), \end{aligned} \quad (2.2)$$

where the first equality is due to the PASTA property, and the second follows from work-conservation. Note that if the policy is not work conserving then (2.2) does not hold, because one may have customers waiting in queue, even if some of the servers are idle.

Let $A(t)$ be the total number of arrivals into the system up to time t (that is, $A(t)$, $t \geq 0$ is a Poisson process with rate λ). Also, for $k = 1, \dots, K$ and for a policy $\pi \in \Pi$, let $A_k(t; \pi)$ be the total number of external arrivals joining pool k upon arrival up to time t , and let $B_k(t; \pi)$ be the total number of customers joining server pool k , up to time t , after being delayed in the queue. The number of arrivals into the queue (and not directly to one of the servers) up to time t is denoted by $A_q(t; \pi)$. In addition, let $T_k(t; \pi)$ denote the total time spent serving customers by all N_k servers of pool k up to time t . In particular, $0 \leq T_k(t; \pi) \leq N_k t$. Respectively, let $I_k(t; \pi)$ be the total idle time experienced by servers of pool k up to time t . Finally, let $D_k(t)$ be a Poisson process with rate μ_k . Then the number of service completions out of server pool k may be written as $D_k(T_k(t; \pi))$. The above definitions allow us to write the following *flow balance equations*:

$$Q(t; \pi) = Q(0; \pi) + A_q(t; \pi) - \sum_{k=1}^K B_k(t; \pi), \quad (2.3)$$

$$\begin{aligned} Z_k(t; \pi) &= Z_k(0; \pi) + A_k(t; \pi) + B_k(t; \pi) \\ &\quad - D_k(T_k(t; \pi)), \quad k = 1, \dots, K, \end{aligned} \quad (2.4)$$

$$T_k(t; \pi) = \int_0^t Z_k(s; \pi) ds \quad (2.5)$$

$$Y(t; \pi) = Y(0; \pi) + A(t) - \sum_{k=1}^K D_k(T_k(t; \pi)), \quad (2.6)$$

$$A(t) = A_q(t; \pi) + \sum_{k=1}^K A_k(t; \pi), \quad (2.7)$$

$$T_k(t; \pi) + I_k(t; \pi) = N_k t. \quad (2.8)$$

Finally, for work conserving policies we have the additional equations:

$$Q(t; \pi) \cdot \left(\sum_{k=1}^K (N_k - Z_k(t; \pi)) \right) = 0, \tag{2.9}$$

$$\int_0^\infty \sum_{k=1}^K (N_k - Z_k(t; \pi)) dA_q(t; \pi) = 0, \tag{2.10}$$

and

$$\sum_{k=1}^K \int_0^\infty Q(t; \pi) dI_k(t; \pi) = 0. \tag{2.11}$$

In words, (2.9) means that there are customers in queue only when *all* servers are busy. The verbal interpretation of (2.10) is that new arrivals wait in the queue only when all servers are busy. Finally, (2.11) states that servers can only be idle when the queue is empty.

2.1. Asymptotic framework

The routing problem defined above is difficult to solve exactly. Specifically, given fixed values of $\mu_1, \mu_2, \dots, \mu_K, \lambda$ and $\vec{N} = (N_1, N_2, \dots, N_K)$, one needs to find a policy $\pi = \pi(\lambda, \vec{N}) \in \Pi$ that stochastically minimizes the queue length and the waiting time. This is hard to do. Even with less ambitious targets, such as minimizing the average steady-state number of customers in the system researchers have not yet been able to *characterize* the optimal routing policy [19]. Instead, we take an asymptotic approach, which finds asymptotically optimal routing rules for systems with many servers and high demand (i.e. large values of λ and N). To this end, we consider a sequence of systems indexed by λ (to appear as a superscript) with increasing arrival rates $\lambda \uparrow \infty$, and increasing total number of servers N^λ but with fixed service rates $\mu_1, \mu_2, \dots, \mu_K$.

Assume that there are K numbers $a_k \geq 0, k = 1, \dots, K$, with $a_1 > 0$ and $\sum_{k=1}^K a_k = 1$, such that the number of servers of each pool $N_k^\lambda, k = 1, 2, \dots, K$, grows with λ as follows:

$$N_k^\lambda = a_k \frac{\lambda}{\mu_k} + o(\lambda), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\mu_k N_k^\lambda}{\lambda} = a_k. \tag{A1}$$

Recall that server pool 1 has the slowest servers. The assumption that $a_1 > 0$ implies that this server pool is of non-negligible size. Its importance will become apparent later (see the proof of Proposition 4.1). Condition (A1) guarantees that the total traffic intensity,

$$\rho^\lambda := \frac{\lambda}{\sum_{k=1}^K \mu_k N_k^\lambda}, \tag{2.12}$$

converges to 1, as $\lambda \rightarrow \infty$, and hence, for large λ , the system is in *heavy traffic*. Also, in view of (A1), the quantity $a_k \lambda / \mu_k$ can be considered as the offered load of server pool k . Let

$$\mu = \left[\sum_{k=1}^K a_k / \mu_k \right]^{-1}, \quad (2.13)$$

then λ / μ is the total offered load of the whole system. Given this definition of μ , (A1) implies that

$$N^\lambda = \frac{\lambda}{\mu} + o(\lambda), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\lambda}{N^\lambda} = \mu, \quad (2.14)$$

where $N^\lambda = \sum_{k=1}^K N_k^\lambda$. Also,

$$\rho^\lambda \approx \frac{\lambda}{N^\lambda \mu}, \quad (2.15)$$

in the sense that $\lim_{\lambda \rightarrow \infty} \rho^\lambda / (\lambda / N^\lambda \mu) = 1$. Finally,

$$\lim_{\lambda \rightarrow \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu} \mu := q_k \geq 0, \quad k = 1, \dots, K, \quad (2.16)$$

where q_k is the limiting fraction of pool k servers out of the total number of servers. The condition $a_1 > 0$ guarantees that $q_1 > 0$, and hence server pool 1 is asymptotically non-negligible in size. Clearly, $\sum_{k=1}^K q_k = 1$ and $\sum_{k=1}^K q_k \mu_k = \mu$.

Fluid Scaling. In view of the above discussion, one observes that assumption (A1) implies that quantities involved in the process such as the arrival rate, the offered load, and the size of the different server pools are all of order $\Theta(N^\lambda)$. Therefore, one expects to get finite limits of these quantities when dividing all of them by N^λ . As it turns out, due to the functional strong law of large numbers (FSLLN), this scaling leads to the fluid dynamics of the system, in the limit as $\lambda \rightarrow \infty$. To see this, for $\lambda \uparrow \infty$, $k = 1, \dots, K$ and a fixed sequence of routing policies $\pi^\lambda := \pi(\lambda, N^\lambda) \in \Pi$ (omitted from the following notation) let $\bar{Q}^\lambda(t) = \frac{Q^\lambda(t)}{N^\lambda}$, and $\bar{Z}_k^\lambda(t) = \frac{Z_k^\lambda(t)}{N^\lambda}$. Similarly, let $\bar{Y}^\lambda(t) = \frac{Y^\lambda(t)}{N^\lambda}$, $\bar{A}^\lambda(t) = \frac{A^\lambda(t)}{N^\lambda}$, $\bar{A}_k^\lambda(t) = \frac{A_k^\lambda(t)}{N^\lambda}$, $\bar{A}_q^\lambda(t) = \frac{A_q^\lambda(t)}{N^\lambda}$, $\bar{B}_k^\lambda(t) = \frac{B_k^\lambda(t)}{N^\lambda}$, $\bar{T}_k^\lambda(t) = \frac{T_k^\lambda(t)}{N^\lambda}$, and $\bar{I}_k^\lambda(t) = \frac{I_k^\lambda(t)}{N^\lambda}$. Finally, let $\bar{D}_k^\lambda(t) = D_k^\lambda(t) = D_k(t)$. That is, as equalities between processes, $(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda) = (Q^\lambda, Z_k^\lambda, Y^\lambda, A^\lambda, A_k^\lambda, A_q^\lambda, B_k^\lambda, T_k^\lambda, I_k^\lambda) / N^\lambda$, and $\bar{D}_k^\lambda = D_k$. Note that D_k^λ need not be divided by N^λ , due to its definition as a Poisson process with rate μ_k , which is independent of λ .

Using standard tools of fluid models (see for example [17], Theorem 2.3.1) one can show that if $(\bar{Q}^\lambda(0), \bar{Z}_k^\lambda(0), k = 1, \dots, K)$ are bounded, then the process $(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda, \bar{D}_k^\lambda)$ is pre-compact as $\lambda \rightarrow \infty$, and hence any sequence has a converging subsequence. Denote any such *fluid limit* with a “bar” over the appropriate letters but with no superscript (for example, let $\bar{Q}(t)$ be a fluid limit of $\bar{Q}^\lambda(t)$). Note that

equations (2.3)–(2.8) imply that the following flow balance equations hold for *any* fluid limit:

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}_q(t) - \sum_{k=1}^K \bar{B}_k(t), \quad (2.17)$$

$$\bar{Z}_k(t) = \bar{Z}_k(0) + \bar{A}_k(t) + \bar{B}_k(t) - \mu_k \bar{T}_k(t), \quad k = 1, \dots, K, \quad (2.18)$$

$$\bar{T}_k(t) = \int_0^t \bar{Z}_k(s) ds \quad (2.19)$$

$$\bar{Y}(t) = \bar{Y}(0) + \mu t - \sum_{k=1}^K \mu_k \bar{T}_k(t), \quad (2.20)$$

$$\mu t = \bar{A}_q(t) + \sum_{k=1}^K \bar{A}_k(t), \quad (2.21)$$

$$\bar{T}_k(t) + \bar{I}_k(t) = q_k t. \quad (2.22)$$

Finally, for work conserving policies, conditions (2.9)–(2.11) imply:

$$\bar{Q}(t) \cdot \left(\sum_{k=1}^K (q_k - \bar{Z}_k(t)) \right) = 0, \quad (2.23)$$

$$\int_0^\infty \sum_{k=1}^K (q_k - \bar{Z}_k(t)) d\bar{A}_q(t) = 0, \quad (2.24)$$

and

$$\sum_{k=1}^K \int_0^\infty \bar{Q}(t) d\bar{I}_k(t) = 0. \quad (2.25)$$

The following proposition shows that for every sequence of work-conserving routing policies and for every fluid limit, the quantities $\bar{Q}(t)$ and $\bar{Z}_k(t)$, $k = 1, \dots, K$, remain constant if starting at time 0 from some appropriate initial conditions.

Proposition 2.1 (fluid limits). For $\lambda > 0$, let $\pi^\lambda := \pi(\lambda, N^\lambda) \in \Pi$ be a sequence of work-conserving policies (omitted from the following notation), and let $(\bar{Q}, \bar{Z}_k, \bar{Y}, \bar{A}, \bar{A}_k, \bar{A}_q, \bar{B}_k, \bar{T}_k, \bar{I}_k, \bar{D}_k)$ be a fluid limit of the processes associated with the system, as $\lambda \rightarrow \infty$. Recall that $q_k = \lim_{\lambda \rightarrow \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k} \mu$, $k = 1, \dots, K$, and suppose that $\bar{Q}(0) = 0$ and $\bar{Z}_k(0) = q_k$, $k = 1, \dots, K$. Then, $\bar{Q}(t) = 0$ and $\bar{Z}_k(t) = q_k$, $k = 1, \dots, K$, for all $t \geq 0$.

In addition to the fluid scaling, we introduce a more refined *diffusion* scaling defined as follows:

Diffusion Scaling. For $\lambda > 0$ and any fixed sequence of work conserving policies $\pi^\lambda = \pi(\lambda, N^\lambda) \in \Pi$ (omitted from the notation), define the centered and scaled process

$\vec{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ as follows:

$$X_1^\lambda(t) := \frac{Q^\lambda(t) + Z_1^\lambda(t) - N_1^\lambda}{\sqrt{N^\lambda}}, \quad (2.26)$$

and, for $k = 2, \dots, K$, let

$$X_k^\lambda(t) := \frac{Z_k^\lambda(t) - N_k^\lambda}{\sqrt{N^\lambda}}. \quad (2.27)$$

Note that for $k = 2, \dots, K$, $X_k^\lambda(t) \leq 0$ for all t , and that for all $k = 1, 2, \dots, K$, $[X_k^\lambda(t)]^- := -\min\{X_k^\lambda(t), 0\}$ corresponds to the number of idle servers, scaled by $1/\sqrt{N^\lambda}$. In addition, $[X_1^\lambda(t)]^+$ corresponds to the total queue length, again, scaled by $1/\sqrt{N^\lambda}$. Finally, let

$$\begin{aligned} X^\lambda(t) &= \sum_{k=1}^K X_k^\lambda(t) = \frac{Q^\lambda(t) + \sum_{k=1}^K Z_k^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} \\ &= \frac{Y^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \sqrt{N^\lambda}(\bar{Y}^\lambda(t) - 1). \end{aligned} \quad (2.28)$$

Note that $X^\lambda(\cdot)$ captures the fluctuations of order $\Theta(1/\sqrt{N^\lambda})$ of $\bar{Y}^\lambda(\cdot)$ about its fluid limit. Also, $[X^\lambda(t)]^-$ is the total number of idle servers, and $[X^\lambda(t)]^+ = [X_1^\lambda(t)]^+$ is the total queue length, both scaled by $1/\sqrt{N^\lambda}$. Finally, note that, from work conservation, if $X_k^\lambda(t) < 0$ for some k , then $X_1^\lambda(t) \leq 0$.

Finally, for all $\lambda > 0$, let $W^\lambda(t)$ be the virtual waiting time of an arbitrary customer who arrives to the system indexed by λ at time t . The scaled waiting time for $\lambda > 0$ is then defined as:

$$\hat{W}^\lambda(t) = \sqrt{N^\lambda} W^\lambda(t). \quad (2.29)$$

As will be shown later, in order for the diffusion scaling to have well defined limits, as $\lambda \rightarrow \infty$, we add the following assumption, in addition to (A1):

$$\sum_{k=1}^K \mu_k N_k^\lambda = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \text{or,} \quad \lim_{\lambda \rightarrow \infty} \frac{\sum_{k=1}^K \mu_k N_k^\lambda - \lambda}{\sqrt{\lambda}} = \delta, \quad (A2)$$

for some $\delta, 0 < \delta < \infty$.

