Design, staffing and control of large service systems: The case of a single customer class and multiple server types.

Mor Armony[1] Avishai Mandelbaum[2]

April 2, 2004

DRAFT

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 simple staffing rules which asymptotically minimize staffing costs. The minimization is subject to constraints on the waiting probability, as demand grows large. The proposed staffing rules add a square-root safety service capacity to the nominal capacity required for system stability. For large values of system demand, the resulting asymptotic regime is what we call the Quality and Efficiency Driven (QED) regime: it achieves high levels of both service quality and system efficiency by carefully balancing between the two. Finally, we propose an asymptotically optimal routing scheme, FSF, which assigns customers to the Fastest Servers First.

Contents

[1]Stern School of Business, New York University, marmony@stern.nyu.edu.

[2]Industrial Engineering and Management, Technion Institute of Technology, avim@ie.technion.ac.il.

1 Introduction

In modern service systems it is common to have multiple classes of customers and multiple server types (skills). The customer classes are differentiated according to their service needs. The server types are characterized by the subset of customer classes that they can adequately serve and the quality of service that they can devote to each such class. An important example of such large scale service system are multi-skill call/contact-centers. Such centers are often characterized by multiple classes of calls (classified according to type or level of service requested, langauge spoken, perceived value of customers, etc.). To match the various service needs of those customers, call centers often consist of hundreds of even thousands of customer service representatives (CSRs). These CSRs have different skills, depending on the call classes that they can handle, and the speed in which they do it.

There are three main issues to address when dealing with the operations management of large-scale service systems. Given a forecast of the customers' arrival rates and their service requirements, these issues are:

- **Design:** The long-term problem of determining the class partitioning of customers, and the types of servers; this typically includes overlapping skills (i.e. servers that can handle more than one class of customers, and classes that can be served by several server types).
- **Staffing:** The short-term problem of determining how many servers are needed of each type, in order to deal with the given demand. These server types may be of overlapping skills. (In addition, there is a scheduling problem which determines the shift structure for the system, as well as determining who are the actual servers that would work in these shifts. The last two issues will not be discussed in this paper.)

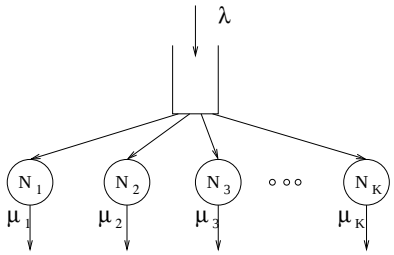

Figure 1.1: The Inverted-V model - single customer class and multiple server types.

- **Control:** The on-line problem of customer routing and server scheduling that involves the assignment of customers to the appropriate server upon service completion or a customer's arrival.

These three problems are all interrelated and should, therefore, be discussed in conjunction with one another. Yet, because of the complexity involved in addressing all these three combined, they are typically addressed hierarchically and unilaterally in the literature.

Even when one addresses the three issues separately, a general solution for all possible system configurations is yet to be achieved. Instead, we approach the problem by studying a relatively simple model in order to gain insight to the more general model. The model we focus on in this work is the $\bigwedge$-design (or the inverted-V design). This is a system design in which customers are homogeneous, with K server types (organized in K pools) that have full overlap of their skills, but differ in the speed in which they serve the customers. Alternatively, one could look at the $V-$design (studied in [7, 29, 57] and elsewhere), which corresponds to a system with a single server pool and multiple customer classes). The $\bigwedge$-design is depicted in figure 1.1.

With respect to the $\bigwedge$-design we ask the following two questions:

1. Given a fixed number of servers of each pool, how to route the customers into the different server pools so as to optimize system performance, and
2. How many servers of each pool are required in order to minimize staffing costs while maintaining pre-specified performance goals.

We address these questions by first characterizing a simple routing scheme which is asymptotically optimal as the arrival rate and the number of servers in each pool increase to infinity. The asymptotic optimality is in the sense that the policy (asymptotically and stochastically) minimizes the steady-state queue length and waiting time (both appropriately scaled). We then identify a simple form for an asymptotic feasible region. This region is the set of all staffing vectors that can obtain a pre-specified waiting probability in steady-state, asymptotically as the arrival rate grows large. Finally, the asymptotic optimality of the staffing vector that minimizes the staffing costs

within the asymptotically feasible region is established for a wide range of cost functions. We conclude by studying the effects of our results on design related issues such as: How many server pools should one have? and, Does having fewer but faster servers affect performance?

The asymptotic framework considered in this paper is the many-server heavy-traffic regime, first appearing in Erlang [18], and formally introduced by Halfin and Whitt [30]. 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. More formally, consider a sequence of systems of a fixed design and an increasing arrival rate λ. Suppose that the total service capacity of each system in the sequence exceeds λ by a safety capacity of order $\sqrt{\lambda}$. In particular, the traffic intensity (or server efficiency) goes to 1 as $\lambda \to \infty$ (ie. the system goes to heavy traffic). On the other hand, the high quality aspect of the QED regime may be seen through the following alternative characterization: Suppose that as $\lambda \to \infty$, the limiting waiting probability is non-trivial (ie. it is in the open interval $(0, 1)$). 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. The two characterizations of the QED regime are shown to be equivalent in various settings (first established in [30]. See the literature review, section 1.2, for more details), including the one considered in this paper (see Section 4).

1.1 Summary of the results

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 policies. 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, $k = 1, ..., K$. Also, let $\vec{N}^\lambda = (N_1^\lambda, N_2^\lambda, ..., N_K^\lambda)$ be the staffing vector, and $N^\lambda = N_1^\lambda + N_2^\lambda + ... + N_K^\lambda$ be the total number of servers. Suppose that the service rates: $\mu_1, ..., \mu_K$ are fixed independently of λ. To be consistent with the QED regime assume that the total service capacity, $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + ... + \mu_K N_K^\lambda$, is equal to the arrival rate plus a square root safety capacity. Formally, suppose that

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

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 $\lambda \to \infty$ (see Theorem 3.1 for further details).

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 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 $\lambda \to \infty$, both policies give rise to the same performance measures. 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.

Fix a customer arrival rate, λ. The associated feasible region for this system is defined as the set of all staffing vectors for which there exists a routing policy under which the steady-state waiting probability does not exceed a pre-specified level. We show that, as the arrival rate grows to infinity, the feasible region is asymptotically linear (see Figure 4.2). Specifically, the total service capacity $\mu_1 N_1 + \mu_2 N_2 + ... + \mu_K N_K$ associated with any staffing vector $\vec{N}$ in the asymptotically feasible set is greater than or equal to the arrival rate plus a square-root safety capacity; that is, the safety capacity is of the form of a constant times a square-root of the arrival rate (the total capacity is equal to $\lambda + \delta\sqrt{\lambda}$, for some positive constant δ). As mentioned earlier, this, in particular, means that the system operates in the QED regime; namely, the QED regime is obtained as an outcome rather than an assumption.

Finally, due to the simple structure of the feasible region, identifying an asymptotically optimal staffing rule may be done by simply finding the lowest cost staffing vector(s) within the linear (asymptotically) feasible region. We show that, by following this procedure, one indeed obtains staffing rules which are asymptotically optimal for various staffing cost functions. For example, we consider staffing costs which are polynomial and homogeneous of the form $C(\vec{N}) = c_1 N_1^p + c_2 N_2^p + ... + c_K N_K^p$, for some $p > 1$. In this case, the staffing vector $\vec{N}$ which is proportional to the vector $(\mu_1/c_1, \mu_2/c_2, ..., \mu_K/c_K)$, and satisfies (1.1) is shown to be asymptotically optimal.

The remainder of the paper is organized as follows: We conclude the introduction by reviewing the relevant literature. In section 2, we detail the single-customer-class multiple-server-types model, and the asymptotic framework used in our analysis. In section 3, we present our proposed routing policy and prove its asymptotic optimality. Section 4 then outlines the form of the asymptotic feasible region, and proves the associated asymptotic feasibility. In section 5, this asymptotic feasibility is finally used to propose an asymptotically optimal staffing rule. The claimed asymptotic optimality is established in this section as well.

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. The QED regime is achieved at by letting $\sqrt{N^\lambda}(1-\rho^\lambda) \to \beta$, 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 [30], GI/D/N [35] and M/M/N with exponential patience [26].