Condition (A2) is a square-root safety staffing rule (similar to [32] and [12]). In particular, the condition $\delta > 0$ guarantees that the system is stable (or can be stable, under reasonable routing) for all λ large enough. Moreover, as is shown later (Proposition 4.6), it guarantees that under the appropriate routing, the fraction of delayed customers is less than 1. Note that (A2) does not specify how the added safety staffing is divided among server pools. In particular, it is possible that one server pool will have fewer servers than

the nominal allocation of $q_k N^\lambda$, while another will compensate for this deficit by having more than the nominal staffing. For $k = 1, \dots, K$, and $\lambda > 0$, let $-\infty < \delta_k^\lambda < \infty$ satisfy:

$$\delta_k^\lambda := \frac{\mu_k N_k^\lambda - a_k \lambda}{\sqrt{\lambda}}. \tag{2.30}$$

Then $\delta_k^\lambda \sqrt{\lambda}$ is the safety capacity associated with server pool k , beyond the nominal allocation of $a_k \lambda$. In particular, one can easily verify that $\delta_k^\lambda \geq 0$ if $a_k = 0$,

$$\delta_k^\lambda = o(\sqrt{\lambda}), \text{ as } \lambda \rightarrow \infty, \quad \forall k = 1, \dots, K, \tag{2.31}$$

and

$$\delta^\lambda := \sum_{k=1}^K \delta_k^\lambda \rightarrow \delta, \text{ as } \lambda \rightarrow \infty. \tag{2.32}$$

Note that we do not require the individual sequences $\{\delta_k^\lambda\}_{\lambda>0}$ to have a limit, for any value of $k = 1, \dots, K$. All that is assumed is that their sum converges to δ . The one exception to this rule is Proposition 4.3, in which the following additional condition is assumed to hold:

$$\theta := \lim_{\lambda \rightarrow \infty} \sum_{k=1}^K \frac{\delta_k^\lambda}{\mu_k} = \lim_{\lambda \rightarrow \infty} \sqrt{\lambda} \left(\frac{N^\lambda}{\lambda} - \frac{1}{\mu} \right), \text{ exists for some finite number } \theta. \tag{A3}$$

3. Optimal preemptive routing

In this section we describe a simple *preemptive* policy FSF_P ((preemptive) Faster Server First), which is optimal within the set of all non-anticipating, but possibly preemptive FIFO policies, with respect to the steady-state distribution of the total number of customers in the system. Section 4 will describe our proposed non-preemptive policy, FSF (Faster Server First), which is also simple, but is not necessarily optimal for any fixed size system. However, it is *asymptotically* optimal as the system grows large (that is, as $\lambda \rightarrow \infty$), in terms of the steady-state queue length and waiting time distributions.

Consider a fixed system with fixed arrival rate and staffing vector. Furthermore, consider the more general family of policies $\Pi_P \supseteq \Pi$, which is the family of all non-anticipating FIFO policies which are preemptive resume (the subscript P is for preemptive). What is meant by preemptive resume in the context of this paper is that a customer who is served by a particular server may be handed-off to another server, who will resume the service from the point it has been discontinued. In addition, we add the following restriction on each policy belonging to this family: It only performs a *finite*

number of *actions* in any finite time interval. An action refers to an assignment of a customer to a certain server, or a hand-off of a customer from one server to another.

Let $\tilde{\Pi}_P \subseteq \Pi_P$ be the family of policies in Π_P which also satisfy the following two properties: For any $\pi \in \tilde{\Pi}_P$ we have

1. *Faster servers are used first.* If $Z_k(t; \pi) < N_k$ then $Z_j(t; \pi) = 0$, for all $j < k$.
2. *Work conservation.* If $Z_1(t; \pi) + Z_2(t; \pi) + \cdots + Z_K(t; \pi) < N$ then $Q(t; \pi) = 0$.

One example of a policy in $\tilde{\Pi}_P$ is the policy FSF_P , which, like other policies in $\tilde{\Pi}_P$ uses faster servers first, and is work conserving; in addition, it only assigns a customer to a server either upon his / her arrival to the system or upon a service completion. Note the non-uniqueness of FSF_P due to the unspecified order of assignments of customer to servers in case more than one option exists. This non-uniqueness can be easily resolved, but does not introduce any complications because all the servers in the same pool are statistically identical. The following proposition establishes the optimality of FSF_P within Π_P .

Proposition 3.1 (Optimal preemptive routing). Consider the preemptive routing policy, FSF_P , that keeps the faster servers busy whenever possible. Then it is optimal in the sense that it stochastically minimizes the total number of customers in the system in steady-state ($t = \infty$) within the family of non-anticipating, possibly preemptive policies. In other words, for all $\pi \in \Pi_P$ and every weak limit $Y(\infty; \pi)$ of $Y(t; \pi)$, as $t \rightarrow \infty$ (or a subsequence thereof), we have $P\{Y(\infty; \pi) > y\} \geq P\{Y(\infty; \text{FSF}_P) > y\}$, for all $y \geq 0$.

Remark 3.1 (Dominance of FSF among work-conserving policies). Shanthikumar and Yao in [51] show that FSF is path-wise optimal among all work-conserving policies. The optimality we establish here is *weaker* in the sense that it only considers the process in steady-state, but is *stronger* in the sense that the optimality is within the set of all non-anticipating policies, including ones that are not work-conserving.

Proof. We prove the Proposition in two steps. The first step establishes that all the policies in $\tilde{\Pi}_P$ share the same steady-state distribution of the total number of customers in the system. The second step shows that any policy in Π_P is path-wise dominated by a policy in $\tilde{\Pi}_P$ in terms of the total number of customers in the system at any point of time (see Lemma 3.1). Both steps together establish that the steady-state distribution of the total number of customers in the system under FSF_P stochastically dominates the corresponding distribution associated with any other policy in Π_P .

Let π be an arbitrary policy in $\tilde{\Pi}_P$, and recall that $Y(t; \pi)$ corresponds to the total number of customers in the system at time t under π . The special properties of the family $\tilde{\Pi}_P$ make the process $Y(\cdot; \pi)$ a birth and death (B&D) Markov process with constant birth rates:

$$\lambda(y) \equiv \lambda, \quad \forall y \geq 0,$$

and a concave piecewise-linear death rate function:

$$\mu(y) = \begin{cases} y\mu_K & \text{if } y \leq N_K \\ (y - N_K)\mu_{K-1} + N_K\mu_K & \text{if } N_K < y \leq N_{K-1} + N_K \\ \cdot & \\ \cdot & \\ (y - (N_2 + \dots + N_K))\mu_1 & \text{if } N_2 + \dots + N_K < y \leq N \\ + N_2\mu_2 + \dots + N_K\mu_K & \\ N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K & \text{if } y > N. \end{cases} \tag{3.1}$$

In particular, the steady-state distribution of $Y(\cdot; \pi)$ exists (recall the stability assumption) and is unique, and the same for all policies in $\tilde{\Pi}_P$. The next lemma (step two of the proof of the proposition) establishes the path-wise dominance of policies in $\tilde{\Pi}_P$ within the larger family Π_P . □

Lemma 3.1. For any policy $\pi \in \Pi_P$, the process $Y(\cdot; \pi)$ which denotes the total number of customers in the system, is path-wise dominated by the total number of customers in the system process $Y(\cdot, \tilde{\pi})$ for some appropriately chosen policy $\tilde{\pi} \in \tilde{\Pi}_P$.

Corollary 3.1. Recall that $Q(t)$ is the queue length at time t , and let $W(t)$ be the virtual waiting time at time t . The preemptive routing policy, FSF_P , that always assigns customers to the faster servers first is also optimal in the sense that it stochastically minimizes the queue length and the waiting time in steady-state ($t = \infty$) within Π_P . In other words, for all $\pi \in \Pi_P$ and all weak limits $Q(\infty; \pi)$ and $W(\infty; \pi)$ of $Q(t; \pi)$ and $W(t; \pi)$, respectively, as $t \rightarrow \infty$ (or a subsequence thereof), we have

$$P\{Q(\infty; \pi) > q\} \geq P\{Q(\infty; \text{FSF}_P) > q\},$$

for all $q \geq 0$, and

$$P\{W(\infty; \pi) > w\} \geq P\{W(\infty; \text{FSF}_P) > w\},$$

for all $w \geq 0$.

Proof. The proof follows from Proposition 3.1 and the work conservation property of FSF_P . For the queue length, the proof directly follows from the relationships:

$$Q(t; \text{FSF}_P) = [Y(t; \text{FSF}_P) - N]^+, \quad \text{a.s.}$$

and

$$Q(t; \pi) \geq [Y(t; \pi) - N]^+, \quad \text{a.s.}$$

for all $t \geq 0$ and $\pi \in \Pi_P$ (the latter inequality is due to the fact that π may not be work-conserving).

For the virtual waiting time, consider a policy $\pi \in \Pi_P$, and suppose that there exists a steady-state distribution, $Y(\infty; \pi)$ for the total number of customers in the system. By conditioning on the state of $Y := Y(\infty; \pi)$ one can easily verify that if π is work conserving and FIFO then the steady-state distribution of $W := W(\infty; \pi)$ exists and it satisfies

$$W \stackrel{\mathcal{D}}{=} \sum_{i=1}^{[Y-N+1]^+} T_i, \quad (3.2)$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution, and T_i are iid exponential random variables with rate $\sum_{k=1}^K \mu_k N_k$, which are independent of Y . If π is FIFO but is not work conserving, then if the steady-state distribution of $W(\cdot; \pi)$ exists it satisfies

$$W(\infty; \pi) \stackrel{st}{\geq} \sum_{i=1}^{[Y-N+1]^+} T_i. \quad (3.3)$$

Since $\sum_{i=1}^{[Y-N+1]^+} T_i$ is an increasing function of Y (for any positive realization of the sequence $\{T_i\}$), the stochastic dominance of FSF_P with respect to Y implies that FSF_P also stochastically minimizes the steady-state waiting time.

Remark 3.2 (Steady-state distributions for the queue length and the waiting time). The proof of Corollary 3.1 suggests a way of computing the steady-state distributions of both the queue length and waiting time for any work-conserving policy $\pi \in \Pi$ (to be omitted for brevity). This computation is possible provided that one knows the steady-state distribution of Y , the total number of customers in the system. Observe that conditioned on the event $Y \geq N$, the process $Y - N$ has transition rates which are like those of an $M/M/1$ system with arrival rate λ and service rate $\sum_{k=1}^K \mu_k N_k$. Hence, since $Q(\infty) = [Y(\infty) - N]^+$ its distribution satisfies:

$$P(Q(\infty) = n) = \alpha \rho^n (1 - \rho), \quad n \geq 1 \quad (3.4)$$

where $\alpha = P(Y(\infty) \geq N)$. Similarly, due to the relationship (3.2), we have

$$\begin{aligned} P(W(\infty) > w) &= \sum_{n=0}^{\infty} P\left(\sum_{i=1}^{n+1} T_i > w\right) P(Y(\infty) = N + n) \\ &= \sum_{n=0}^{\infty} P\left(\sum_{i=1}^{n+1} T_i > w\right) \alpha \rho^n (1 - \rho) \\ &= \alpha e^{-(1-\rho)\left(\sum_{k=1}^K \mu_k N_k\right)w}, \quad \forall w \geq 0. \end{aligned} \quad (3.5)$$

In particular, $(W(\infty) | W(\infty) > 0) \sim \exp((1-\rho) \sum_{k=1}^K \mu_k N_k) = \exp(\sum_{k=1}^K \mu_k N_k - \lambda)$.

Remark 3.3 (State-space collapse under FSF_P). Note the *state-space collapse* associated with the policy FSF_P (and all other policies in $\tilde{\Pi}_P$). For a work conserving policy, the state-space is generally K dimensional. However, under this policy it is sufficient to know the total number of customers in the system in order to specify exactly how they are distributed between the server pools and the queue, as is demonstrated by the death rates (3.1). Hence, the state-space reduces to one dimension.

4. Faster server first (FSF) is asymptotically optimal

In this section we describe a simple non-preemptive policy FSF which is also work-conserving. This policy may be described simply as follows: Upon a customer arrival or a service completion, assign the first customer in the queue (or the one that has just arrived, if the queue is empty) to the fastest available server (which is the server with the largest index k). Judging by the literature on the slow server problem (e.g. [39]), this policy is not likely to be optimal. However, as we show in this section, it is *asymptotically* optimal as the arrival rate λ grows to ∞ and the number of servers per pool grow according to (A1) and (A2); the asymptotic optimality is in terms of the steady-state distribution of the queue length and the waiting time. The main premise of this section is the asymptotic optimality of FSF within the family of non-preemptive non-anticipating FIFO policies. This is summarized in Theorem 4.1 and proved at the end of this section via Propositions 3.1–4.6.

Theorem 4.1 (Asymptotic optimality of FSF). Consider a sequence of systems indexed by the arrival rate λ , that satisfy conditions (A1) and (A2). Then the non-preemptive policy FSF that assigns customers to the fastest server available whenever a customer arrives, or upon service completion, is asymptotically optimal within the set Π of all non-preemptive, non-anticipating FIFO policies. The asymptotic optimality is in terms of stochastic minimization of the steady-state distributions of the (centered and scaled) total number of customers in the system ($X^\lambda(\infty)$), the scaled queue length ($\hat{X}_0^\lambda(\infty) := Q^\lambda(\infty)/\sqrt{N^\lambda}$), and the scaled waiting time ($\hat{W}^\lambda(\infty)$), as $\lambda \rightarrow \infty$.

Remark 4.1 (Sojourn time optimization). Note that we focus our attention on optimality criteria which relate to delayed customers (namely, queue length and waiting time), rather than the total number of customers in the system, or the total sojourn time. If one is interested in the latter two as optimality criteria, then, within the asymptotic framework considered here, any work conserving policy would be asymptotically optimal. This is apparent from Proposition 2.1, where it was shown that any work conserving policy will result in the same fluid limit for the total number of customers in the system. The optimality criteria we consider are more refined, and hence, require more careful policy selection and analysis.

Remark 4.2 (Routing in “conventional” heavy-traffic). The asymptotic optimality of FSF within the family Π underlines an important difference between the QED regime, and the so-called conventional heavy-traffic. Teh and Ward [55] study a routing problem in a model similar to ours, with a single customer class, and two servers only, one of each type. Each server has its own queue, and the decision of which queue a customer should be routed to is made upon the customer’s arrival. For their model they show that a threshold policy similar to the one identified in [39] is also asymptotically optimal as the traffic intensity goes to 1, in terms of the total number of customers in the system. Moreover, they show that the asymptotically optimal threshold must grow logarithmically to infinity as the traffic intensity approaches 1. This is different in our case. Here, we show that one needs no thresholds (or can use thresholds of size 0) in order to achieve asymptotic optimality. Of course, in order to get a fair comparison between the two asymptotic regimes, one needs to look at comparable models (single queue vs. multiple queues - one per each server pool, and a growing number of servers vs. a fixed number of servers). This will not be broached further here.

To prove the asymptotic optimality of FSF, as $\lambda \rightarrow \infty$, we will show that as λ grows, the process $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ (recall the diffusion scaling in Section 2.1) under FSF becomes close to the same process under the preemptive policy FSF_p , and in the limit as $\lambda \rightarrow \infty$ the two processes coincide. Taking the limits as $t \rightarrow \infty$ we will also show that the corresponding steady-state processes become close, and hence, the optimality of FSF_p in steady-state (see Corollary 3.1) will imply the asymptotic optimality of FSF. The crucial step in the proof of the equivalence between the two processes is the state-space collapse of the process $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$ under FSF, into a one dimensional process as $\lambda \rightarrow \infty$. Recall, that such state-space collapse holds for every λ under FSF_p (Remark 3.3). When FSF is used, this is no longer true, but the state-space collapse is attained when $\lambda \rightarrow \infty$, as will be shown in Proposition 4.1 below.

4.1. State-space collapse

In this section we establish the state-space collapse result with respect to the policy FSF and the process $\bar{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), \dots, X_K^\lambda(\cdot))$. Since the policy here is fixed we omit FSF from all notation. Essentially, the state-space collapse result indicates that, as λ grows, the one-dimensional process $X^\lambda(\cdot)$ (see (2.28)) becomes sufficient in describing the whole K -dimensional process $\bar{X}^\lambda(\cdot)$. Specifically, we show that as $\lambda \rightarrow \infty$, all the faster servers (from pools $k = 2, \dots, K$) are constantly busy (or, more accurately, the number of idle servers in these pools is of order $o(\sqrt{N^\lambda})$), and the only possible idleness is within the slowest servers (pool 1). Hence, as λ grows, the processes $X_2^\lambda(\cdot), \dots, X_K^\lambda(\cdot)$ become identically zero, while the processes $X^\lambda(\cdot)$ and $X_1^\lambda(\cdot)$ become close. This result is presented in the following proposition:

Proposition 4.1 (State-space collapse). Suppose that conditions (A1) and (A2) hold as $\lambda \rightarrow \infty$, and that the work-conserving non-preemptive policy FSF is used. In addition,

suppose that $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, \dots, x_K)$, in probability, as $\lambda \rightarrow \infty$. Then for all $t > 0$ we have,

$$X_k^\lambda(t) \xrightarrow{p} 0, \text{ uniformly on compact intervals, as } \lambda \rightarrow \infty, \forall k \geq 2.$$

Remark 4.3 (Instantaneous transition into the collapsed state-space). Note that while $X_k(0)$ could be non-zero for some $k = 2, \dots, K$, according to Proposition 4.1, in the limit, $X_k(t) = 0$ for all $t > 0$. In other words, in the limit, there can be an instantaneous transition into the collapsed state-space. Intuitively, this happens because the faster servers face an increasing arrival rate which is well beyond their service capacity, and therefore, very quickly, they all become busy. Further note that the uniform convergence in the statement of the proposition is on compact intervals which do not include $t = 0$.

Remark 4.4 (State-space collapse in the V-model with priorities). Note the similarities and the differences between our state-space collapse result and the ones established in [2,3,46], for a multi-class, single server type system (the V-design) with service priority. The state-space collapse established in [2,3,46] essentially shows that whenever one customer class has priority in receiving service over the other classes, its respective queue length and waiting time are zero (both with the appropriate scaling). This is provided that the arrival rate into the lower priority classes is non-negligible. In such cases, the higher priority class “sees” a system which is in *light traffic*. Hence, the total queue length includes customers of lower priority classes only. In our system, faster servers get priority over slower servers. Hence, the number of idle fast servers and the amount of time such a fast server waits between two consecutive customers is zero (again, with the appropriate scalings). Here, the required condition for this to happen is that the number of slow servers is non-negligible. What the latter implies is that the faster servers experience a system which is *over-loaded*, and hence are continuously busy.

Remark 4.5 (State-space collapse for FSF_p in the QED regime). Proposition 4.1 is also true if the preemptive policy FSF_p is used. Here the proof is much simpler. (See the appendix for details).

4.2. Transient diffusion limit

In this section we establish the form of the diffusion limit of the scaled process \vec{X}^λ . Since the paper is mostly concerned with the optimization of steady-state performance measures, the transient diffusion limit can be regarded as an intermediate step in obtaining the steady-state limit (based on the equivalence between the transient diffusion limits of FSF_p and FSF). However, the form of diffusion process obtained in the limit is also interesting in its own right. Especially, when compared with the diffusion limit obtained

by Halfin and Whitt [32] for the $M/M/N$ system (see Remark 4.1). In a nutshell, this comparison reveals that given a certain level of service capacity, the multi-type system of this paper (stochastically) outperforms the single type $M/M/N$ system.

We note that the state-space collapse result of Proposition 4.1 essentially shows that it is sufficient to find the diffusion limit of the total count of customers (centered and scaled) X^λ . Denoting this limit by X , we have that the limit of X_k^λ , for $k \geq 2$, is identically zero, and the limit of X_1^λ is hence equal to X .

Proposition 4.2 (Transient diffusion limit). Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$, as $\lambda \rightarrow \infty$, for $k = 1, \dots, K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (A1) and (A2) hold, and that the policy FSF is used. Recall that $\mu_1 < \mu_2 < \dots < \mu_K$, and $\mu = [\sum_{k=1}^K a_k / \mu_k]^{-1}$. Then, $X^\lambda \Rightarrow X$, as $\lambda \rightarrow \infty$, where X is a diffusion process with an infinitesimal drift

$$m(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0, \\ -\delta\sqrt{\mu} - \mu_1 x & x < 0, \end{cases} \quad (4.1)$$

and infinitesimal variance

$$\sigma^2(x) = 2\mu. \quad (4.2)$$

Remark 4.6 (The infinitesimal drift). The drift term (4.1) has two components: $-\delta\sqrt{\mu}$ and $-\mu_1 x$. The first component is due to the difference between the overall available service capacity $\sum_{k=1}^K \mu_k N_k$ and the arrival rate. This difference is of order $\Theta(\sqrt{\lambda}) = \Theta(\sqrt{N})$. The second component is a drift that is due to idle servers. The state-space collapse result implies that, in the limit, only the slowest servers can be idle, and hence, this term is only affected by their service rate μ_1 .

Corollary 4.1 (Outperforming the single server type system). Consider, in comparison to our system, a sequence of systems with a single customer class and a single server pool, instead of K types. Suppose that all these servers have service rate μ . In addition, suppose that the sequence of arrival rates, $\{\lambda\}$, is identical for both models, and that the number of servers in the single pool model, $N^{S,\lambda}$, satisfies $N^{S,\lambda}\mu = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$, as $\lambda \rightarrow \infty$. That is, in both models the excess capacity is approximately equal to $\delta\sqrt{\lambda}$. For this model, let $Y^{S,\lambda}(t)$ be the total number of customers in the system at time t , and $X^{S,\lambda}(t) = (Y^{S,\lambda}(t) - N^{S,\lambda})/\sqrt{N^{S,\lambda}}$. Then, by [32], if $X^{S,\lambda}(0) \Rightarrow X^S(0)$, as $\lambda \rightarrow \infty$, then $X^{S,\lambda} \Rightarrow X^S$, as $\lambda \rightarrow \infty$, where X^S is a diffusion process with an infinitesimal drift

$$m^S(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0, \\ -\delta\sqrt{\mu} - \mu x & x < 0, \end{cases} \quad (4.3)$$

and infinitesimal variance

$$(\sigma^S)^2(x) = 2\mu. \quad (4.4)$$

In particular, the diffusion limits of both processes are of the same form, with the exception that $-\mu x$ replaces $-\mu_1 x$ in the drift component that applies when there are idle servers. This is to be expected, because, clearly, in the single server type model all servers are identical each with service rate μ . The comparison between the two diffusion processes indicates that the limiting process associated with the \wedge -design stochastically dominates the process associated with the I -design. This result is surprising because one would expect that more variance in service time would lead to worse performance. Additionally, it is an asymptotic result and examples can be identified in which the homogeneous server system actually outperforms its heterogenous server counterpart. However, in the QED regime, the FSF policy uses the servers efficiently, and therefore the heterogenous system is indeed better. The managerial implication from this result on the design and staffing of such large service systems is that heterogeneity in the server population is an asset and not a liability.

Remark 4.7 (Transient diffusion limit for FSF_p). Proposition 4.2 remains true if the preemptive policy FSF_p is used instead. The proof remains unchanged due to Remark 4.5.

We conclude this section by establishing the transient diffusion limit of the scaled waiting time process, which turns out to be a simple linear function of the corresponding limit of the queue length process. Note that here, in addition to the two assumptions (A1) and (A2), one also needs to assume that (A3) holds. In essence, (A3) guarantees that the arrival process converges to a diffusion limit when scaled and centered around μ and not just around λ/N^λ .

Proposition 4.3 (Transient diffusion limit of the waiting time process). Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$ as $\lambda \rightarrow \infty$, for $k = 1, \dots, K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (A1), (A2) and (A3) hold, and that the policy FSF is used. Then, $\hat{W}^\lambda := \sqrt{N^\lambda} W^\lambda \Rightarrow \hat{W}$, as $\lambda \rightarrow \infty$, where $\hat{W} = [X]^+ / \mu$, and X is the diffusion limit of X^λ as $\lambda \rightarrow \infty$, given in Proposition 4.2.

4.3. Stationary diffusion limit

In this section we establish that the stationary distributions of the process \vec{X}^λ , under both FSF_p and FSF, converge to the stationary distribution of \vec{X} , as $\lambda \rightarrow \infty$. In particular, this implies the asymptotic optimality of FSF within Π in terms of the steady-state queue length and waiting time, due to the optimality of FSF_p in Π_p .

First we spell out the stationary distribution of X , the limiting diffusion process, given in Proposition 4.2. Next we show that the stationary distribution of X^λ under FSF_p converges to this stationary distribution. Finally, we use the transient convergence results (Proposition 4.2 and Remark 4.7), and the sample path optimality of $\vec{\Pi}_p$ to establish the

convergence of the stationary distribution of X^λ under FSF. In all processes we use ∞ in place of the time argument to denote steady-state.

Proposition 4.4 (Stationary distribution of the diffusion process). Let $X(\cdot)$ be the diffusion process described in Proposition 4.2, with infinitesimal drift and variance as in (4.1) and (4.2). Then the steady-state distribution of X has a density $f(\cdot)$ given by:

$$f(x) = \begin{cases} \frac{\delta}{\sqrt{\mu}} \exp\{-\delta x/\sqrt{\mu}\} \alpha, & \text{if } x \geq 0, \\ \frac{\sqrt{\frac{\mu_1}{\mu}} \phi\left(\sqrt{\frac{\mu_1}{\mu}} x + \frac{\delta}{\sqrt{\mu_1}}\right)}{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right)} (1 - \alpha), & \text{if } x < 0, \end{cases} \quad (4.5)$$

where $\alpha := \alpha(\delta/\sqrt{\mu_1}) = [1 + \frac{\delta/\sqrt{\mu_1} \Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})}]^{-1} = P\{X(\infty) \geq 0\}$.

Remark 4.8 (Slow servers are good for wait minimization). Note that α depends on the model parameters only through δ and μ_1 . In particular, it is an increasing function of μ_1 . Proposition 4.6 later shows that under FSF, the steady-state distribution of the total number in the system (centered and scaled) converges to the distribution (4.5) as $\lambda \rightarrow \infty$. In view of this, α can be regarded as the limit of the steady-state waiting probability. This interpretation gives rise to an interesting insight; given a fixed total service capacity of $\lambda + \delta\sqrt{\lambda}$ one is able to *decrease* the waiting probability by choosing to staff the system with a non-negligible fraction of *slow* servers. This is to be expected given the well known tradeoff between a single fast server and multiple slower ones (see for example [50]). To the best of our knowledge, this is the first time such a result is shown for a heterogeneous server system.

Remark 4.9 (Heterogeneity is good for wait minimization). Recall that in Corollary 4.1 it is established that the heterogeneous system outperforms its homogeneous counterpart if in the latter the service rate of each server is equal to μ , which was given in (2.13). The comparison there is path-wise for the transient diffusion process. If one is concerned with steady-state waiting probability only, a stronger statement can be made; in this case the waiting probability of the heterogeneous system is less than the waiting probability of a homogeneous system for *any* service rate which is greater than μ_1 , as long as the total service capacity is the same. This can be seen by comparing the expression for α in Proposition 4.4, and the expression for the waiting probability in the homogeneous system given by Halfin and Whitt in [32].

We now turn to showing that under the preemptive policy FSF_p , the stationary distribution of $X^\lambda(\cdot)$ weakly converges to the stationary distribution of X , given in (4.5). Note that this convergence does not immediately follow from Remark 4.7 and Proposition 4.4. In particular, a double limit interchange (as both time and arrival rate go to infinity) needs to be justified. Recall that the process $X^\lambda(\cdot)$ under FSF_p admits a

state-space collapse. In particular, it is sufficient to know the total number of customers in the system, $Y^\lambda(t)$, in order to know the whole $K + 1$ dimensional state space. In addition, the process $Y^\lambda(\cdot)$ is a B&D process with birth rates $\lambda(y) = \lambda$ and death rates $\mu^\lambda(y)$ as given in (3.1). Under conditions (A1) and (A2) the system is stable for all λ , and the stationary distribution is given by $p_n^\lambda := P(Y^\lambda(\infty) = n) = p_0^\lambda \pi_n^\lambda, n = 0, 1, \dots$, where $\pi_n^\lambda = \frac{\lambda^n}{\prod_{i=1}^n \mu^\lambda(i)}$, $n = 0, 1, \dots$, and $p_0^\lambda = [\sum_{n=0}^\infty \pi_n^\lambda]^{-1}$. Clearly, the stationary distribution of $X^\lambda = \frac{Y^\lambda - N^\lambda}{\sqrt{N^\lambda}}$, can be easily obtained from the stationary distribution of Y^λ .

Proposition 4.5 (Convergence of the preemptive process in steady-state). Suppose that conditions (A1) and (A2) hold, and that the preemptive policy FSF_P is used. Then the stationary distribution of X^λ weakly converges to the stationary distribution of X given in (4.5), as $\lambda \rightarrow \infty$.