Due to the desirable features of the QED regime, it has enjoyed recently considerable attention in the literature. Yet the regime was explicitly recognized already in Erlang's 1923 paper (that appeared in [18]) 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 [51] (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 [34]; see also [53], and then [42] 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 [30]: Within the GI/M/N framework, they analyzed the scaled number of customers, both in steady state and as a stochastic process. Recent generalizations are [55, 56]. Convergence of the scaled queueing process, in the more general GI/PH/N setting, was established in [45]. Application of QED queues to modelling and staffing of telephone call centers and communication networks, taking into account customers' impatience, can be found in [26] and [21], respectively. The optimality of the QED regime, under revenue maximization or constraint satisfaction, is discussed in [10, 40, 3, 4]. Readers are referred to Sections 4 and 5.1.4 of [22] 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 [15], 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 [22].

Skill-based routing

Of the three issues related to the management of large-scale service system, the control problem has received the most attention in the literature. Specifically, for a given design, and staffing levels,

researchers have proposed routing and / or scheduling schemes that are either optimal or near-optimal. Alternatively, researchers have considered commonly used routing schemes (such as fixed priority rules, or dedicated servers per customer class) and computed the relevant performance measures. Examples for both criteria include: **Exact analysis** (Kella and Yechiali [37], Federgruen and Groenvelt [20], Brandt and Brandt [13], Gans and Zhou [25], Armony and Bambos [2], Rykov [47], Luh and Viniotis [39], and de Véricourt and Zhou [17] ([47] and [39] are concerned with the $\wedge-$model, and will be expanded on in section 3.1)), **Asymptotic analysis - "conventional" heavy traffic** (Harrison [31], Bell and Williams [9], Glazebrook and Niño-Mora [27], Teh and Ward [52], Mandelbaum and Stolyar [41] and Stolyar [50]) and **Asymptotic analysis - QED regime** (Armony and Maglaras [3, 4], Harrison and Zeevi [32], Atar et. al. [7], and Atar [5, 6]).

Staffing Rules

The staffing problem in the single-class, single-type case has also gained a lot of attention in the literature. With multi-type, however, things are quite different. The problem of determining how many servers of each type are required is very difficult. This is especially true if skills overlap. In the latter case, one wants to take advantage of the flexibility of the servers who have multiple skills, but these servers are typically more costly. The most common approaches taken by researchers to tackle the staffing problem are: **Heuristical bounds:** Using heuristics to achieve performance bounds by analyzing simpler (but related) systems (Examples include Borst and Seri [11], Whitt [54], and Jennings et al. [36]), **Stability Staffing:** Staffing levels that guarantee system stability (Examples include Bambos and Walrand [8], Gans and van Ryzin [23], Armony and Bambos [2]), and **Cost minimizing staffing:** For a given routing scheme, find the staffing level that minimizes personnel costs while guaranteeing certain performance bounds, or alternatively, such staffing levels that minimize personnel costs plus operating costs (Examples include Borst et al. [10] (QED regime), Perry and Nilsson [43], Stanford and Grassmann [49], Shumsky [48] and Harrison and Zeevi [33]).

Design

On the design front, even less has been done. Ganz and Zhou [24] develop a dynamic programming (DP) model of long term server hiring that admits a general class of controls. There, the lower level routing problem is explicitly modelled as the core of the DP's one-period cost function, and the optimal hiring policies are characterized as analogues to "order-up-to" policies in the inventory literature. Other studies we are aware of focus on design for flexibility that results from the cross-training of service reps (see Aksin and Karaesmen [1] and references therein).

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, where the service rate depends on the pool (type) of the particular server. Specifically, the average service time of a customers that is served by a server of type k ($k = 1, 2, ..., K$) is $1/\mu_k$. We assume that the service rates are ordered as follows: $\mu_1 < \mu_2 < ... < \mu_K$. 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.

We seek to determine the number N_k of servers required of each type k, $k = 1, 2, .., K$. In choosing the staffing levels N_k we require that, at the very least, N_k are sufficiently large to ensure *stability*. Specifically, we require the following necessary condition for stability:

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

that is, the total service capacity is larger than the arrival rate. The cost of staffing the system with N_k servers of type k is denoted by $C_k(N_k)$. The total staffing cost is, hence, $C(N_1, N_2, ..., N_K) = C_1(N_1) + C_2(N_2) + ... + C_K(N_K)$. By determining the number of servers required of each type, we wish to minimize the staffing cost while maintaining a target service level constraint. The service performance measure that we study is the steady-state probability that a customer waits before starting service. Equivalently, we focus on the long-term proportion of customers who are delayed before their service starts. Denote this steady-state probability by $P(wait > 0)$, and let $0 < \alpha < 1$ be the target waiting probability. The staffing problem is then stated as:

$$\begin{array}{ll} \text{minimize} & C_1(N_1) + C_2(N_2) + ... + C_K(N_K) \\ \text{subject to} & P(wait > 0) \leq \alpha \\ & N_1, N_2, ..., N_K \in \mathbb{Z}_+. \end{array} \tag{2.2}$$

In order to solve (2.2), one needs to be able to evaluate $P(wait > 0)$ given any server staffing vector $\vec{N} = (N_1, N_2, ..., N_K)$ (here and elsewhere, $\vec{x}$ is used to denote a vector whose elements are $x_1, x_2, ...$). This requires knowing the actual routing policy that is used to determine which type of server will handle each customer. In particular, different routing policies can result in different waiting probabilities. Let Π be the set of all non-preemptive non-anticipative 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 $P_\pi(wait > 0)$ be the steady state probability that a customer is delayed before his service starts.[3] Then a more precise definition of the staffing problem (2.2) is as follows:

$$\begin{array}{ll} \text{minimize} & C_1(N_1) + C_2(N_2) + ... + C_K(N_K) \\ \text{subject to} & P_\pi(wait > 0) \leq \alpha, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & N_1, N_2, ..., N_K \in \mathbb{Z}_+. \end{array} \tag{2.3}$$

[3]If steady-state does not exist, consider $P_\pi(wait > 0)$ as the random variable corresponding to the essential limsup of the long term proportion of customers who are delayed before receiving service.

As mentioned in the introduction, solving the staffing and control problems concurrently is usually too difficult. Hence, researchers commonly end up solving one while assuming the solution to the other is fixed. A distinguishing feature of our solution to (2.3) is that we identify a policy which is near-optimal given *any* staffing level, and therefore, are able to solve the staffing and the control problems concurrently.

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 ($k = 1, 2, ..., K$) 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, $Y(t;\pi) = Z_1(t;\pi) + Z_2(t;\pi) + ...Z_K(t;\pi) + Q(t;\pi)$. 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 $Z_1(t;\pi) + Z_2(t;\pi) + ... + Z_K(t;\pi) = N$, where

$$N = N_1 + N_2 + ... + N_K$$

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 $Z_1(t;\pi), ..., Z_K(t;\pi)$. However, for work conserving policies, the state space can be described by the $K-$dimensional vector $(Z_1(t;\pi)+Q(t;\pi), Z_2(t;\pi), ..., Z_K(t;\pi))$. 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 $Z_k(t;\pi) = [Q(t;\pi) + Z_k(t;\pi) - N_k]^-$ (where $[x]^- := -min\{x, 0\}$). 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 for its underlying processes, then

$$P_\pi(wait > 0) = P(Z_1(\infty;\pi) + Z_2(\infty;\pi) + ... + Z_K(\infty;\pi) = N) = P(Y(\infty;\pi) \geq N), \quad (2.4)$$

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 (2.4) 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 $k = 1, ..., K$ 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 customer 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 $A_q(t;\pi)$. 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*:

$$Q(t;\pi) = Q(0;\pi) + A_q(t;\pi) - \sum_{k=1}^{K} B_k(t;\pi), \tag{2.5}$$

$$Z_k(t;\pi) = Z_k(0;\pi) + A_k(t;\pi) + B_k(t;\pi) - D_k(T_k(t;\pi)), \;\; k = 1, ..., K, \tag{2.6}$$

$$T_k(t;\pi) = \int_0^t Z_k(s;\pi)ds \tag{2.7}$$

$$Y(t;\pi) = Y(0;\pi) + A(t) - \sum_{k=1}^{K} D_k(T_k(t;\pi)), \tag{2.8}$$

$$A(t) = A_q(t;\pi) + \sum_{k=1}^{K} A_k(t;\pi), \tag{2.9}$$

$$T_k(t;\pi) + I_k(t;\pi) = N_k t. \tag{2.10}$$

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

$$Q(t;\pi) \cdot \left(\sum_{k=1}^{K} (N_k - Z_k(t;\pi)) \right) = 0, \tag{2.11}$$

$$\int_0^\infty \sum_{k=1}^{K} (N_k - Z_k(t;\pi)) dA_q(t;\pi) = 0, \tag{2.12}$$

and

$$\sum_{k=1}^{K} \int_0^\infty Q(t;\pi) dI_k(t;\pi) = 0. \tag{2.13}$$

In words, (2.11) means that there are customers in queue only when *all* servers are busy. The verbal interpretation of (2.12) is that new arrivals wait in the queue only when all servers are busy. Finally, (2.13) states that servers can only be idle when the queue is empty.