Remark 4.10 (Convergence of the K -dimensional processes). Note that Proposition 4.5 also implies the weak convergence of the stationary distribution of \bar{X}^λ to \bar{X} , which are both K -dimensional processes. This is due to the state-space collapse that holds, in fact, for all $\lambda > 0$ (see Remark 3.3), as well as in the limit as $\lambda \rightarrow \infty$.

In order to establish the asymptotic optimality of FSF with respect to the queue length distribution in steady-state, we need to show the convergence of the steady-state distribution of X^λ under FSF to the steady-state distribution of X . We have already shown in Proposition 4.2 that if $X_k^\lambda(0) \Rightarrow X_k(0)$ for all $k = 1, \dots, K$, then $X^\lambda(\cdot) \Rightarrow X(\cdot)$ for $0 \leq t < \infty$. Our goal is to show that this convergence is also true at $t = \infty$. This result is stated in the next proposition.

Proposition 4.6 (Convergence of the non-preemptive process in steady-state). Suppose that conditions (A1) and (A2) hold, and that the non-preemptive policy FSF is used. Then the stationary distribution of X^λ exists for all λ , and it weakly converges to the stationary distribution of X given in (4.5), as $\lambda \rightarrow \infty$.

We are now finally in a position to prove the asymptotic optimality of FSF which is stated in Theorem 4.1. See details in the appendix.

Remark 4.11 (Sufficiency of optimality with respect to the waiting probability). The proof of Theorem 4.1 essentially shows that in order to establish asymptotic optimality of the waiting time for our model in the QED regime, it suffices to show asymptotic optimality with respect to $\tilde{\alpha}$, the probability that there are at least N^λ customers in the system. For work conserving FIFO policies this implies that asymptotic optimality with respect to the waiting time is *equivalent* to the asymptotic optimality with respect to the waiting probability (both in steady-state). Figure 2 shows a diagram of the asymptotic optimality relationships between the four entities included in the proof of Theorem 4.1.

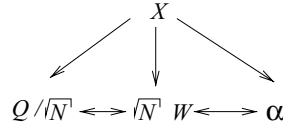


Figure 2. Asymptotic optimality equivalence relationships for work conserving FIFO policies. These relationships are elaborated on in Remark 4.11.

We conclude this section by presenting a lemma which provides a simple relationship between the steady-state queue length and waiting time distributions for work conserving policies, which we refer to as an asymptotic distributional Little's law. This relationship is of the same form as the one shown in Proposition 4.3 for the transient limits of the queue length and waiting time processes.

Lemma 4.1 (Asymptotic distributional Little's law). Suppose that conditions (A1) and (A2) hold, and consider a sequence of policies $\{\pi^\lambda\} \subseteq \Pi$, $\lambda > 0$. Let $X(\infty)$, $\hat{X}_0(\infty)$, and $\hat{W}(\infty)$, be the weak limits as $\lambda \rightarrow \infty$ for the steady-state random variables: $X^\lambda(\infty; \pi^\lambda)$, $Q^\lambda(\infty; \pi^\lambda)/\sqrt{N^\lambda}$, and $\sqrt{N^\lambda}W^\lambda(\infty; \pi^\lambda)$, respectively (assuming that those steady-state distributions, and limits as $\lambda \rightarrow \infty$ exist and are unique). Then,

$$\hat{W}(\infty) \stackrel{st}{\geq} \frac{[X(\infty)]^+}{\mu}, \quad (4.6)$$

and if π^λ is work conserving for all λ large enough, then

$$\hat{W}(\infty) \stackrel{D}{=} \frac{[X(\infty)]^+}{\mu} = \frac{\hat{X}_0(\infty)}{\mu}. \quad (4.7)$$

5. Conclusions and extensions

We study the routing problem in large scale service systems with a single customer class and multiple server types. For this model, the FSF policy is proposed which assigns customers to the fastest available server. This policy is shown to be asymptotically optimal as demand and the number of servers grow large according to the QED regime. Our results show that, in this regime, a square-root safety service capacity is equivalent to a non-trivial waiting probability (in the open interval $(0,1)$). In a follow-up paper [4] we show that this equivalence has an important implication on staffing of such large-scale heterogeneous server systems. In a nutshell, if one is interested in maintaining a certain waiting probability $\alpha \in (0, 1)$, then it is necessary for the total service capacity $N_1\mu_1 + N_2\mu_2 + \dots + N_K\mu_K$ to be of the form $\lambda + \delta\sqrt{\lambda}$ for some positive number $\delta = \delta(\alpha)$. In particular, if this waiting probability requirement is the constraint of a staffing optimization problem, then the feasible region is *linear*. This simple form of the

feasible region facilitates the optimization of various staffing costs subject to waiting probability constraint.

Our proposed routing scheme is useful in other settings as well. In [18] a model is considered for a help-desk type call-center in which heterogeneous servers differ not only in their service rate μ but also in the probability of a call resolution p . The authors in [18] suggest that a sensible routing rule there is the so called $p\mu$ rule which assigns calls to the available server with the highest value of the product $p \cdot \mu$ first. Provided that callers with an unresolved problem call the center back right away, it is straightforward to extend our analysis to show that indeed the $p\mu$ is asymptotically optimal in a setting where there are multiple server pools, each characterized by a unique pair of p and μ .

Several additional extensions are interesting as directions for further research. These include:

1. Considering the modified FSF policy, proposed in [48] and [41] where customers are served by a particular server only if all the faster servers are busy *and* the queue length exceeds a certain threshold. How do the threshold values affect the performance of the system?
2. Considering a system in which the queues are attached to each server pool (that is, customers are routed to a server pool upon arrival). What is the optimal routing in this case? How does this setting affect system performance? Stolyar in [53] studies this problem under “classical” heavy-traffic. There, he shows that under appropriate policies, the separate queues system and the single queue one are equivalent. This setting is a natural to consider in distributed (multi-location) call-center, and therefore, it would be interesting to generalize the results of [53] to the QED regime.
3. With customers abandonment of the system if their service does not start after a certain time, what is the optimal routing in this case?
4. Do any of the insights gained from studying the \wedge -design and the V -design (studied in [30]) extend to more complex system topologies (such as N -design)?
5. Finally, in practice, even if a call center has a \wedge -design, it is not practical, and is also unfair, to keep the faster servers busy at all times, while the slower servers get more idle time. Hence, in many call centers, the service rep with the longest idle time is the one to handle the next call. This latter rule is clearly sub-optimal, though, with respect to the performance measures studied in this paper. It is, therefore, natural to study the extension of this paper to the question of optimizing system performance subject to some fairness criteria with respect to servers work utilization. □

Acknowledgement

I thank Avi Mandelbaum for sharing his knowledge, ideas and enthusiasm for service systems research. I am also grateful to Rami Atar, Assaf Zeevi and Tolga Tezcan for assistance with the asymptotic optimality proof. Furthermore, I appreciate the thoughtful

comments made by the anonymous referees and AE who have helped improve the paper. Finally, most of this work was done while I was visiting the Industrial Engineering and Management Faculty at the Technion. I am thankful for their hospitality.

6. Appendix: Proofs

Proof of Proposition 2.1. Let $f(t) = |\bar{Y}(t) - 1| = |\sum_{k=1}^K (\bar{Z}_k(t) - q_k) + \bar{Q}(t)|$, then $f(t) \geq 0$ and $f(t) = 0$ if and only if $\bar{Q}(t) = 0$ and $\bar{Z}_k(t) = q_k$ for all $k = 1, \dots, K$. By an argument similar to lemma 2.4.5 of [17], and from the fact that $f(\cdot)$ is absolutely continuous, it is sufficient to show that whenever $t \geq 0$ is such that f is differentiable at t , we have $\dot{f}(t) \leq 0$. Suppose that t is such that $\bar{Y}(t) \geq 1$. Then, by (2.23) $\bar{Z}_k(t) = q_k$, for all k . In particular, if f is differentiable at t , then

$$\dot{f}(t) = \dot{\bar{Y}}(t) = \mu - \sum_{k=1}^K \mu_k \bar{Z}_k(t) = \mu - \sum_{k=1}^K \mu_k q_k = 0.$$

If t is such that $\bar{Y}(t) < 1$, then $\bar{Z}_k(t) < q_k$ for at least one k , and hence, by (2.23), $\bar{Q}(t) = 0$. If f is differentiable at t then,

$$\dot{f}(t) = -\dot{\bar{Y}}(t) = \sum_{k=1}^K \mu_k \bar{Z}_k(t) - \mu < \sum_{k=1}^K \mu_k q_k - \mu = 0.$$

□

Proof of Lemma 3. For simplicity, we prove the Lemma for the special case $K = 2$. The general case follows similarly. The proof is based on sample-path coupling arguments. Suppose that the j th customer to arrive into the system arrives at time t_j and has a service requirement of η_j . The interpretation of η_j is that if this customer is served exclusively by a server of pool k , $k = 1, 2$, her service time is η_j/μ_k . Note that the sequence $\{(t_j, \eta_j)\}_{j=1}^{\infty}$ is random. In fact, given the routing policy, this sequence is the only random element in the system. Consider an arbitrary policy $\pi \in \Pi_p$, and focus only the customers $i = 1, 2, \dots, n$, for some finite number n (the lemma will follow by induction on n). Fix a sample-path of $\{(t_j, \eta_j)\}_{j=1}^{\infty}$. Suppose that on this sample-path, for some $i \in \{1, \dots, n\}$, the customers $j = i + 1, \dots, n$, satisfy the following two properties which agree with the family $\tilde{\Pi}_p$:

1. *Use fast servers first.* During the sojourn time of customer j in the system it is never served by a slow server if there is a fast server available.
2. *Work conservation.* During the sojourn time of customer j in the system it is never held in the queue if there is any idle server.

Let $d_j(\pi)$ be the departure time of customer j from the system according to the policy π . Also let $D_n(\pi)$ be the time by which all the customers $j = 1, \dots, n$ have departed.

Let $S = \{0 \leq s_1 < s_2 < \dots < s_M = D_n(\pi)\}$ be the set of all event time points for the policy π . In particular this set includes all arrival times, departure times and action times such as assignment of customers to servers or hand-offs of customers from one server to another. According to the definition of Π_P , M has to be finite.

We will construct a new policy $\pi' \in \Pi_P$ which will satisfy properties 1 and 2 for $j = i, i + 1, \dots, n$, which will have at most as many total number of customers in the system at any time $t \geq 0$ as π . By backward induction on i , this will complete the proof of the lemma. Let l_0 be such that $s_{l_0} = t_i$. Now perform the procedure $FIX(i, l_0)$ defined as follows:

Procedure $FIX(j, l)$: For customer j and time interval $[s_l, s_{l+1})$ do the following:

- If property 1 is violated for customer j during the interval $[s_l, s_{l+1})$, that is, the customer is served by a slow server and there is a fast server available, assign this customer to this fast server for the duration of this interval.
- If property 2 is violated with respect to customer j during the interval $[s_l, s_{l+1})$, that is, customer j is held in the queue and there are idle servers, assign this customer to a fast server if available. Otherwise, assign this customer to a slow server.
- If none of these properties is violated do nothing.
- If, after performing the previous steps of this procedure, customer j has departed during the interval $[s_l, s_{l+1})$, add its new departure time d_j to S , and renumber the other points in S (including the value of M) accordingly.

Repeat this process for customer i and $l = l_0 + 1, \dots, M - 1$. Note that the set S may only change by adding the new departure time of customer i , $d_i(\pi')$ in the appropriate place in the sequence. Therefore the sequence S remains finite. Also, note that after performing the procedure the total number of customers in the system at any point in time is at most the number it was before, because only customer i is handled differently, and his service time may only get shorter. Finally, note that after performing the procedure $FIX(i, l)$, for $l = l_0, \dots, M - 1$, customer i satisfies Properties 1 and 2 for all $t \geq 0$.

In order to complete the improvement of the policy π , one needs to examine the effect of the procedure performed on customer i over the customers $i + 1, \dots, n$. In this respect, note that the procedure $FIX(i, l)$ may not induce a violation of either properties 1 and 2 with respect to customers $i + 1, \dots, n$ as long as customer i is in the system. However, if customer i now departs earlier than before, it may free up some servers, and hence some of these customers may violate one or both of these properties. To take care of these violations, first perform the procedure $FIX(j, l)$ for $j = i + 1$ and $l = l_1, \dots, M - 1$ with l_1 satisfying $s_{l_1} = d_i(\pi')$. Note that customer i is not affected at all, because the procedure starts with her departure. Proceed with the same procedure for $j = i + 2, \dots, n$ in increasing order of the index j , always starting with the interval that begins with the new departure time of customer $j - 1$. One can easily verify that at the end of the process we have a new policy π' that:

- a. Satisfies properties 1. and 2. for customers $j = i, i + 1, \dots, N$.
- b. $Y(t; \pi') \leq Y(t; \pi)$ for all $t \geq 0$.
- c. The number of action points is finite in any finite interval.

□

Proof of Proposition 4.1. Our goal is to establish that under the conditions of the proposition, for all $\epsilon > 0$ and $T > 0$, as $\lambda \rightarrow \infty$,

$$P\left(\sup_{0 < t \leq T} \left| \sum_{k=2}^K X_k^\lambda(t) \right| > \epsilon\right) \rightarrow 0, \quad \text{or} \quad P\left(\inf_{0 < t \leq T} \sum_{k=2}^K X_k^\lambda(t) < -\epsilon\right) \rightarrow 0. \quad (\text{A.1})$$

We prove the Proposition for $K = 2$. The general case follows similarly. For $K = 2$, (A.1) translates into

$$P\left(\sup_{0 < t \leq T} \left| X_2^\lambda(t) \right| > \epsilon\right) \rightarrow 0, \quad \text{or} \quad P\left(\inf_{0 < t \leq T} X_2^\lambda(t) < -\epsilon\right) \rightarrow 0. \quad (\text{A.2})$$

We claim that in order to establish (A.2) it is sufficient to show the existence of a sequence $\{b^\lambda\}$, with $b^\lambda \rightarrow 0$ as $\lambda \rightarrow \infty$, such that

$$\lim_{\lambda \rightarrow \infty} P\left(\inf_{0 < t \leq T} X_2^\lambda(t + b^\lambda) < -\epsilon\right) = 0. \quad (\text{A.3})$$

The sufficiency of (A.3) has been established in [2]. Essentially, it follows from a random time change argument (see [29, Prop. 5]). Recall that x_2 is the weak limit of $X_2^\lambda(0)$ as $\lambda \rightarrow \infty$. Then, since x_2 may be less than $-\epsilon$ we consider the interval $(0, b^\lambda]$ in proving (A.3) separately from $(b^\lambda, T + b^\lambda]$. Specifically, we show that both summands on the right hand side of the following inequality converge to 0 as $\lambda \rightarrow \infty$.

$$\begin{aligned} P\left(\inf_{0 < t \leq T} X_2^\lambda(t + b^\lambda) < -\epsilon\right) &\leq P\left(X_2^\lambda(b^\lambda) < -\frac{\epsilon}{2}\right) \\ &\quad + P\left(\inf_{0 < t \leq T} X_2^\lambda(t + b^\lambda) < -\epsilon \mid X_2^\lambda(b^\lambda) \geq -\frac{\epsilon}{2}\right) \end{aligned} \quad (\text{A.4})$$

□

This is proved in the next two lemmas.

Lemma A.1. Suppose that $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, \dots, x_K)$, in probability, as $\lambda \rightarrow \infty$. Then, under the conditions of Proposition 4.1, there exists a sequence $\{b^\lambda\}_{\lambda > 0}$ (which is a function of \vec{x}) with $b^\lambda \rightarrow 0$ as $\lambda \rightarrow \infty$, such that

$$(X_2^\lambda(b^\lambda), \dots, X_K^\lambda(b^\lambda)) \xrightarrow{P} 0, \quad \text{as } \lambda \rightarrow \infty. \quad (\text{A.5})$$

Proof. The lemma is proved for $K = 2$. The proof for the general case is similar. To prove the lemma we define a new fluid-scale process (different from \bar{Z}), which is identical to the diffusion-scale process, except that time is scaled by $1/\sqrt{N^\lambda}$. We will show that the fluid limit reaches the goal of $x_2 = 0$ in finite time, and hence, the diffusion limit will get there instantaneously. This argument mimics the one proposed by Bramson in [13], although does not make a direct use of his results.

Let $\vec{X}^\lambda(t) := \vec{X}^\lambda(t/\sqrt{N^\lambda})$, then

$$\vec{X}^\lambda(t) = (\vec{X}_1^\lambda(t), \vec{X}_2^\lambda(t)) = \left(\frac{Q^\lambda(t/\sqrt{N^\lambda}) + Z_1^\lambda(t/\sqrt{N^\lambda}) - N_1^\lambda}{\sqrt{N^\lambda}}, \frac{Z_2^\lambda(t/\sqrt{N^\lambda}) - N_2^\lambda}{\sqrt{N^\lambda}} \right),$$

and note that $\vec{X}_2^\lambda(0) = \vec{X}_2^\lambda(0)$. Hence, if $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, x_2)$ as $\lambda \rightarrow \infty$, then, we also have $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, x_2)$ as $\lambda \rightarrow \infty$. We show that if $x_2 < 0$ then there exists $s^* = s^*(x_2)$ such that

$$\vec{X}_2^\lambda(s^*) \xrightarrow{P} 0, \quad \text{as } \lambda \rightarrow \infty. \tag{A.6}$$

Setting $b^\lambda = s^*/\sqrt{N^\lambda}$ will then complete the proof.

The proof follows three steps:

1. Establishing that $\vec{X}(t) = x_1 + x_2$ for all $t \geq 0$, for all fluid limits \vec{X} of \vec{X}^λ .
 2. Establishing the existence of a fluid limit \vec{X}_2 of \vec{X}_2^λ .
 3. Finding s^* such that $\vec{X}_2(s^*) = 0$.
1. To prove (A.6) consider the sequence of initial conditions $X_1^\lambda(0) = x_1$ and $X_2^\lambda(0) = x_2 < 0$. Recall the definitions of Section 2.1, and let $\tilde{T}_k^\lambda(t) = \frac{T_k^\lambda(t/\sqrt{N^\lambda})}{\sqrt{N^\lambda}}$, $k = 1, 2$. Note that for $k = 1, 2$ the process $T_k^\lambda(\cdot)$ is uniformly Lipschitz with constant N_k^λ , and thus $\tilde{T}_k^\lambda(\cdot)$ is Lipschitz with constant $N_k^\lambda/N^\lambda \leq 1$. Hence, there exists an increasing subsequence λ_j for which $\tilde{T}_k^{\lambda_j}(\cdot) \rightarrow \tilde{T}_k(\cdot)$ as $j \rightarrow \infty$, where \tilde{T}_k is a limiting allocation process, and the convergence is almost surely (a.s.), uniformly on compact intervals (u.o.c). Without loss of generality assume that the whole sequence converges. Using the functional strong law of large numbers, (2.14) and the key renewal theorem we have that as $\lambda \rightarrow \infty$,

$$\frac{A^\lambda(s/\sqrt{N^\lambda})}{\sqrt{N^\lambda}} \rightarrow \mu s \quad \text{and} \quad \frac{D_k(T_k^\lambda(s/\sqrt{N^\lambda}))}{\sqrt{N^\lambda}} \rightarrow \mu_k \tilde{T}_k(s), \quad \text{a.s., u.o.c.}$$

Now, note that

$$\begin{aligned}\tilde{X}^\lambda(s) &= \tilde{X}_1^\lambda(s) + \tilde{X}_2^\lambda(s) \\ &= x_1 + x_2 + \frac{A^\lambda(s/\sqrt{N^\lambda})}{\sqrt{N^\lambda}} - \sum_{k=1}^2 \frac{D_k(T_k^\lambda(s/\sqrt{N^\lambda}))}{\sqrt{N^\lambda}} \\ &\rightarrow \mu s - \mu_1 \tilde{T}_1(s) - \mu_2 \tilde{T}_2(s).\end{aligned}$$

To find $\tilde{T}_1(s)$ and $\tilde{T}_2(s)$, note that $\tilde{T}_1(s) \leq q_1 s$ and $\tilde{T}_2(s) \leq q_2 s$, with an equality in both simultaneously, if and only if $\tilde{T}_1(s) + \tilde{T}_2(s) = s$. But, notice also that,

$$\begin{aligned}\tilde{T}_1^\lambda(s) + \tilde{T}_2^\lambda(s) &= \int_0^{s/\sqrt{N^\lambda}} \frac{Z_1^\lambda(\tau) + Z_2^\lambda(\tau)}{\sqrt{N^\lambda}} d\tau \\ &= s + \frac{1}{\sqrt{N^\lambda}} \int_0^s \left(\tilde{X}^\lambda(\tau) - \frac{Q(\tau/\sqrt{N^\lambda})}{\sqrt{N^\lambda}} d\tau \right) \rightarrow s, \quad \text{as } \lambda \rightarrow \infty.\end{aligned}$$

Therefore, we have

$$\tilde{X}(s) = x_1 + x_2 + \mu s - \mu_1 q_1 s - \mu_2 q_2 s = x_1 + x_2. \quad (\text{A.7})$$

- Note that if $x_2 < 0$, then $x_1 \leq 0$ (work conservation), and hence (A.7) implies that $\tilde{X}(s) < 0$ for all s , which implies that $Q^\lambda(s/\sqrt{N^\lambda}) = 0$ for all λ large enough. Specifically, $B_1^\lambda(s) + B_2^\lambda(s) = 0$, for all s and all λ large enough (no queue implies only external arrivals to the servers). Note that since $A_2^\lambda(s) \leq A^\lambda(s)$ for all s , there is also an increasing subsequence λ_j such that

$$\frac{A_2^\lambda(s/\sqrt{N^{\lambda_j}})}{\sqrt{N^{\lambda_j}}} \rightarrow \tilde{A}_2(s), \quad \text{as } j \rightarrow \infty,$$

(WLOG, assume that λ_j is the whole sequence). Hence, we have,

$$\begin{aligned}\tilde{X}_2^\lambda(s) &= \tilde{X}_2(0) + \frac{A_2^\lambda(s/\sqrt{N^\lambda})}{\sqrt{N^\lambda}} + \frac{B_2^\lambda(s/\sqrt{N^\lambda})}{\sqrt{N^\lambda}} - \frac{D_2(T_2^\lambda(s/\sqrt{N^\lambda}))}{\sqrt{N^\lambda}} \\ &\rightarrow \tilde{X}_2(s) = x_2 + \tilde{A}_2(s) - \mu_2 q_2 s, \quad \text{as } \lambda \rightarrow \infty.\end{aligned} \quad (\text{A.8})$$

- Let $s^*(x_2) = \inf\{s \geq 0 \mid \tilde{X}_2(s) = 0\}$ (where, $s^*(x_2) = \infty$ if $\tilde{X}_2(s) < 0$ for all s). Then for all $0 \leq s \leq s^*(x_2)$, we have $\tilde{X}_2(s) < 0$, and in particular, according to FSF, $\tilde{A}_2(s) = \tilde{A}(s)$ (all arrivals join the fast server pool, as long as such servers are available). Hence, for all $0 \leq s \leq s^*(x_2)$, (A.8) implies that, as $\lambda \rightarrow \infty$,

$$\tilde{X}_2^\lambda(s) \rightarrow \tilde{X}_2(s) = x_2 + \tilde{A}_2(s) - \mu_2 q_2 s = x_2 + \tilde{A}(s) - \mu_2 q_2 s = x_2 + (\mu - \mu_2 q_2) s.$$

Solving for $\tilde{X}_2(s^*(x_2)) = 0$ we get that $s^*(x_2) = \frac{[x_2]^-}{\mu - \mu_2 q_2}$. In particular, the case of $s^*(x_2) = \infty$ is ruled out, because $q_2 < 1$ (recall our assumption that $a_1 > 0$, hence, $q_1 > 0$ as well).

□

Lemma A.2. Suppose that $\vec{X}^\lambda(0) \rightarrow \vec{X}(0) = \vec{x} = (x_1, \dots, x_K)$, in probability, as $\lambda \rightarrow \infty$. Then, under the conditions of Proposition 4.1, and for every sequence $\{b^\lambda\}_{\lambda>0}$ such that $\lim_{\lambda \rightarrow \infty} b^\lambda = 0$ we have for all $\epsilon > 0$

$$P\left(\inf_{0 < t \leq T} \sum_{k=2}^K X_k^\lambda(t + b^\lambda) < -\epsilon \mid \sum_{k=2}^K X_k^\lambda(b^\lambda) \geq -\epsilon/2\right) \rightarrow 0, \text{ as } \lambda \rightarrow \infty.$$

Proof. We prove the lemma for $K = 2$. The general case is similar. Fix $\epsilon > 0$ and a sequence $\{b^\lambda\}_{\lambda>0}$ such that $\lim_{\lambda \rightarrow \infty} b^\lambda = 0$. Suppose that $X_2^\lambda(b^\lambda) \geq -\epsilon/2$, and let

$$\tau^\lambda = \inf\{t \geq b^\lambda : X_2^\lambda(t) < -\epsilon\},$$

and

$$\tau'^\lambda = \sup\left\{b^\lambda \leq t \leq \tau^\lambda : X_2^\lambda(t) \geq -\frac{\epsilon}{2}\right\}.$$

Under FSF, for $\tau'^\lambda \leq t \leq \tau^\lambda$ all new arrivals join server pool 2, so that

$$\begin{aligned} X_2^\lambda(t) &= X_2^\lambda(\tau'^\lambda -) \\ &\quad + \frac{1}{\sqrt{N^\lambda}}[A^\lambda(t) - A^\lambda(\tau'^\lambda -)] \\ &\quad - \frac{1}{\sqrt{N^\lambda}}[D_2(T_2^\lambda(t)) - D_2(T_2^\lambda(\tau'^\lambda -))], \end{aligned} \quad (\text{A.9})$$

where A , T_2 and D_2 are the arrival, service allocation and departure processes defined in Section 2. In particular, considering the extreme case where all the servers in pool 2 are busy during $[\tau'^\lambda, t]$, we have,

$$\begin{aligned} X_2^\lambda(t) &\geq X_2^\lambda(\tau'^\lambda -) \\ &\quad + \frac{1}{\sqrt{N^\lambda}}[A^\lambda(t) - A^\lambda(\tau'^\lambda -)] \\ &\quad - \frac{1}{\sqrt{N^\lambda}}[D_2(T_2^\lambda(\tau'^\lambda -) + N_2^\lambda(t - \tau'^\lambda)) - D_2(T_2^\lambda(\tau'^\lambda -))]. \end{aligned} \quad (\text{A.10})$$

Therefore, we can write,

$$\begin{aligned}
& P\{\inf_{b^\lambda \leq t \leq T+b^\lambda} X_2^\lambda(t) < -\epsilon\} \leq \\
& P\{\inf_{b^\lambda \leq s < t \leq T+b^\lambda, 0 \leq u \leq s} \frac{1}{\sqrt{N^\lambda}}[A^\lambda(t) - \lambda t - A^\lambda(s) + \lambda s] + \frac{\lambda(t-s)}{\sqrt{N^\lambda}} \\
& \quad - \frac{1}{\sqrt{N^\lambda}}[D_2(N_2^\lambda(u+t-s)) - N_2^\lambda \mu_2(u+t-s)] \\
& \quad - \frac{1}{\sqrt{N^\lambda}} \mu_2 N_2^\lambda(u+t-s) \\
& \quad + \frac{1}{\sqrt{N^\lambda}}[D_2(N_2^\lambda u) - N_2^\lambda \mu_2 u] + \frac{\mu_2 N_2^\lambda u}{\sqrt{N^\lambda}} < -\frac{\epsilon}{2}\} = \quad (\text{A.11}) \\
& P\{\inf_{b^\lambda \leq s < t \leq T+b^\lambda, 0 \leq u \leq s} -\sqrt{\frac{a_2 \mu}{\mu_2}}(\hat{D}_2^\lambda(u+t-s) - \hat{D}_2^\lambda(u)) + (\hat{A}^\lambda(t) - \hat{A}^\lambda(s)) \\
& \quad - a_2 \mu \sqrt{N^\lambda}(t-s) + \frac{\lambda(t-s)}{\sqrt{N^\lambda}} + o(\sqrt{N^\lambda}) < -\frac{\epsilon}{2}\} = \\
& P\{\inf_{b^\lambda \leq s < t \leq T+b^\lambda, 0 \leq u \leq s} -\sqrt{\frac{a_2 \mu}{\mu_2}}(\hat{D}_2^\lambda(u+t-s) - \hat{D}_2^\lambda(u)) + (\hat{A}^\lambda(t) - \hat{A}^\lambda(s)) \\
& \quad + (t-s)\sqrt{N^\lambda} a_1 \mu + o(\sqrt{N^\lambda}) < -\frac{\epsilon}{2}\},
\end{aligned}$$

where $\hat{A}^\lambda(t) := \frac{A^\lambda(t) - \lambda t}{\sqrt{N^\lambda}}$ and $\hat{D}_2^\lambda(t) = \frac{D_2(N_2^\lambda t) - N_2^\lambda \mu_2 t}{\sqrt{N_2^\lambda}}$. Note that, due to the functional central limit theorems for renewal processes, $\hat{A}^\lambda \Rightarrow BM(0, \mu)$ and $\hat{D}_2^\lambda \Rightarrow BM(0, \mu_2)$, as $\lambda \rightarrow \infty$, where $BM(0, \mu)$ stands for a driftless Brownian motion with infinitesimal variance μ . Finally, an argument analogous to the Proof of Theorem 3.2 in [47] shows that the right-hand-side of (A.11) goes to 0 as $\lambda \rightarrow \infty$. \square