2.1 Asymptotic Framework

Although the staffing problem (2.3) is well defined, it is difficult to be solved exactly. Specifically, given fixed values of $\mu_1, \mu_2, ..., \mu_K, \lambda$ and α, one would need to find the *feasible region* of all those vectors $(N_1, N_2, ..., N_K)$ for which there exists a policy that satisfies $P_\pi(wait > 0) \leq \alpha$, and then find the vector(s) that minimizes the staffing costs within this feasible region. Instead, we take an asymptotic approach, which finds asymptotically optimal staffing rules for systems with high demand (i.e. large values of λ). To this end, we consider a sequence of systems and routing policies indexed by λ (to appear as a superscript) with increasing arrival rates $\lambda \uparrow \infty$, but with fixed service rates $\mu_1, \mu_2, ..., \mu_K$ and a fixed target waiting probability α.

The appropriate staffing levels will be determined according to the staffing costs and the desired service level. For the time being we assume (this assumption will, in fact, be established later as a result under some general conditions) that there are K numbers $a_k \geq 0$, $k = 1, ..., K$, with $a_1 > 0$ and $\sum_{k=1}^{K} a_k = 1$, such that the number of servers of each pool N_k^λ, $k = 1, 2, ..., K$, grows with λ as follows:

$$N_k^\lambda = a_k \frac{\lambda}{\mu_k} + o(\lambda), \text{ as } \lambda \to \infty, \quad \text{or,} \quad \lim_{\lambda \to \infty} \frac{\mu_k N_k^\lambda}{\lambda} = a_k. \tag{2.14}$$

Condition (2.14) guarantees that the total traffic intensity,

$$\rho^\lambda \triangleq \frac{\lambda}{\sum_{k=1}^{K} \mu_k N_k^\lambda}, \tag{2.15}$$

converges to 1, as $\lambda \to \infty$, and hence, for large λ, the system is in *heavy traffic*. Also, in view of (2.14), the quantity $a_k \lambda / \mu_k$ can be considered as the offered load of server pool k. Let

$$\mu = \left[\sum_{k=1}^{K} a_k / \mu_k \right]^{-1}, \tag{2.16}$$

then λ/μ is the total offered load of the whole system. Given this definition of μ, (2.14) implies that

$$N^\lambda = \frac{\lambda}{\mu} + o(\lambda), \text{ as } \lambda \to \infty, \quad \text{or,} \quad \lim_{\lambda \to \infty} \frac{\lambda}{N^\lambda} = \mu, \tag{2.17}$$

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

$$\rho^\lambda \approx \frac{\lambda}{N^\lambda \mu}, \tag{2.18}$$

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

$$\lim_{\lambda \to \infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k} \mu \triangleq q_k \geq 0, \quad k = 1, ..., K, \tag{2.19}$$

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 (2.14) 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 the functional strong law of large numbers (FSLLN), this scaling leads to the fluid dynamics of the system, in the limit as $\lambda \to \infty$. To see this, for $\lambda \uparrow \infty$, $k = 1, ..., K$ and a fixed sequence of routing policies $\pi^\lambda \in \Pi$ (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 [16], Theorem 2.3.1) one can show that if $(\bar{Q}^\lambda(0), \bar{Z}_k^\lambda(0),\ k = 1, ..., K)$ 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 $\lambda \to \infty$, 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 (2.5)-(2.10) imply that the following flow balance equations hold for *any* fluid limit:

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}_q(t) - \sum_{k=1}^{K} \bar{B}_k(t), \tag{2.20}$$

$$\bar{Z}_k(t) = \bar{Z}_k(0) + \bar{A}_k(t) + \bar{B}_k(t) - \mu_k \bar{T}_k(t), \quad k = 1, ..., K, \tag{2.21}$$

$$\bar{T}_k(t) = \int_0^t \bar{Z}_k(s) ds \tag{2.22}$$

$$\bar{Y}(t) = \bar{Y}(0) + \mu t - \sum_{k=1}^{K} \mu_k \bar{T}_k(t), \tag{2.23}$$

$$\mu t = \bar{A}_q(t) + \sum_{k=1}^{K} \bar{A}_k(t), \tag{2.24}$$

$$\bar{T}_k(t) + \bar{I}_k(t) = q_k t. \tag{2.25}$$

Finally, for work conserving policies, conditions (2.11)-(2.13) imply:

$$\bar{Q}(t) \cdot \left(\sum_{k=1}^{K} (q_k - \bar{Z}_k(t)) \right) = 0, \tag{2.26}$$

$$\int_0^\infty \sum_{k=1}^{K} (q_k - \bar{Z}_k(t)) d\bar{A}_q(t) = 0, \tag{2.27}$$

and

$$\sum_{k=1}^{K} \int_0^\infty \bar{Q}(t) d\bar{I}_k(t) = 0. \tag{2.28}$$

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

Proposition 2.1 ***(fluid limits)*** *For $\lambda > 0$, let $\pi^\lambda \in \Pi$ 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 $\lambda\to\infty$. Recall that $q_k = \lim_{\lambda\to\infty} \frac{N_k^\lambda}{N^\lambda} = \frac{a_k}{\mu_k}\mu,\ k = 1, ..., K$, and suppose that $\bar{Q}(0) = 0$ and $\bar{Z}_k(0) = q_k,\ k = 1, ..., K$. Then, $\bar{Q}(t) = 0$ and $\bar{Z}_k(t) = q_k,\ k = 1, ..., K$, for all $t \geq 0$.

Proof: Let $f(t) = \left|\bar{Y}(t) - 1\right| = \left|\sum_{k=1}^K (\bar{Z}_k(t) - q_k) + \bar{Q}(t)\right|$, 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 $k = 1, ..., K$. By an argument similar to lemma 2.4.5 of [16], 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 (2.26) $\bar{Z}_k(t) = q_k$, 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 (2.26), $\bar{Q}(t) = 0$. If f is differentiable at t then,

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

∎

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 policy $\pi^\lambda \in \Pi$ (omitted from the notation), define the centered and scaled process $\vec{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), ..., X_K^\lambda(\cdot))$ as follows:

$$X_1^\lambda(t) := \frac{Q^\lambda(t) + Z_1^\lambda(t) - N_1^\lambda}{\sqrt{N^\lambda}}, \tag{2.29}$$

and, for $k = 2, ..., K$, let

$$X_k^\lambda(t) := \frac{Z_k^\lambda(t) - N_k^\lambda}{\sqrt{N^\lambda}}. \tag{2.30}$$

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

$$X^\lambda(t) = \sum_{k=1}^K X_k^\lambda(t) = \frac{Q^\lambda(t) + \sum_{k=1}^K Z_k^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \frac{Y^\lambda(t) - N^\lambda}{\sqrt{N^\lambda}} = \sqrt{N^\lambda}\left(\bar{Y}^\lambda(t) - 1\right). \tag{2.31}$$

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, $\left[X^\lambda(t)\right]^-$ is the total number of idle servers, and $\left[X^\lambda(t)\right]^+ = \left[X_1^\lambda(t)\right]^+$ 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:

$$\hat{W}^\lambda(t) = \sqrt{N^\lambda} W^\lambda(t). \tag{2.32}$$

As will be shown later, in order for the diffusion scaling to have well defined limits, as $\lambda \to \infty$, we add the following assumption, in addition to (2.14):

$$\sum_{k=1}^{K} \mu_k N_k^\lambda = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda}), \text{ as } \lambda \to \infty, \quad \text{or,} \quad \lim_{\lambda\to\infty} \frac{\sum_{k=1}^{K} \mu_k N_k^\lambda - \lambda}{\sqrt{\lambda}} = \delta, \tag{2.33}$$

for some δ, $0 < \delta < \infty$.