Proof of Remark 4.5. Lemma A2 remains unchanged, while the argument for lemma A1 is trivially the following: suppose that for $k = 1, 2$, $\tilde{X}_k^\lambda(0) \rightarrow x_k$ in probability, as $\lambda \rightarrow \infty$. We show that $x_2 = 0$, and then the lemma is true with $b^\lambda \equiv 0$. By contradiction, suppose that $x_2 < 0$, then for λ large enough, and with probability close to 1, we have $\tilde{X}_2^\lambda(0) < 0$. In particular, $Z_2^\lambda(0) < N_2^\lambda$. But from the ‘‘faster servers used first’’ and the work conservation properties of the policy FSF_P we then have, $Z_1^\lambda(0) + Q^\lambda(0) = 0$, which is a contradiction to the assumption that $\tilde{X}_1^\lambda(0) = \frac{Z_1^\lambda(0) + Q^\lambda(0) - N_1^\lambda}{\sqrt{N^\lambda}}$ converges, in probability, to a finite limit. \square

Proof of Proposition 4.2. We prove the proposition for the case $K = 2$. The general case will follow similarly. We introduce the following notation (adapted from [46]). Consider the Poisson processes:

$$S_k^l = S_k^l(t), t \geq 0 \quad \text{with rate } \mu_k, \quad k = 1, 2, \quad l = 1, 2, \dots$$

The interpretation of these processes is as follows: the process S_k^l corresponds to the number of service completions of the l^{th} server of pool k that is currently busy. When there are fewer than l customers being served in pool k at the moment of a jump in S_k^l , the jump has no effect on the system state. The total number of customers in the system

process admits the following dynamics:

$$\begin{aligned}
 Y^\lambda(t) &:= Q^\lambda(t) + Z_1^\lambda(t) + Z_2^\lambda(t) \\
 &= Q^\lambda(0) + Z_1^\lambda(0) + Z_2^\lambda(0) + A^\lambda(t) - \sum_{k=1}^2 \sum_{l=1}^{N_k} \int_0^t 1_{\{Z_k^\lambda(s-) \geq l\}} dS_k^l(s). \quad (\text{A.12})
 \end{aligned}$$

Define $\mathcal{F}^\lambda(t)$ to be the following σ -algebra:

$$\mathcal{F}^\lambda(t) = \sigma \{ Q^\lambda(0), Z_k^\lambda(0), A^\lambda(s), S_k^l(s); \quad k = 1, 2, \quad l \geq 1, 0 \leq s \leq t \} \vee \mathcal{N},$$

where \mathcal{N} denotes the family of P -null sets, and introduce the filtration $\mathbb{F}^\lambda = (\mathcal{F}^\lambda(t), t \geq 0)$. Clearly, the processes Q^λ and $Z_k^\lambda, k = 1, 2$, are \mathbb{F}^λ -adapted.

We claim that $Y^\lambda(t)$ admits the following decomposition:

$$Y^\lambda(t) = Y^\lambda(0) + \lambda t - \sum_{k=1}^2 \mu_k \int_0^t Z_k^\lambda(s) ds + M^\lambda(t), \quad (\text{A.13})$$

where $M^\lambda = (M^\lambda(t), t \geq 0)$ is an \mathbb{F}^λ -locally square-integrable martingale, that satisfies $M^\lambda = M_A^\lambda - \sum_{k=1}^2 M_{S_k}^\lambda$, where M_A^λ and $M_{S_k}^\lambda, k = 1, 2$, are three independent \mathbb{F}^λ -locally square-integrable martingales with respective predictable quadratic variations:

$$\langle M_A^\lambda \rangle(t) = \lambda t, \quad (\text{A.14})$$

$$\langle M_{S_k}^\lambda \rangle(t) = \mu_k \int_0^t Z_k^\lambda(s) ds, \quad k = 1, 2. \quad (\text{A.15})$$

To show the validity of the decomposition (A.13), note that the Poisson processes A^λ and S_k^l admit the representations [46, (3.8)–(3.11)]:

$$A^\lambda(t) = \lambda t + M_A^\lambda(t), \quad (\text{A.16})$$

$$S_k^l(t) = \mu_k t + M_k^l(t), \quad k = 1, 2, \quad l \geq 1, \quad (\text{A.17})$$

where M_A^λ and M_k^l are independent locally square-integrable martingales relative to the associated natural filtrations (as well as relative to \mathbb{F}^λ) with respective predictable quadratic variations (A.14) and

$$\langle M_k^l \rangle(t) = \mu_k t. \quad (\text{A.18})$$

Now, from (A.16), (A.17), (A.14), (A.18), we get that (A.12) may be represented as (A.13). The latter implies that:

$$\begin{aligned}
X^\lambda(t) &= X^\lambda(0) + \frac{\sum_{k=1}^2 \mu_k N_k^\lambda}{\sqrt{N^\lambda}} t - \delta \frac{\sqrt{\sum_{k=1}^2 \mu_k N_k^\lambda}}{\sqrt{N^\lambda}} t \\
&\quad + \sum_{k=1}^2 \mu_k \int_0^t [X_k^\lambda(s)]^- ds - \frac{\sum_{k=1}^2 \mu_k N_k^\lambda}{\sqrt{N^\lambda}} t + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1) \\
&= X^\lambda(0) - \delta \sqrt{\mu} t + \sum_{k=1}^2 \mu_k \int_0^t [X_k^\lambda(s)]^- ds + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1) \\
&= X^\lambda(0) - \delta \sqrt{\mu} t + \mu_1 \int_0^t [X^\lambda(s)]^- ds + \epsilon^\lambda(t) + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1), \quad (\text{A.19})
\end{aligned}$$

where $\sup_{t \leq T} |\epsilon^\lambda(t)| \xrightarrow{p} 0$, and the last equality follows from Proposition 4.1. Now note that from (A.14), (A.15) and Proposition 2.1 we have

$$\left\langle \frac{1}{\sqrt{N^\lambda}} M_A^\lambda \right\rangle(t) \xrightarrow{p} \mu t, \quad \text{and} \quad \left\langle \frac{1}{\sqrt{N^\lambda}} M_{S_k}^\lambda \right\rangle(t) \xrightarrow{p} q_k \mu_k t,$$

and by Theorem 8.3.1 in [40] the processes $\{M_A^\lambda/\sqrt{N^\lambda}, M_{S_k}^\lambda/\sqrt{N^\lambda}, k = 1, 2\}$ converge jointly in distribution to $\{\sqrt{\mu} b_A, \sqrt{q_k \mu_k} b_k, k = 1, 2\}$, where $b_A, b_k, k = 1, 2$, are independent standard Brownian motions. Therefore, by the continuous mapping theorem the process M^N/\sqrt{N} converges to $b = \sqrt{\mu} b_A - \sqrt{q_1 \mu_1} b_1 - \sqrt{q_2 \mu_2} b_2$. It is easy to verify that b is a Brownian motion with zero drift and variance 2μ . Applying the continuous mapping theorem to the process X^λ completes the proof of the Proposition. \square

Proof of Proposition 4.3. The proof is a result of a corollary by Puhalskii [45] which deals with limits of the first passage time. Specifically, it further follows straightforwardly from Lemma A.2 in Puhalskii and Reiman [46] (which is based on [45]), and from assumption (A3). \square

Proof of Proposition 4.4. The proof follows from [15]. Note that the process $X(\cdot)$, restricted to $[0, \infty)$, is a reflected Brownian motion with infinitesimal drift $-\delta\sqrt{\mu}$ and variance 2μ . Hence, according to [15, (18.33)], its steady-state density conditional on $X(\infty) \geq 0$ is exponential with rate $\delta/\sqrt{\mu}$. Similarly, the process $X(\cdot)$ restricted to the negative half-line is an O-U process with infinitesimal drift $-\delta\sqrt{\mu} - \mu_1 x$ and variance 2μ . Therefore, its stationary density conditional on $X(\infty) < 0$ is the density of a normal random variable with mean $-\delta\sqrt{\mu}/\mu_1$, and variance μ/μ_1 conditioned on having negative values only (see [15, (18.28)]). Putting these two densities together, establishes that $f(x)$ is indeed the steady-state density of X , with $\alpha = P(X(\infty) \geq 0)$.

To find the value of α , note that $f(\cdot)$ is continuous because the infinitesimal variance is continuous on the whole real line (see [15, p. 471]). Hence, α may be solved for by a smooth fit, namely, by equating the limits of $f(\cdot)$ at 0 from both left and right. \square

Proof of Proposition 4.5. We prove the Proposition for $K = 2$. The general proof follows similarly. We need to show that for all $-\infty < x < \infty$, we have

$$P(X^\lambda(\infty) \leq x) \rightarrow P(X(\infty) \leq x), \quad \text{as } \lambda \rightarrow \infty. \tag{A.20}$$

Note that for all x , $P(X^\lambda(\infty) \leq x) = P(Y^\lambda(\infty) \leq N^\lambda + \sqrt{N^\lambda}x) = \sum_{n \leq N^\lambda + \sqrt{N^\lambda}x} p_n^\lambda$. Recall that for $n = 0, 1, \dots$, $p_n^\lambda = p_0^\lambda \pi_n^\lambda$. For $K = 2$, π_n^λ satisfies:

$$\pi_n^\lambda = \begin{cases} \frac{\lambda^n}{\mu_2^n n!}, & \text{if } 0 \leq n \leq N_2^\lambda, \\ \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \prod_{i=N_2^\lambda+1}^n (\mu_2 N_2^\lambda + (i - N_2^\lambda)\mu_1)}, & \text{if } N_2^\lambda < n \leq N^\lambda - 1, \\ \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! (N_1^\lambda \mu_1 + N_2^\lambda \mu_2)^{(n - N_2^\lambda + 1)} \prod_{i=N_2^\lambda+1}^{N^\lambda-1} (\mu_2 N_2^\lambda + (i - N_2^\lambda)\mu_1)}, & \text{if } N^\lambda \leq n. \end{cases} \tag{A.21}$$

The proof of (A.20) is based on the expression A.21, but the details are tedious. Hence, for clarity, we only describe its three main steps:

1. Let $\alpha^\lambda = P(X^\lambda(\infty) \geq 0)$ which is (due to work conservation and the PASTA property) the steady-state probability that an arbitrary customer will have to wait before starting service. Then, $\alpha^\lambda \rightarrow \alpha$, as $\lambda \rightarrow \infty$. To prove this, we explicitly write down the steady-state waiting probability for every fixed $\lambda > 0$, and show, that as $\lambda \rightarrow \infty$, this expression converges to α . The main result used in establishing this convergence is the central limit theorem (CLT).
2. For all $x < 0$, we show that (A.20) holds at x . This is done by first establishing that, due to 1., it is sufficient to show that for all $x < 0$, $P(X^\lambda(\infty) \leq x \mid X^\lambda(\infty) < 0) \rightarrow P(X(\infty) \leq x \mid X(\infty) < 0)$, as $\lambda \rightarrow \infty$. Second, we explicitly spell out the steady-state probabilities: $P(X^\lambda(\infty) \leq x \mid X^\lambda(\infty) < 0)$ for $\lambda > 0$. Finally, by an extensive use of the CLT we establish the desired convergence, as $\lambda \rightarrow \infty$.
3. For all $x \geq 0$, we show that $P(X^\lambda(\infty) > x) \rightarrow P(X(\infty) > x)$, as $\lambda \rightarrow \infty$. This is the simplest step of all three. First, we note that, due to 1., it is sufficient to establish that, for all $x \geq 0$, $P(X^\lambda(\infty) \leq x \mid X^\lambda(\infty) \geq 0) \rightarrow P(X(\infty) \leq x \mid X(\infty) \geq 0)$, as $\lambda \rightarrow \infty$. Second, we note that for all $\lambda > 0$ the process $X^\lambda(\cdot)$, restricted to non-negative values, is a Birth and Death process with constant birth and death rates, and

hence, the resulting steady-state distribution is geometric. The resulting convergence as $\lambda \rightarrow \infty$ is then straightforward.

□

Proof of Proposition 4.6. The proof is based on Ethier and Kurtz [21, Theorem 9.10 and Remark 9.11, p. 244]. According to [21] and based on our Propositions 4.1 and 4.2, it suffices to show that:

1. There exists a stationary distribution of $\vec{X}^\lambda(\cdot)$ for all λ .
2. The sequence of stationary distributions of $\vec{X}^\lambda(\cdot)$ is tight.

We establish 1. and 2. for $K = 2$. The general case follows similarly.

1. Fix $\lambda > 0$. To show the existence of a stationary distribution of \vec{X}^λ , it is sufficient to establish that the state $(0, 0)$ is positive recurrent, due to the irreducibility of the process (the superscript λ is omitted from the following notation for brevity). Equivalently, let $T_{(0,0)}$ be the time of first returning to the state $(0, 0)$, given that the process starts there. Then it is sufficient to show that $ET_{(0,0)} < \infty$. We will establish the finiteness of this expectation by showing that $ET_{(0,0)} \leq ET_{(0,0)}^P$, where $T_{(0,0)}^P$ is the corresponding returning time to the state $(0, 0)$ under FSF_P . The finiteness of $ET_{(0,0)}^P$ is known due to the existence of the stationary distribution of \vec{X} under FSF_P (which can be obtained from (A.21)). In particular,

$$ET_{(0,0)}^P = \frac{1}{P(\vec{X}(\infty; \text{FSF}_P) = (0, 0))}.$$

Recall the definition of $\tilde{\Pi}_P$ (given in Section 3) as the family of all work-conserving preemptive policies which always use the faster servers first. According to Lemma 3.1, there exists a policy $\tilde{\pi} \in \tilde{\Pi}_P$ such that $X(t; \text{FSF}) \geq X(t; \tilde{\pi})$ for all t , with probability 1. In addition, from the second part of the proof of Proposition 3.1, we have that $\tilde{\pi}$ and FSF_P share the same steady-state distribution. Particularly, if $\tilde{T}_{(0,0)}$ is the returning time to the state $(0, 0)$ under the policy $\tilde{\pi}$, then $E\tilde{T}_{(0,0)} = ET_{(0,0)}^P < \infty$. We will show that $ET_{(0,0)} \leq E\tilde{T}_{(0,0)}$. The latter is true due to the following observations:

- (a) Let $S_- := \mathbb{R}_-^2$ and $S_+ := \mathbb{R}_+ \times \{0\}$. Then, the processes $\vec{X}(\cdot; \tilde{\pi})$ and $\vec{X}(\cdot; \text{FSF})$ both have state spaces which are subsets of $S = S_- \cup S_+$ (due to work conservation).
- (b) Under both policies, in order to have a transition from S_- to S_+ or back, the process has to visit the state $(0, 0)$ first.
- (c) Let T, \tilde{T} be the time of the first transition out of the state $(0, 0)$ under FSF and $\tilde{\pi}$, respectively. Then, according to a) and b), we have

$$ET_{(0,0)} = ET + P(\vec{X}(T; \text{FSF}) \in S_-) E(T_{(0,0)} - T \mid \vec{X}(T; \text{FSF}) \in S_-)$$

$$\begin{aligned}
 & + P(\vec{X}(T; \text{FSF}) \in S_+) E(T_{(0,0)} - T \mid \vec{X}(T; \text{FSF}) \in S_+) \\
 = & E\tilde{T} + P(\vec{X}(\tilde{T}; \tilde{\pi}) \in S_-) E[T_{(0,0)} - T \mid \vec{X}(T; \text{FSF}) \in S_-] \\
 & + P(\vec{X}(\tilde{T}; \tilde{\pi}) \in S_+) E[T_{(0,0)} - T \mid \vec{X}(T; \text{FSF}) \in S_+],
 \end{aligned}$$

where the second equality follows from the fact that the transition rates out of the state $(0, 0)$ are the same under both policies.

- (d) Note that the transition rates of both processes restricted to S_+ are the same. Hence,

$$E[T_{00} - T \mid \vec{X}(T; \text{FSF}) \in S_+] = E[\tilde{T}_{00} - \tilde{T} \mid \vec{X}(\tilde{T}; \tilde{\pi}) \in S_+].$$

- (e) Due to the pathwise dominance of $\tilde{\pi}$ over FSF with respect to $X(\cdot)$, we have

$$(X(t; \text{FSF}) \mid \vec{X}(T; \text{FSF}) \in S_-) \geq (X(t; \tilde{\pi}) \mid \vec{X}(\tilde{T}; \tilde{\pi}) \in S_-)$$

for all $t \geq 0$, with probability 1. In particular,

$$(X(\tilde{T}_{(0,0)}; \text{FSF}) \mid \vec{X}(T; \text{FSF}) \in S_-) \geq (X(\tilde{T}_{(0,0)}; \tilde{\pi}) \mid \vec{X}(\tilde{T}; \tilde{\pi}) \in S_-) = 0.$$

Specifically, at time $\tilde{T}_{(0,0)}$, $\vec{X}(\tilde{T}_{(0,0)}; \text{FSF}) \in S_+$. From observation *b*), it follows that

$$(T_{(0,0)} \mid \vec{X}(T; \text{FSF}) \in S_-) \leq (\tilde{T}_{(0,0)} \mid \vec{X}(\tilde{T}; \tilde{\pi}) \in S_-),$$

which implies that

$$E[T_{(0,0)} - T \mid \vec{X}(T; \text{FSF}) \in S_-] \leq E[\tilde{T}_{(0,0)} - \tilde{T} \mid \vec{X}(\tilde{T}; \tilde{\pi}) \in S_-].$$

- (f) From (c), (d) and (e), it follows that $E[T_{(0,0)}] \leq E[\tilde{T}_{(0,0)}]$. This establishes the existence of a stationary distribution of $\vec{X}^\lambda(\cdot)$ for all λ .
2. Now that the existence of a stationary distribution for \vec{X}^λ has been established for all λ , we need to show that the resulting sequence of stationary distributions is tight. For any measurable set $K \subseteq S$, let $\nu^\lambda(K) := P(\vec{X}^\lambda(\infty; \text{FSF}) \in K)$ and let $\eta^\lambda(K) := P(\vec{X}^\lambda(\infty; \text{FSF}_p) \in K)$. By Proposition 4.5, $\eta^\lambda(\cdot)$ is tight. Hence, given $\epsilon > 0$, there is a compact set K_0 such that $\eta^\lambda(K_0) \geq 1 - \tilde{\epsilon} \stackrel{\Delta}{=} 1 - \frac{\alpha}{2+\alpha}\epsilon$, for all λ and $\alpha = \alpha(\delta/\sqrt{\mu_1})$. Our goal is to find another compact set, \tilde{K} , such that $\nu^\lambda(\tilde{K}) \geq 1 - \epsilon$, for all λ large enough.

Let $K^+ := \{(x_1, x_2) \in S \mid \exists (y_1, y_2) \in K_0 \text{ with } y_1 + y_2 \leq x_1 + x_2\}$. That is, K^+ is the set of all points in the state space, whose total sum of their elements weakly dominates the sum of the elements of at least one point from K_0 (see Figure A.1 for illustration). From Proposition 3.1, we have $\nu^\lambda(K^+) \geq \eta^\lambda(K^+) \geq \eta^\lambda(K_0) \geq 1 - \tilde{\epsilon}$. Hence, K^+ satisfies the probability requirement. However, it is not compact, because it is unbounded from above.

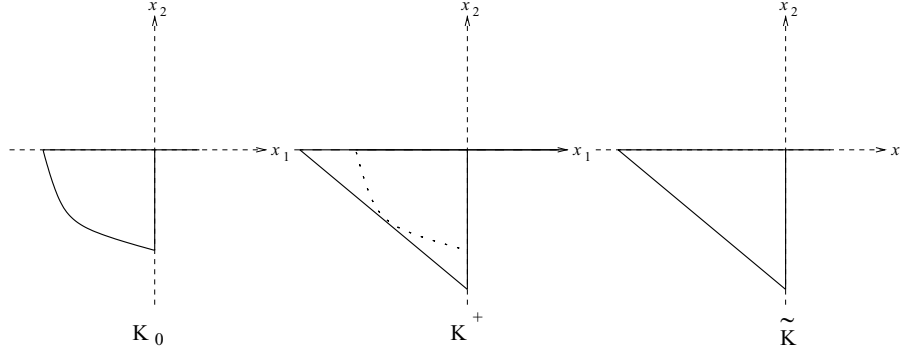


Figure A.1. Illustration of the tightness proof.

Let $\bar{K}^+ = \{(x_1, 0) \in S \mid x_1 \geq 0, (x_1, 0) \notin K_0, \text{ and } \exists (y_1, y_2) \in K_0 \text{ with } y_1 + y_2 \leq x_1\}$. Then, $\bar{K}^+ \subseteq K^+$, and $K^+ \setminus \bar{K}^+$ is compact. \bar{K}^+ is the part of K^+ we wish to remove in order to obtain compactness. Before it is removed, we need to make sure that its measure is small enough not to spoil tightness. Recall that the transition rates of \bar{X}^λ restricted to $\mathbb{R}_+ \times \{0\}$ are the same for both FSF_p and FSF . Hence, $v^\lambda(K \mid \mathbb{R}_+ \times \{0\}) = \eta^\lambda(K \mid \mathbb{R}_+ \times \{0\})$ for all λ and any measurable set K . Specifically,

$$\begin{aligned} v^\lambda(\bar{K}^+) &= v^\lambda(\bar{K}^+ \mid \mathbb{R}_+ \times \{0\})v^\lambda(\mathbb{R}_+ \times \{0\}) = \eta^\lambda(\bar{K}^+ \mid \mathbb{R}_+ \times \{0\})v^\lambda(\mathbb{R}_+ \times \{0\}) \\ &= \frac{\eta^\lambda(\bar{K}^+)}{\eta^\lambda(\mathbb{R}_+ \times \{0\})}v^\lambda(\mathbb{R}_+ \times \{0\}) = \eta^\lambda(\bar{K}^+) \cdot \frac{\alpha_{NP}^\lambda}{\alpha_p^\lambda} \\ &\leq \tilde{\epsilon} \cdot \frac{\alpha_{NP}^\lambda}{\alpha_p^\lambda} \leq \tilde{\epsilon} \frac{1}{\alpha/2}, \quad \text{for all } \lambda \text{ large enough, independently of } \tilde{\epsilon}. \end{aligned}$$

Here, α_{NP}^λ and α_p^λ are the steady-state waiting probabilities for the λ system, under FSF and FSF_p , respectively. The first inequality follows from the fact that $\bar{K}^+ \cap K_0 = \emptyset$. The second inequality is due to Proposition 4.5, and particularly, the fact that $\alpha_p^\lambda \rightarrow \alpha$ as $\lambda \rightarrow \infty$. Finally, let $\tilde{K} = K^+ \setminus \bar{K}^+$, then \tilde{K} is compact and

$$v^\lambda(\tilde{K}) = v^\lambda(K^+ \setminus \bar{K}^+) \geq 1 - \tilde{\epsilon} - \tilde{\epsilon} \frac{1}{\alpha/2} = 1 - \tilde{\epsilon} \left(\frac{2 + \alpha}{\alpha} \right) = 1 - \epsilon.$$

□

Proof of Theorem 4.1. Let $\{\pi^\lambda\}_{\lambda > 0} \subseteq \Pi$ be a sequence of policies, and suppose that the steady-state distributions of $X^\lambda(\cdot; \pi^\lambda)$, $Q^\lambda(\cdot; \pi^\lambda)$ and $\hat{W}^\lambda(\cdot; \pi^\lambda)$ exist for all $\lambda > 0$. In addition, suppose that the weak limits, $X(\infty; \{\pi^\lambda\})$, $\hat{X}_0(\infty; \{\pi^\lambda\})$ and $\hat{W}(\infty; \{\pi^\lambda\})$ of $X^\lambda(\infty; \pi^\lambda)$, $\hat{X}_0^\lambda(\infty; \pi^\lambda) := Q^\lambda(\infty; \pi^\lambda)/\sqrt{N^\lambda}$ and $\hat{W}^\lambda(\infty; \pi^\lambda)$, respectively, exist as $\lambda \rightarrow \infty$.

We prove the theorem in four steps:

1. First we show asymptotic optimality of FSF in terms of $X^\lambda(\infty)$, as $\lambda \rightarrow \infty$.
2. The asymptotic optimality of FSF with respect to X^λ is used to show its asymptotic optimality with respect to the queue length.
3. The asymptotic optimality with respect to X^λ is trivially shown to imply the asymptotic optimality with respect to the probability of having at least N^λ customers in the system. For work conserving policies the latter is equal to the probability that all servers are busy, or the waiting probability.
4. The asymptotic optimality of FSF with respect to the waiting probability is shown to imply its asymptotic optimality with respect to the waiting time distribution.

1. We need to show that

$$P(X(\infty; \text{FSF}) > x) \leq P(X(\infty; \{\pi^\lambda\}) > x) \text{ for all } x, \quad -\infty < x < \infty.$$

This includes establishing the existence of (i) the steady-state distribution of $X^\lambda(\cdot; \text{FSF})$ for all λ , and (ii) the existence of $X(\infty; \text{FSF})$, the limit of $X^\lambda(\infty; \text{FSF})$ as $\lambda \rightarrow \infty$. Recall that both (i) and (ii) were established in Proposition 4.6. The latter together with Proposition 4.5 also established that $X(\infty; \text{FSF}) = X(\infty; \text{FSF}_p) = \lim_{\lambda \rightarrow \infty} X^\lambda(\infty; \text{FSF}_p)$. Finally, the optimality of FSF_p with respect to $X^\lambda(\infty)$ for all λ (see Proposition 3.1) implies that indeed FSF is asymptotically optimal with respect to $X^\lambda(\infty)$, as $\lambda \rightarrow \infty$.

2. We wish to show that for all $q \geq 0$,

$$P(\hat{X}_0(\infty; \text{FSF}) > q) \leq P(\hat{X}_0(\infty; \{\pi^\lambda\}) > q), \tag{A.22}$$

The proof follows directly from 1. and from the facts that $\hat{X}_0^\lambda(\infty; \text{FSF}) = [X^\lambda(\infty; \text{FSF})]^+$, a.s. (work conservation) and that $\hat{X}_0^\lambda(\infty; \pi^\lambda) \geq [X^\lambda(\infty; \pi^\lambda)]^+$, a.s. for all $\lambda > 0$. (For work conserving policies $\tilde{\alpha}^\lambda(\pi^\lambda) = \alpha^\lambda(\pi^\lambda) := P_{\pi^\lambda}^\lambda(\text{wait} > 0)$.)

3. For any sequence of policies $\{\pi^\lambda\}$, for which the steady-state distribution of $X^\lambda(\cdot; \pi^\lambda)$ exists for all λ , let $\tilde{\alpha}^\lambda(\pi^\lambda) = P(X^\lambda(\infty; \pi^\lambda) \geq N^\lambda)$, be the probability of having at least N^λ customers in the system. Suppose that $X(\infty; \{\pi^\lambda\}) = \lim_{\lambda \rightarrow \infty} X^\lambda(\infty; \pi^\lambda)$ exists. Then 1. implies that

$$\alpha(\text{FSF}) = \lim_{\lambda \rightarrow \infty} \alpha^\lambda(\text{FSF}) \leq \lim_{\lambda \rightarrow \infty} \tilde{\alpha}^\lambda(\pi^\lambda) =: \tilde{\alpha}(\{\pi^\lambda\}).$$

4. We wish to show that for all $w \geq 0$ we have

$$P(\hat{W}(\infty; \text{FSF}) > w) \leq P(\hat{W}(\infty; \{\pi^\lambda\}) > w). \tag{A.23}$$

To prove (A.23) it suffices to show that

- (i) The steady-state distribution of $\hat{W}^\lambda(\infty; \text{FSF})$ exists for all $\lambda > 0$.