Condition (2.33) is a square-root safety staffing rule (similar to [30] and [10]). In particular, the condition $\delta > 0$ guarantees that the system is stable (or can be stable, under reasonable routing) for all λ large enough. Note that (2.33) 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 $k = 1, ..., K$, and $\lambda > 0$, let $-\infty < \delta_k^\lambda < \infty$ satisfy:

$$\delta_k^\lambda := \frac{\mu_k N_k^\lambda - a_k \lambda}{\sqrt{\lambda}}. \tag{2.34}$$

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 $a_k = 0$,

$$\delta_k^\lambda = o(\sqrt{\lambda}), \text{ as } \lambda \to \infty, \ \forall k = 1, ..., K, \tag{2.35}$$

and

$$\delta^\lambda := \sum_{k=1}^{K} \delta_k^\lambda \to \delta, \text{ as } \lambda \to \infty. \tag{2.36}$$

Note that we do not require the individual sequences $\{\delta_k^\lambda\}_{\lambda>0}$ to have a limit, for any value of $k = 1, ..., K$. All that is assumed is that their sum converges to δ. The one exception to this rule is Proposition 3.4, in which the following additional condition is assumed to hold:

$$\theta := \lim_{\lambda\to\infty} \sum_{k=1}^{K} \frac{\delta_k^\lambda}{\mu_k}, \text{ exits for some finite number } \theta. \tag{2.37}$$

3 Routing Policies

In this section we describe three routing policies. The first one, $\pi^* \in \Pi$, is an optimal policy that minimizes the long-term average of the total number of customers in the system and the average

sojourn time, given any fixed values of system parameters. This policy is simple to describe but its implementation requires the computation of certain threshold values which are a function of the model parameters and system state. The second one, FSF$_P$, is a simple *preemptive* policy which is optimal within the set of all non-anticipative, but possibly preemptive policies, with respect to the steady-state distribution of the total number of customers in the system. Finally, we describe a third policy, FSF, 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 $\lambda \to \infty$), in terms of the steady-state queue length and waiting time distributions.

3.1 Background: Optimal Non-Preemptive Routing

In this section we describe an optimal policy π^* within the set Π, and some of its properties. The policy is based on two recent papers [47] and [39]. Both these papers study systems with heterogenous servers, which may each have his/her own service rate. We describe their policy as adapted to our case of K server pools, with $\mu_1 < \mu_2 < ... < \mu_K$. Both papers show that for the optimality criterion of minimizing the average steady-state number of customers in the system, there exists an optimal policy of a *threshold* type. According to this policy, one should assign a customer to an idle server of pool k if:

1. It is the fastest idle server, and
2. the number of customers in queue is equal to or exceeds a threshold m_k, $m_k \geq 0$.

The thresholds have the following properties:

- m_k may depend on the state of the other servers (current pool k and slower ones in pools $1, ..., k-1$),
- they are non-increasing in the service rates; that is, $m_1 \geq m_2 \geq ... \geq m_K$.

Note that π^* minimizes the average total number of customers in the system in steady-state. However, this does not imply that it minimizes the average steady-state queue length or waiting time. The reason is that this policy is *not* work conserving, and hence the queue length is not a well defined function of the total number of customers in the system. Also note that this policy should actually be denoted as $\pi^{*\lambda}$, because the threshold values may, conceivably, depend on the actual values of λ and $\vec{N} = (N_1, N_2, ...N_K)$.

3.2 Optimal Preemptive Routing

In this section we describe a policy which is optimal within a greater family of policies $\Pi_P \supseteq \Pi$, namely the family of all non-anticipative policies which are preemptive resume (the subscript P is for preemptive). What is meant by preemptive resume in our context 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 actions at a finite number of time points in any finite time interval, where an action includes 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 $Z_1(t;\pi) + Z_2(t;\pi) + ... + Z_K(t;\pi) < N$ 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; however, it only assigns a customer to a server upon customer arrivals and service completions. 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. 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 Π_P. In other words, for all $\pi \in \Pi_P$ and every weak limit $Y(\infty;\pi)$ of $Y(t;\pi)$, as $t \to \infty$ (or a subsequence thereof), we have $P\{Y(\infty;\pi) > y)\} \geq P\{Y(\infty;FSF_P) > y)\}$, for all $y \geq 0$.*

Proof: We prove the Proposition in two steps. The first step will establish 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 will show 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 of distribution of the total number of customers in the system under FSF$_P$ stochastically dominates the steady-state distribution of the total number of customers in the system 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:

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

In particular, the steady-state of $Y(\cdot;\pi)$ exists (recall the stability assumption) and is unique under 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 $\tilde{\pi} \in \tilde{\Pi}_P$.*

Proof: 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 $i = 1, 2, ..., n$, 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 $1 \leq i \leq n$, the customers $j = i + 1, ..., n$, 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 ($j = i+1, ..., n$) 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 ($j = i + 1, ..., n$) 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 $j = 1, ..., n$ have departed. Let $S = \{0 \leq s_1 < s_2 < ... < S_M = D_n(\pi)\}$ 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 $j = i, i+1, ..., n$, which will have at most as many total number of customers in the system at any time $t \geq 0$ as π. By backwards 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 $l = l_0 + 1, ..., M - 1$. 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 $l = l_0, ..., M - 1$, 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 $i + 1, ..., n$. 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 $i + 1, ..., n$ 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 $l = l_1, ..., M - 1$ 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 $j = i + 2, ..., n$ 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 $j = i, i+1, ..., N$.

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.

∎

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 $t \to \infty$ (or a subsequence thereof), we have $P\{Q(\infty;\pi) > q)\} \geq P\{Q(\infty;$ FSF$_P) > q)\}$, for all $q \geq 0$, and $P\{W(\infty;\pi) > w)\} \geq P\{W(\infty;$ 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:

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

and

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

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 then the steady state of $W := W(\infty;\pi)$ exists and it satisfies

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

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 not work conserving, then if the steady state distribution of $W(\cdot;\pi)$ exists it satisfies

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

Hence, a stochastic dominance of FSF$_P$ within Π_P with respect to the steady-state of the process Y implies that FSF$_P$ also stochastically minimizes both the queue length and the waiting time in steady-state. ∎

Remark 3.1 ***(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 there exists a steady state distribution Y for the total number of customers in the system. Observe that conditioned on the event $Y \geq N$, $Y - N$ has transition rates which are like and $M/M/1$ system with arrival rate λ and service rate $\sum_{k=1}^{K} \mu_k N_k$. Hence, since $Q(\infty) = [Y - N]^+$ its distribution satisfies:*

$$P(Q(\infty) = n) = \alpha \rho^n (1 - \rho), \;\; n \geq 1 \tag{3.4}$$

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

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

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

Remark 3.2 ***(State-space collapse for 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 know 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.*

3.3 Asymptotically Optimal Non-preemptive Routing

In this section we describe a simple non-preemptive policy FSF which is also work-conserving. This policy is identical to the non-preemptive policy π^* described in section 3.1, except that all the thresholds m_k are equal to zero (and hence the policy is work-conserving). It 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). Since the thresholds m_k are not chosen optimally here, 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 (2.14) and (2.33); 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 policies. This is summarized in Theorem 3.1 and proved at the end of this section via Propositions 3.1-3.7.

Theorem 3.1 *Consider a sequence of systems indexed by the arrival rate λ, that satisfy conditions (2.14) and (2.33). 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 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 waiting time ($\hat{W}^\lambda(\infty)$), as $\lambda\to\infty$.*

Remark 3.3 *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 3.4 *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 [52] 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 as to 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 π^* 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 $\lambda\to\infty$, we will show that as λ grows, the process $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), ..., X_K^\lambda(\cdot))$ (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 $\lambda\to\infty$ the two processes coincide. Taking the limits as $t\to\infty$ 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 $(X_1^\lambda(\cdot), X_2^\lambda(\cdot), ..., X_K^\lambda(\cdot))$ under FSF, into a one dimensional process as $\lambda\to\infty$. Recall, that such state-space collapse holds for every λ under FSF$_P$ (Remark 3.2). When FSF is used, this is no longer true, but the state-space collapse is attained when $\lambda\to\infty$, as will be shown in Proposition 3.2 below.