- (ii) The weak limit $\hat{W}(\infty; \text{FSF})$ of $\hat{W}^\lambda(\infty; \text{FSF})$ as $\lambda \rightarrow \infty$ exists.
- (iii) $\hat{W}(\infty; \text{FSF}) \stackrel{st}{\leq} \hat{W}(\infty; \{\pi^\lambda\})$.
- (i) The existence of a steady-state distribution of $\hat{W}^\lambda(\infty; \text{FSF})$ for all $\lambda > 0$ follows from Corollary 3.1.
- (ii) To show the existence of a weak limit $\hat{W}(\infty; \text{FSF})$ of $\hat{W}^\lambda(\infty; \text{FSF})$ as $\lambda \rightarrow \infty$, recall that by (3.5),

$$P(W^\lambda(\infty; \text{FSF}) > w) = \alpha^\lambda(\text{FSF}) e^{-\left(\sum_{k=1}^K \mu_k N_k^\lambda - \lambda\right)w}, \quad \forall w \geq 0,$$

where $\alpha^\lambda(\text{FSF}) = P(X^\lambda(\infty; \text{FSF}) \geq 0)$. In particular,

$$\begin{aligned} P(\hat{W}^\lambda(\infty; \text{FSF}) > w) &= P(\sqrt{N^\lambda} W^\lambda(\infty; \text{FSF}) > w) \\ &= \alpha^\lambda(\text{FSF}) e^{-\frac{\left(\sum_{k=1}^K \mu_k N_k^\lambda - \lambda\right)w}{\sqrt{N^\lambda}}} \\ &\rightarrow \alpha(\text{FSF}) e^{-\delta \sqrt{\mu} w}, \quad \forall w \geq 0, \text{ as } \lambda \rightarrow \infty. \end{aligned}$$

The convergence of $\alpha^\lambda(\text{FSF})$ as $\lambda \rightarrow \infty$ was established in 3.

- (iii) To show that $\hat{W}(\infty; \text{FSF}) \stackrel{st}{\leq} \hat{W}(\infty; \{\pi^\lambda\})$, note that since the sequence $\{\pi^\lambda\}$ may contain some policies which are not work-conserving, (3.5) may not hold any more, but instead, (3.3) implies that

$$\begin{aligned} P(W^\lambda(\infty; \{\pi^\lambda\}) > w) &\geq \tilde{\alpha}^\lambda(\pi^\lambda) e^{-\frac{\left(\sum_{k=1}^K \mu_k N_k^\lambda - \lambda\right)w}{\sqrt{N^\lambda}}} \rightarrow \tilde{\alpha}(\{\pi^\lambda\}) e^{-\delta \sqrt{\mu} w}, \\ &\forall w \geq 0, \text{ as } \lambda \rightarrow \infty. \end{aligned}$$

Now, since $\tilde{\alpha}(\{\pi^\lambda\}) \geq \alpha(\text{FSF})$ (by 3.), the asymptotic optimality of the steady-state waiting time then immediately follows. \square

Proof of Lemma 4.1. We prove (4.7). The relationship (4.6) follows similarly. Suppose that π^λ is work conserving for all $\lambda > 0$. We omit the policy and time arguments from all notation for brevity. Recall that Y^λ is the steady-state total number of customers in the system. From (3.2) we have

$$W^\lambda \stackrel{\mathcal{D}}{=} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} T_i^\lambda, \quad \lambda > 0, \quad (\text{A.24})$$

where T_i^λ are iid random variables distributed $\exp(\sum_{k=1}^K \mu_k N_k^\lambda)$, and are independent of Y^λ . It is easy to see that $\frac{[Y^\lambda - N^\lambda + 1]^+}{\sqrt{N^\lambda}} \Rightarrow [X]^+$, as $\lambda \rightarrow \infty$. Let Y^λ , and X be versions of the original random variables such that the latter convergence is almost surely. For

sample paths such that $Y^\lambda - N^\lambda + 1 \rightarrow \infty$ we have:

$$\begin{aligned} \sqrt{N^\lambda} W^\lambda &\stackrel{\mathcal{D}}{=} \sqrt{N^\lambda} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} T_i^\lambda \\ &= \frac{[Y^\lambda - N^\lambda + 1]^+}{\sqrt{N^\lambda}} \frac{1}{[Y^\lambda - N^\lambda + 1]^+} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} N^\lambda T_i^\lambda \rightarrow \frac{[X]^+}{\mu}, \end{aligned}$$

almost surely, as $\lambda \rightarrow \infty$. The convergence follows from the strong law of large numbers applied to $N^\lambda T_i^\lambda$. If Y^λ does not diverge to ∞ then, in particular, $\lim_{\lambda \rightarrow \infty} \frac{[Y^\lambda - N^\lambda + 1]^+}{\sqrt{N^\lambda}} = [X]^+ = 0$. In this case, for any subsequence $\{\lambda_j\}$ for which $\{[Y^{\lambda_j} - N^{\lambda_j} + 1]^+\}$ is bounded, we have $[Y^{\lambda_j} - N^{\lambda_j} + 1]^+ \leq \log(N^{\lambda_j})$ for all j large enough. Hence, for all j large enough

$$\sqrt{N^{\lambda_j}} W^{\lambda_j} \stackrel{\mathcal{D}}{=} \sqrt{N^{\lambda_j}} \sum_{i=1}^{[Y^{\lambda_j} - N^{\lambda_j} + 1]^+} T_i^{\lambda_j} \leq \frac{\log(N^{\lambda_j})}{\sqrt{N^{\lambda_j}}} \frac{1}{\log(N^{\lambda_j})} \sum_{i=1}^{\log(N^{\lambda_j})} N^{\lambda_j} T_i^{\lambda_j} \rightarrow 0, \text{ as } j \rightarrow \infty.$$

□

References

- [1] M. Armony and N. Bambos, Queueing dynamics and maximal throughput scheduling in switched processing systems, *Queueing Systems* 44 (2003) 209–252.
- [2] M. Armony and C. Maglaras, On customer contact centers with a call-back option: Customer decisions, routing rules and system design, *Operations Research* 52(2) (2004) 271–292.
- [3] M. Armony and C. Maglaras, Contact centers with a call-back option and real-time delay information, *Operations Research* 52(4) (2004) 527–545.
- [4] M. Armony and A. Mandelbaum, Routing and staffing in large-scale service systems with heterogeneous servers and impatient customers, Preprint (2005).
- [5] R. Atar, A diffusion model of scheduling control in queueing systems with many servers, *Ann. Appl. Probab.* 15(1B) (2005) 820–852.
- [6] R. Atar, Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic, *Ann. Appl. Probab.*, to appear (2005).
- [7] R. Atar, A. Mandelbaum and M. Reiman, Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy traffic, *Ann. Appl. Prob.* 14(3) (2004) 1084–1134.
- [8] A. Bassamboo, J.M. Harrison and A. Zeevi, Design and control of a large call center: Asymptotic analysis of an LP-based method, preprint (2004).
- [9] A. Bassamboo, J.M. Harrison and A. Zeevi, Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits, preprint (2004).
- [10] S.L. Bell and R.J. Williams, Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy, *Annals of Applied Probability* 11 (2001) 608–649.
- [11] S. Bhulai and G. Koole, A queueing model for call blending in call centers, *IEEE Transactions on Automatic Control* 48 (2003) 1434–1438.

- [12] S. Borst, A. Mandelbaum and M. Reiman, Dimensioning large call centers, *Operations Research* 52(1) (2004) 17–34.
- [13] M. Bramson, State space collapse with applications to heavy-traffic limits for multiclass queueing networks, *Queueing Systems* 30 (1997) 89–148.
- [14] A. Brandt and M. Brandt, On a two-queue priority system with impatience and its application to a call center, *Methodology and Computing in Applied Probability* 1 (1999) 191–210.
- [15] S. Browne and W. Whitt, Piecewise-linear diffusion processes, in: *Advances in Queueing. Theory, Methods, and Open Problems*, ed. J.H. Dshalalow, (CRC Press, 1995), Chapter 18, pp. 463–480.
- [16] H. Chen and D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, (Springer, New-York, 2001).
- [17] J.G. Dai, Stability of fluid and stochastic processing networks, *MaPhySto*, 9 (1999).
- [18] F. de Véricourt and Y.-P. Zhou, A routing problem for call centers with customer callbacks after service failure, *Operations Research*, to appear, (2004).
- [19] F. de Véricourt and Y.-P. Zhou, On the incomplete results for the multiple-server slow-server problem, Technical report, Duke University, The Fuqua School of Business (2004).
- [20] A.K. Erlang, On the rational determination of the number of circuits, in: *The Life and Works of A.K. Erlang*, eds. E. Brockmeyer, H.L. Halstrom, and A. Jensen, (Copenhagen: The Copenhagen Telephone Company, Copenhagen, 1948).
- [21] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence* (John Wiley & Sons, 1985).
- [22] A. Federgruen and H. Groenevelt, $M/G/c$ systems with multiple customer classes: Characterization and achievable performance under nonpreemptive priority rules, *Management Science* 34 (1988) 1121–1138.
- [23] P. Fleming, A. Stolyar and B. Simon, Heavy traffic limit for a mobile phone system loss model, in: *Proceedings of 2nd Int'l Conf. on Telecomm. Syst. Mod. and Analysis*, Nashville, TN (1994).
- [24] G.J. Foschini, On heavy traffic diffusion analysis and dynamic routing in packet switched networks, in: *Computer Performance*, eds. K.M. Chandy and M. Reiser (North Holland, 1977).
- [25] N. Gans, G. Koole and A. Mandelbaum, Telephone call centers: Tutorial, review and research prospects, *Manufacturing & Service Operations Management* 5(2) (2003) 79–141.
- [26] N. Gans and Y.-P. Zhou, A call-routing problem with service-level constraints, *Operations Research* 51(2) (2003) 255–271.
- [27] O. Garnett, A. Mandelbaum and M. Reiman, Designing a call center with impatient customers, *Manufacturing & Service Operations Management* 4(3) (2002) 208–227.
- [28] K. Glazebrook and J. Niño-Mora, Parallel scheduling of multiclass $M/M/m$ queues: Approximate and heavy-traffic optimization of achievable performance, *Operations Research* 49(4) (2001) 609–623.
- [29] P.W. Glynn, Diffusion approximations, in: *Stochastic Models, Handbooks in OR & MS*, eds. D. Heyman and M. Sobel (North-Holland, 1990) vol. 2, pp. 145–198.
- [30] I. Gurvich, Design and control of the $M/M/N$ queue with multi-class customers and many servers, Masters Thesis, Technion Institute of Technology, Israel (2004).
- [31] I. Gurvich, M. Armony and A. Mandelbaum, Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers, preprint, (2004).
- [32] S. Halfin and W. Whitt, Heavy-traffic limits for queues with many exponential servers, *Operations Research* 29(3) (1981) 567–588.
- [33] J.M. Harrison, Heavy traffic analysis of a system with parallel servers: Asymptotic analysis of discrete-review policies, *Annals of Applied Probability* 8 (1998) 822–848.
- [34] J.M. Harrison and A. Zeevi, Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime, *Operations Research* 52(2) (2004) 243–257.
- [35] D.L. Jagerman, Some properties of the Erlang loss function, *Bell Systems Technical Journal* 53(3) (1974) 525–551.

- [36] P. Jelenkovic, A. Mandelbaum and P. Momcilović, The GI/D/N queue in the QED regime, *Queueing Systems* 47 (2004) 53–69.
- [37] O. Kella and U. Yechiali, Waiting times in the nonpreemptive priority $M/M/c$ queue, *Stochastic Models* 1(2) (1985) 257–262.
- [38] F.P. Kelly and C.N. Laws, Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling, *Queueing Systems* 13 (1993) 47–86.
- [39] W. Lin and P.R. Kumar, Optimal control of a queueing system with two heterogeneous servers, *IEEE Trans. Automat. Control* 29 (1984) 696–703.
- [40] R.Sh. Lipster and A.N. Shiryaev, *Theory of Martingales*, (Kluwer, Amsterdam, 1989).
- [41] H.P. Luh and I. Viniotis, Threshold control policies for heterogeneous server systems, *Math Meth Oper Res* 55 (2002) 121–142.
- [42] C. Maglaras and A. Zeevi, Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions, *Management Science* 49(8) (2003) 1018–1038.
- [43] A. Mandelbaum and A.L. Stolyar, Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$ -rule, *Operations Research* 52(6) (2004) 836–855.
- [44] W.A. Massey and R.B. Wallace, An optimal design of the $M/M/C/K$ queue for call centers, *Queueing Systems* to appear (2004).
- [45] A. Puhalskii, On the invariance principle for the first passage time, *Mathematics of Operations Research* 19(4) (1994) 946–954.
- [46] A.A. Puhalskii and M.I. Reiman, The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime, *Advances in Applied Probability* 32 (2000) 564–595.
- [47] M.I. Reiman, Some diffusion approximations with state space collapse, in: *Modelling and Performance Evaluation Methodology*, eds. F. Baccelli and G. Fayolle (Springer-Verlag, 1984) pp. 209–240.
- [48] V.V. Rykov, Monotone control of queueing systems with heterogeneous servers, *Queueing Systems* 37 (2001) 391–403.
- [49] V.V. Rykov and D. Efrosinin, Optimal control of queueing systems with heterogeneous servers, *Queueing Systems* 46, (2004) 389–407.
- [50] A.A. Scheller-Wolf, Necessary and sufficient conditions for delay moments in FIFO multiserver queues with an application comparing slow servers with one fast one, *Operations Research* 51 (2003) 748–758.
- [51] J.G. Shanthikumar and D.D. Yao, Comparing ordered-entry queues with heterogeneous servers, *Queueing Systems* 2 (1987) 235–244.
- [52] R.A. Shumsky, Approximation and analysis of a call center with flexible and Specialized servers, *OR Spectrum* 26(3) (2004) 307–330.
- [53] S. Stolyar, Optimal routing in output-queues flexible server systems, *Probability in the Engineering and Informational Sciences* 19 (2005) 141–189.
- [54] D.Y. Sze, A queueing model for telephone operator staffing, *Operations Research* 32 (1984) 229–249.
- [55] Y.-Ch. Teh and A.R. Ward, Critical thresholds for dynamic routing in queueing networks, *Queueing Systems* 42 (2002) 297–316.
- [56] R.B. Wallace and W. Whitt, A Staffing Algorithm for Call Centers with Skill-Based Routing, working paper (2004).
- [57] W. Whitt, Heavy traffic approximations for service systems with blocking, *AT & T Bell Lab. Tech. Journal* 63 (1984) 689–708.
- [58] W. Whitt, *Stochastic-process limits: An introduction to stochastic-process limits and their application to queues*, Springer (2002).
- [59] W. Whitt, A diffusion approximation for the $G/GI/n/m$ queue, *Operations Research* 52(6) (2004) 922–941.
- [60] W. Whitt, Heavy-traffic limits for the $G/H_2^*/n/m$ queue, *Mathematics of Operations Research* 30(1) (2005) 1–27.