3.3.1 State-Space Collapse

In this section we establish the state-state collapse result with respect to the policy FSF and the process $\vec{X}^\lambda(\cdot) = (X_1^\lambda(\cdot), ..., X_2^\lambda(\cdot))$. 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 (2.31)) becomes sufficient in describing the whole $K-$dimensional process $\vec{X}^\lambda(\cdot)$. Specifically, we show that as $\lambda\to\infty$, all the faster servers (from pools $k = 2, ..., K$) 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 $X_2^\lambda(\cdot), ..., X_K^\lambda(\cdot)$ become identically zero, while the processes $X^\lambda(\cdot)$ and $X_1^\lambda(\cdot)$ become close. This result is presented in Proposition 3.2.

Proposition 3.2 ***(State-Space Collapse)*** *Suppose that conditions (2.14) and (2.33) hold as $\lambda\to\infty$, and that the work-conserving non-preemptive policy FSF is used. In addition, suppose that $\vec{X}^\lambda(0) \to \vec{X}(0) = \vec{x} = (x_1, ..., x_K)$, in probability, as $\lambda\to\infty$. Then for all $t > 0$ we have,*

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

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

$$P\left(\sup_{0<t\leq T}\left|\sum_{k=2}^{K} X_k^\lambda(t)\right| > \epsilon\right) \to 0, \text{ or } P\left(\inf_{0<t\leq T}\sum_{k=2}^{K} X_k^\lambda(t) < -\epsilon\right) \to 0. \tag{3.6}$$

We prove the Proposition for $K = 2$. The general case follows similarly. For $K = 2$, (3.6) translates into

$$P\left(\sup_{0<t\leq T}\left|X_2^\lambda(t)\right| > \epsilon\right) \to 0, \quad \text{or} \quad P\left(\inf_{0<t\leq T} X_2^\lambda(t) < -\epsilon\right) \to 0. \tag{3.7}$$

We claim that in order to establish (3.7) it is sufficient to show the existence of a sequence $\{b^\lambda\}$, with $b^\lambda\to 0$ as $\lambda\to\infty$, such that

$$\lim_{\lambda\to\infty} P\left(\inf_{0<t\leq T} X_2^\lambda(t + b^\lambda) < -\epsilon\right) = 0. \tag{3.8}$$

The sufficiency of (3.8) has been established in [3], and it essentially follows from a random time change argument (see [28, Prop. 5]). Recall that x_2 is the weak limit of $X_2^\lambda(0)$ as $\lambda\to\infty$. Then, as in [45, Lemma 3.3] and [3, (29)], we have, for all $C > 0$,

$$P\left(\inf_{0<t\leq T} X_2^\lambda(t + b^\lambda) < -\epsilon\right) \leq P\left(\sup_{0\leq t\leq T}\left|X_2^\lambda(t)\right| > C\right) + P\left(\inf_{|x_2|\leq C} X_2^\lambda(b^\lambda) < -\epsilon\right). \tag{3.9}$$

Hence, it is sufficient to show that both the summands on the right hand side of (3.9) converge to 0 as $C\to\infty$ and $\lambda\to\infty$. This will be shown in two steps: The first step (Lemma 3.2) will

establish that $\lim_{C\to\infty}\limsup_{\lambda\to\infty} P\left(\sup_{0\le t\le T}\left|X_2^\lambda(t)\right|>C\right)=0$. The second step (Lemma 3.3) will identify the sequence $\{b^\lambda\}$ (as a function the bound C) with $b^\lambda\to 0$ as $\lambda\to\infty$, for which $P\left(\inf_{|x_2|\le C} X_2^\lambda(b^\lambda)<-\epsilon\right)\to 0$, as $\lambda\to\infty$.

Lemma 3.2 *Suppose that $\vec{X}^\lambda(0)\to\vec{X}(0)=(x_1,...,x_K)$, in probability, as $\lambda\to\infty$. Then, under the conditions of Proposition 3.2,*

$$\lim_{C\to\infty}\limsup_{\lambda\to\infty} P\left(\sup_{0\le t\le T}\left|\sum_{k=2}^{K} X_k^\lambda(t)\right|>C\right)=0, \quad \textit{for all } T>0. \tag{3.10}$$

Proof: The proof is provided for $K=2$. The general case is similar. We introduce the following notation (adapted from [45]). Consider the Poisson processes:

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

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 affect on the system state. The total number of customers in the system process admits the following dynamics:

$$\begin{aligned} Y^\lambda(t) &:= Q^\lambda(t)+Z_1^\lambda(t)+Z_2^\lambda(t) \\ &= Q^\lambda(0)+Z_1^\lambda(0)+Z_2^\lambda(0)+A^\lambda(t)-\sum_{k=1}^{2}\sum_{l=1}^{N_k}\int_0^t 1_{\{Z_k^\lambda(s-)\ge l\}} dS_k^l(s). \end{aligned} \tag{3.11}$$

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

$$\mathcal{F}^\lambda(t)=\sigma\left\{Q^\lambda(0),Z_k^\lambda(0),A^\lambda(s),S_k^l(s);\ k=1,2,\ \ l\ge 1, 0\le s\le t\right\}\vee\mathcal{N},$$

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

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

$$Y^\lambda(t)=Y^\lambda(0)+\lambda t-\sum_{k=1}^{2}\mu_k\int_0^t Z_k^\lambda(s)ds+M^\lambda(t), \tag{3.12}$$

where $M^\lambda=(M^\lambda(t), t\ge 0)$ is an $\mathbb{F}^\lambda-$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 $\mathbb{F}^\lambda-$locally square-integrable martingales with respective predictable quadratic variations:

$$\left\langle M_A^\lambda\right\rangle(t)=\lambda t, \tag{3.13}$$

$$\left\langle M_{S_k}^\lambda\right\rangle(t)=\mu_k\int_0^t Z_k^\lambda(s)ds, \quad k=1,2. \tag{3.14}$$

To show the validity of the decomposition (3.12), note that the Poisson processes A^λ and S_k^l admit the representations [45, (3.8)-(3.11)]:

$$A^\lambda(t) = \lambda t + M_A^\lambda(t), \tag{3.15}$$

$$S_k^l(t) = \mu_k t + M_k^l(t), \quad k = 1, 2, \;\; l \geq 1, \tag{3.16}$$

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 $\mathbb{F}^\lambda$) with respective predictable quadratic variations (3.13) and

$$\left\langle M_k^l \right\rangle (t) = \mu_k t. \tag{3.17}$$

With respect to the decomposition (3.12), we also claim that there exists a constant $b > 0$ such that for all $t \geq 0$ and all λ large enough,

$$\left\langle M^\lambda \right\rangle (t) \leq b N^\lambda t. \tag{3.18}$$

To show the validity of (3.18) we use the fact that given two locally square-integrable martingales $M_1 = (M_1(t), t \geq 0)$ and $M_2 = (M_2(t), t \geq 0)$, their predictable covariation $\langle M_1, M_2 \rangle$ satisfies the inequality $2\langle M_1, M_2 \rangle \leq \langle M_1 \rangle + \langle M_2 \rangle$ (see [38, Problem 1.8.9]). Consequently, and since $M^\lambda = M_A^\lambda - M_{S_1}^\lambda - M_{S_2}^\lambda$, we have,

$$\begin{aligned} \left\langle M^\lambda \right\rangle (t) &\leq 3\left(\left\langle M_A^\lambda \right\rangle (t) + \left\langle M_{S_1}^\lambda \right\rangle (t) + \left\langle M_{S_2}^\lambda \right\rangle (t)\right) \\ &= 3\left(\lambda^\lambda t + \mu_1 \textstyle\int_0^t Z_1^\lambda(s) ds + \mu_2 \int_0^t Z_2^\lambda(s) ds\right) \\ &\leq 3\left((\mu N^\lambda + o(N^\lambda))t + \mu_1 N^\lambda t + \mu_2 N^\lambda t\right) \leq b t N^\lambda, \end{aligned}$$

for $b = 3(\mu + 1 + \mu_1 + \mu_2)$ and all λ large enough such that $\lambda \leq (\mu + 1) N^\lambda$ (exists due to (2.17)).

Now, from (3.15), (3.16), (3.13), (3.17), we get that (3.11) may be represented as (3.12). The latter implies that:

$$\begin{aligned} X^\lambda(t) &= X^\lambda(0) + \frac{\sum_{k=1}^2 \mu_k N_k^\lambda}{\sqrt{N^\lambda}} t - \delta \frac{\sqrt{\sum_{k=1}^2 \mu_k N_k^\lambda}}{\sqrt{N^\lambda}} t \\ &\quad + \sum_{k=1}^2 \mu_k \int_0^t \left[X_k^\lambda(s)\right]^- ds - \frac{\sum_{k=1}^2 \mu_k N_k^\lambda}{\sqrt{N^\lambda}} t + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1) \\ &= X^\lambda(0) - \delta\sqrt{\mu} t + \sum_{k=1}^2 \mu_k \int_0^t \left[X_k^\lambda(s)\right]^- ds + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1). \end{aligned} \tag{3.19}$$

For $k = 1, 2$, let $\hat{X}_k^\lambda(t) := (Z_k^\lambda(t) - N_k^\lambda)/\sqrt{N^\lambda}$, and let $\hat{X}_0^\lambda(t) := Q^\lambda(t)/\sqrt{N^\lambda}$. Then, the following relationships hold: $\hat{X}_0^\lambda = [X_1^\lambda]^+$, $\hat{X}_1^\lambda = -[X_1^\lambda]^-$ and $\hat{X}_2^\lambda = X_2^\lambda$. In addition, due to work conservation, we have $\left|X^\lambda\right| = \sum_{k=0}^2 \left|\hat{X}_k^\lambda\right|$. Putting all these observations together with (3.19) implies that,

$$\sum_{k=0}^2 \left|\hat{X}_k^\lambda(t)\right| \leq \sum_{k=0}^2 \left|\hat{X}_k^\lambda(0)\right| + \delta\sqrt{\mu} t + \frac{\left|M^\lambda(t)\right|}{\sqrt{N^\lambda}} + A \int_0^t \sum_{k=0}^2 \left|\hat{X}_k^\lambda(s)\right| ds + o(1),$$

for some large enough $A > 0$. Gronwall's inequality then yields

$$\sup_{0\leq t\leq T}\sum_{k=0}^{2}\left|\hat{X}_k^\lambda(t)\right| \leq \left(\sum_{k=0}^{2}\left|\hat{X}_k^\lambda(0)\right| + \delta\sqrt{\mu}T + \frac{\sup_{0\leq t\leq T}\left|M^\lambda(t)\right|}{\sqrt{N^\lambda}} + o(1)\right)\cdot e^{AT}. \tag{3.20}$$

Since $\vec{X}^\lambda(0)\to(x_1,x_2)$ in probability, as $\lambda\to\infty$, we have

$$\lim_{C\to\infty}\limsup_{\lambda\to\infty} P\left(\sum_{k=0}^{2}\left|\hat{X}_k^\lambda(0)\right| > C\right) = 0.$$

It is left to show that $\lim_{C\to\infty}\limsup_{\lambda\to\infty} P\left(\sup_{0\leq t\leq T}\left|M^\lambda(t)\right|/\sqrt{N^\lambda} > C\right) = 0$. To show this, note that since M^λ is a locally square-integrable martingale, by the Lenglart-Rebolledo inequality (see [38]) for any $B > 0$,

$$P\left(\sup_{0\leq t\leq T}\frac{\left|M^\lambda(t)\right|}{\sqrt{N^\lambda}} > C\right) \leq \frac{B}{C^2} + P\left(\frac{\left\langle M^\lambda\right\rangle(T)}{N^\lambda} > B\right). \tag{3.21}$$

Thus, from (3.18) we have,

$$\lim_{C\to\infty}\limsup_{\lambda\to\infty} P\left(\sup_{0\leq t\leq T}\left|M^\lambda(t)\right|/\sqrt{N^\lambda} > C\right) = 0. \tag{3.22}$$

■

Lemma 3.3 *Suppose that $\vec{X}^\lambda(0)\to\vec{X}(0)=\vec{x}=(x_1,...,x_K)$, in probability, as $\lambda\to\infty$. Then, under the conditions of Proposition 3.2, if $|x_k| < C,\ k\geq 2$, there exists a sequence $\left\{b^\lambda\right\}_{\lambda>0}$ (which is a function of C) with $b^\lambda\to 0$ as $\lambda\to\infty$, such that*

$$(X_2^\lambda(b^\lambda),...,X_K^\lambda(b^\lambda))\xrightarrow{p}0, \ \text{ as } \lambda\to\infty. \tag{3.23}$$

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 $\bar{Z}$ above), 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 [12], although does not make a direct use of his results.

Let

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

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

$$\tilde{X}_2^\lambda(s^*)\xrightarrow{p}0, \ \text{ as } \lambda\to\infty. \tag{3.24}$$

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

The proof follows three steps:

1. Establishing that $\tilde{X}(t) = x_1 + x_2$ for all $t \geq 0$, for all fluid limits $\tilde{X}$ of $\tilde{X}^\lambda$.
2. Establishing the existence of a fluid limit $\tilde{X}_2$ of $\tilde{X}_2^\lambda$.
3. Finding s^* such that $\tilde{X}_2(s^*) = 0$.

1. To prove (3.24) 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 $\tilde{T}_k^{\lambda_j}(\cdot) \to \tilde{T}_k(\cdot)$ as $j \to \infty$, 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, (2.17) and the key renewal theorem we have that as $\lambda \to \infty$,

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

Now, note that

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

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 and if $\tilde{T}_1(s) + \tilde{T}_2(s) = s$. But, notice also that,

$$\tilde{T}_1^\lambda(s) + \tilde{T}_2^\lambda(s) = \int_0^{s/\sqrt{N^\lambda}} \frac{Z_1^\lambda(\tau) = s + Z_2^\lambda(\tau)}{\sqrt{N^\lambda}} d\tau + \frac{1}{\sqrt{N^\lambda}} \int_0^s \tilde{X}^\lambda(\tau) d\tau \to s, \text{ as } \lambda \to \infty.$$

Therefore, we have

$$\tilde{X}(s) = x_1 + x_2 + \mu s - \mu_1 q_1 s - \mu_2 q_2 s = x_1 + x_2. \tag{3.25}$$

2. Note that if $x_2 < 0$, then $x_1 \leq 0$ (work conservation), and hence (3.25) 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

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

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

$$\begin{aligned}\tilde{X}_2^{\lambda}(s) &= \tilde{X}_2(0) + \frac{A_2^{\lambda}(s/\sqrt{N^{\lambda}})}{\sqrt{N^{\lambda}}} + \frac{B_2^{\lambda}(s/\sqrt{N^{\lambda}})}{\sqrt{N^{\lambda}}} - \frac{D_2(T_2^{\lambda}(s/\sqrt{N^{\lambda}})}{\sqrt{N^{\lambda}}} \\ &\rightarrow \tilde{X}_2(s) = x_2 + \tilde{A}_2(s) - \mu_2 q_2 s, \text{ as } \lambda\rightarrow\infty.\end{aligned} \tag{3.26}$$

3. 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 $0 \leq s \leq s^*(x_2)$, (3.26) 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).

It is still left to show that there exists $s^* = s^*(C)$ (independent of x_2) for which $\tilde{X}_2(s^*) = 0$. In view of the latter argument, if we show that $\tilde{X}_2(s) = 0$ for all $s > s^*(x_2)$, then setting $s^* = \frac{C}{\mu - \mu_2 q_2}$ will conclude the proof. Suppose, by contradiction, that there exists $\tau > s^*(x_2)$ such that $\tilde{X}_2(\tau) < 0$. Let $\tau_0 = \sup\left\{s^*(x_2) \leq t \leq \tau \mid \tilde{X}_2(t) \geq \tilde{X}_2(\tau)/2\right\}$. Note that along the interval $(\tau_0, \tau]$, $\tilde{X}_2(t) < 0$, and hence, along this interval $\tilde{A}_2(t) - \tilde{A}_2(\tau_0) = \tilde{A}(t) - \tilde{A}_2(\tau_0)$. In particular,

$$\tilde{X}_2(\tau) = \frac{\tilde{X}_2}{2} + (\mu - \mu_2 q_2)(\tau - \tau_0) \geq \frac{\tilde{X}_2}{2},$$

which contradicts the assumption that $\tilde{X}_2(\tau) < 0$.

∎

Remark 3.5 *Note the similarities and the differences between our state-space collapse result and the ones established in [45, 3, 4], for a multi-class, single server type system (the V-design) with service priority. The state-space collapse established in [45, 3, 4] 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. This results in a set of idle servers which includes slow servers only.*

Remark 3.6 *Proposition 3.2 is also true if the preemptive policy FSF$_P$ is used. Here the proof is even simpler. Lemma 3.2 remains unchanged, while the argument for lemma 3.3 is trivially the following: suppose that for $k = 1, 2$, $\tilde{X}_k^\lambda(0) \to x_k$ in probability, as $\lambda \to \infty$. 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.*

3.3.2 Transient Diffusion limit

In this section we establish the form of the diffusion limit of the scaled process $\vec{X}^\lambda$. The main purpose of presenting this transient limit here, is that it will be used later to establish the steady-state equivalence between the policies FSF$_P$ and FSF. However, the form of diffusion process obtained in the limit is also interesting in its right. Especially, when compared with the diffusion limit obtained by Halfin and Whitt [30] for the M/M/N system.

We note that the state-space collapse result of Proposition 3.2 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 3.3 ***(Transient diffusion limit)*** *Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$, as $\lambda \to \infty$, for $k = 1, ..., K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (2.14) and (2.33) hold, and that the policy FSF is used. Recall that $\mu_1 < \mu_2 < ... < \mu_K$, and $\mu = \left[\sum_{k=1}^K a_k/\mu_k\right]^{-1}$. Then, $X^\lambda \Rightarrow X$, as $\lambda \to \infty$, where X is a diffusion process with an infinitesimal drift*

$$m(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0, \\ -\delta\sqrt{\mu} - \mu_1 x & x < 0, \end{cases} \tag{3.27}$$

and infinitesimal variance

$$\sigma^2(x) = 2\mu. \tag{3.28}$$

Remark 3.7 ***(The infinitesimal drift)*** *The drift term (3.27) 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.*

Remark 3.8 ***(Drift in 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^λ, satisfies $N^\lambda\mu = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$, as $\lambda\to\infty$. That is, in both models the excess capacity is approximately equal to $\delta\sqrt{\lambda}$. For this model, let $Y^\lambda(t)$ be the total number of customers in the system at time t, and $X^\lambda(t) = (Y^\lambda(t) - N^\lambda)/\sqrt{N^\lambda}$. Then, by [30], if $X^\lambda(0) \Rightarrow X(0)$, as $\lambda\to\infty$, then $X^\lambda \Rightarrow X$, as $\lambda\to\infty$, where X is a diffusion process with an infinitesimal drift

$$m(x) = \begin{cases} -\delta\sqrt{\mu} & x \geq 0, \\ -\delta\sqrt{\mu} - \mu x & x < 0, \end{cases} \tag{3.29}$$

and infinitesimal variance

$$\sigma^2(x) = 2\mu. \tag{3.30}$$

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, and hence all can be idle at times. The comparison between the two diffusion processes reveals that the limiting process associated with the $\bigwedge-$design stochastically dominates the process associated with the $I-$design. Hence, if one is interested in determining staffing levels based on transient *performance measures, less overall capacity is required when there are multiple server types. Remark 4.2 will describe the implications of this difference on staffing which is based on steady-state performance measures.*

Proof: We prove the proposition for the case $K = 2$. The general case will follow similarly. We use the notation presented in the proof of Lemma 3.2.

Note that (3.19) implies that:

$$\begin{aligned} X^\lambda(t) &= X^\lambda(0) - \delta\sqrt{\mu}t + \sum_{k=1}^{2}\mu_k\int_0^t \left[X_k^\lambda(s)\right]^- ds + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1) \\ &= X^\lambda(0) - \delta\sqrt{\mu}t + \mu_1\int_0^t \left[X^\lambda(s)\right]^- ds + \epsilon^\lambda(t) + \frac{M^\lambda(t)}{\sqrt{N^\lambda}} + o(1), \end{aligned} \tag{3.31}$$

where $\sup_{t\leq T}\left|\epsilon^\lambda(t)\right| \xrightarrow{p} 0$, and the second equality follows from Proposition 3.2. Now note that from (3.13), (3.14) 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 [38] the processes $\left\{M_A^\lambda/\sqrt{N^\lambda},\ M_{S_k}^\lambda/\sqrt{N^\lambda},\ \ k = 1,2\right\}$ converge jointly in distribution to $\left\{\sqrt{\mu}b_A,\ \sqrt{q_k\mu_k}b_k,\ \ k = 1,2\right\}$, 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. ∎

Remark 3.9 *Proposition 3.3 remains true if the preemptive policy FSF_P is used instead. The proof remains unchanged due to Remark 3.6 and the fact that the dynamics of the total number of customers in the system is the same under both policies.*

We conclude this section by establishing the transient diffusion limit of the scaled waiting time process, which turns out to have simple linear form of the corresponding limit of the queue length process.

Proposition 3.4 *Suppose that $X_k^\lambda(0) \Rightarrow X_k(0)$ as $\lambda\to\infty$, for $k = 1, ..., K$, and let $X(0) = \sum_{k=1}^K X_k(0)$. Assume further that (2.14), (2.33) and (2.37) hold, and that the policy FSF is used. Then, $\hat{W}^\lambda := \sqrt{N^\lambda} W^\lambda \Rightarrow \hat{W}$, as $\lambda\to\infty$, where $\hat{W} = [X]^+/\mu$, and X is the diffusion limit of X^λ as $\lambda\to\infty$, given in Proposition 3.3.*

Proof: The proof is a result of a corollary by Puhalskii [44] which deals with limits of the first passage time. The result in [44] was first adapted to the QED regime by Garnett et. al. [26]. This proof further adapts the one in [26] to our setting.

Let

$$Y^\lambda = \{Y^\lambda(t), t \geq 0\}, \;\; A^\lambda = \{A^\lambda(t), t \geq 0\}, \;\; D^\lambda = \{D^\lambda(t), t \geq 0\},$$

be the total number of customers in the system, arrival and departure processes, respectively. Since FSF is work conserving and service is FIFO, $W^\lambda(t)$ can be written as:

$$W^\lambda(t) = \inf\{s \geq 0 \; : \; D^\lambda(s+t) \geq Y^\lambda(0) + A^\lambda(t) - (N^\lambda - 1)\}.$$

We define the re-scaled processes

$$\bar{Y}^\lambda(t) = \frac{1}{N^\lambda} Y^\lambda(t), \;\; \bar{A}^\lambda(t) = \frac{1}{N^\lambda} A^\lambda(t), \;\; \bar{D}^\lambda(t) = \frac{1}{N^\lambda} D^\lambda(t),$$

and an additional process $K^\lambda(t)$ characterized via $W^\lambda(t) = [K^\lambda(t) - t]^+$, or, equivalently,

$$K^\lambda(t) = \inf\{s \geq 0 \; : \; \bar{D}^\lambda(s) \geq \bar{Y}^\lambda(0) + \bar{A}^\lambda(t) - (1 - 1/N^\lambda)\}.$$

Now introduce

$$\bar{D}(t) = \mu t, \;\; \bar{Y}(0) = 1, \;\; \bar{A} = \mu t,$$

and a first passage time

$$K(t) = \inf\{s \geq 0 \; : \; \bar{D}(s) \geq \bar{A}(t)\},$$

noting that $K(t) \equiv t$. Finally, let $\theta = \lim_{\lambda\to\infty} \sum_{k=1}^K \frac{\delta_k^\lambda}{\mu_k}$, and

$$V(t) = X(0) - (\mu)^{3/2}\theta t + \sqrt{\mu} b(t),$$

and

$$U(t) = X(0) - (\mu)^{3/2}\theta t + \sqrt{\mu}b(t) - X(t),$$

where $b(t)$ is a standard Brownian motion. Then one can verify that

$$\sqrt{N^\lambda}\left((\bar{Y}^\lambda(0) + \bar{A}^\lambda - (1 - 1/N^\lambda)) - (\bar{Y}(0) + \bar{A} - 1)\right) \Rightarrow V,$$

and that

$$\sqrt{N^\lambda}\left(\bar{D}^\lambda - \bar{D}\right) \Rightarrow V.$$

Hence, by the corollary in [44], we have

$$\sqrt{N^\lambda}\left(K^\lambda - K\right) \Rightarrow \hat{K},$$

where $\hat{K}(t) = \frac{V(t) - U(K(t))}{\bar{D}'(K(t))} = \frac{X(t)}{\mu}$. In particular, due to the continuous mapping theorem, we have

$$\hat{W}^\lambda(t) = \sqrt{N^\lambda}W^\lambda(t) = \sqrt{N^\lambda}[K^\lambda(t) - t]^+ \Rightarrow \frac{[X(t)]^+}{\mu}.$$

■

3.3.3 Stationary diffusion limit

In this section we establish that the stationary distributions of the process $\vec{X}^\lambda$, under both FSF$_P$ and FSF, converge to the stationary distribution of $\vec{X}$, as $\lambda \to \infty$. 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 3.3. 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 3.3 and Remark 3.9), and the sample path optimality of $\tilde{\Pi}_P$ 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 3.5 ***(Stationary distribution of the diffusion process)*** *Let $X(\cdot)$ be the diffusion process described in Proposition 3.3, with infinitesimal drift and variance as in (3.27) and (3.28). Then the steady-state distribution of X has a density $f(\cdot)$ given by:*

$$f(x) = \begin{cases} \frac{\delta}{\sqrt{\mu}} \exp\{-\delta x/\sqrt{\mu}\}\alpha, & \text{if } x \geq 0, \\ \frac{\sqrt{\frac{\mu_1}{\mu}}\phi\left(\sqrt{\frac{\mu_1}{\mu}}x + \frac{\delta}{\sqrt{\mu_1}}\right)}{\Phi\left(\frac{\delta}{\sqrt{\mu_1}}\right)}(1-\alpha), & \text{if } x < 0, \end{cases} \tag{3.32}$$

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

Proof: The proof follows from [14]. 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 [14, (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 [14, (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 [14, p. 471]). Hence, α may be solved for by smooth fit, namely, by equating the limits of $f(\cdot)$ at 0 from both left and right. ∎

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 (3.32). 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 $\lambda^\lambda(y) = \lambda$ and death rates $\mu^\lambda(y)$ as given in (3.1). Under conditions (2.14) and (2.33) the system is stable for all λ, and the stationary distribution is given by $p_n^\lambda := P(Y^\lambda(\infty) = n) = p_0^\lambda \pi_n^\lambda$, $n = 0, 1, ...$, where $\pi_n^\lambda = \frac{\lambda^n}{\prod_{i=1}^n \mu^\lambda(i)}$, $n = 0, 1, ...$, and $p_0^\lambda = \left[\sum_{n=0}^\infty \pi_n^\lambda\right]^{-1}$. 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 3.6 ***(Convergence of the preemptive process in steady-state)*** *Suppose that conditions (2.14) and (2.33) 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 (3.32), as $\lambda \to \infty$.*

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

$$P(X^\lambda(\infty) \leq x) \to P(X(\infty) \leq x), \text{ as } \lambda \to \infty. \tag{3.33}$$

The proof of (3.33) is tedious, hence, for clarity, we first 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, $\alpha^\lambda \to \alpha$, as $\lambda \to \infty$. To prove this, we explicitly write down the steady-state waiting probability for every fixed $\lambda > 0$, and show, that as $\lambda \to \infty$, 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 (3.33) holds at x. This is done by first establishing that, due to 1., it is sufficient to show that for all $x < 0$, $P(X^\lambda(\infty) \leq x \mid X^\lambda(\infty) < 0) \to P(X(\infty) \leq$

$x \mid X(\infty) < 0)$, as $\lambda \to \infty$. Second, we explicitly spell out the steady-state probabilities: $P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) < 0)$ for $\lambda \ge 1$. Finally, by an extensive use of the CLT we establish the desired convergence, as $\lambda \to \infty$.

3. For all $x \ge 0$, we show that $P(X^\lambda(\infty) > x) \to P(X(\infty) > x)$, as $\lambda \to \infty$. This is the simplest step of all three. First, we note that, due to 1., it is sufficient to establish that, for all $x \ge 0$, $P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) \ge 0) \to P(X(\infty) \le x \mid X(\infty) \ge 0)$, as $\lambda \to \infty$. 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 $\lambda \to \infty$ is then straightforward.

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

$$\pi_n^\lambda = \begin{cases} \frac{\lambda^n}{\mu_2^n n!}, & \text{if } 0 \le n \le N_2^\lambda, \\ \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \prod_{i=N_2^\lambda+1}^{n} \left(\mu_2 N_2^\lambda + (i - N_2^\lambda)\mu_1\right)}, & \text{if } N_2^\lambda < n \le N^\lambda - 1, \\ \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \left(N_1^\lambda \mu_1 + N_2^\lambda \mu_2\right)^{(n - N^\lambda + 1)} \prod_{i=N_2^\lambda+1}^{N^\lambda - 1} \left(\mu_2 N_2^\lambda + (i - N_2^\lambda)\mu_1\right)}, & \text{if } N^\lambda \le n. \end{cases} \tag{3.34}$$

1. For $\lambda > 0$, let $\alpha^\lambda = P(X^\lambda(\infty) \ge 0) = P(Y^\lambda(\infty) \ge N^\lambda) = \sum_{n \ge N^\lambda} p_n^\lambda$. It is then easy to see that

$$\alpha^\lambda = \frac{\sum_{n=N^\lambda}^{\infty} \pi_n^\lambda}{\sum_{n=0}^{\infty} \pi_n^\lambda} = \left[1 + \frac{\sum_{n=0}^{N_2^\lambda} \pi_n^\lambda + \sum_{n=N_2^\lambda+1}^{N^\lambda - 1} \pi_n^\lambda}{\sum_{n=N^\lambda}^{\infty} \pi_n^\lambda} \right]^{-1}.$$

Let

$$A^\lambda := \sum_{n=0}^{N_2^\lambda} \pi_n^\lambda,$$

$$B^\lambda := \sum_{n=N_2^\lambda+1}^{N^\lambda - 1} \pi_n^\lambda,$$

and

$$C^\lambda := \sum_{n=N^\lambda}^{\infty} \pi_n^\lambda,$$

then we need to show that $\left[1 + \frac{A^\lambda + B^\lambda}{C^\lambda}\right]^{-1} \to \alpha$ as $\lambda \to \infty$, or, equivalently, that $\frac{A^\lambda + B^\lambda}{C^\lambda} \to \frac{\delta/\sqrt{\mu_1}\Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})}$, as $\lambda \to \infty$. We look at C^λ first. Let $M^\lambda = \left[\mu_2 N_2^\lambda / \mu_1\right]$, $\rho^\lambda = \frac{\lambda}{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda}$, and let '$\approx$' denote two

quantities whose ratio goes to 1 in the limit, then,

$$C^\lambda = \sum_{n=N^\lambda}^{\infty} \pi_n^\lambda$$

$$= \sum_{n=N^\lambda}^{\infty} \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \left(N_1^\lambda \mu_1 + N_2 \mu_2^\lambda\right)^{(n-N^\lambda+1)} \prod_{i=N_2^\lambda+1}^{N^\lambda-1} \left(\mu_2 N_2^\lambda + (i-N_2^\lambda)\mu_1\right)}$$

$$\approx \frac{\lambda^{(N^\lambda-1)}}{\mu_2^{N_2^\lambda} N_2^\lambda! \mu_1^{(N_1^\lambda-1)} (M^\lambda+N_1^\lambda-1)!/M^\lambda!} \cdot \sum_{n=N^\lambda}^{\infty} \frac{\lambda^{(n-N^\lambda+1)}}{\left(\mu_1 N_1^\lambda + \mu_2 N_2^\lambda\right)^{(n-N^\lambda+1)}}$$

$$= \frac{\lambda^{(N^\lambda-1)} M^\lambda! \rho^\lambda}{\mu_2^{N_2^\lambda} N_2^\lambda! \mu_1^{(N_1^\lambda-1)} (M^\lambda+N_1^\lambda-1)!(1-\rho^\lambda)}$$

$$\approx \frac{\lambda^{(N^\lambda-1)} \sqrt{2\pi(\mu_2 N_2^\lambda/\mu_1)} \left(\mu_2 N_2^\lambda/\mu_1\right)^{(\mu_2 N_2^\lambda/\mu_1)} e^{-(\mu_2 N_2^\lambda/\mu_1)} \rho}{\mu_2^{N_2^\lambda} \sqrt{2\pi N_2^\lambda} (N_2^\lambda)^{N_2^\lambda} e^{-N_2^\lambda} \mu_1^{(N_1^\lambda-1)} \sqrt{2\pi(M^\lambda+N_1^\lambda-1)} \left(M^\lambda+N_1^\lambda-1\right)^{(M^\lambda+N_1^\lambda-1)} e^{-(M^\lambda+N_1^\lambda-1)} (1-\rho^\lambda)}$$

$$\approx \frac{\sqrt{\mu_2} \lambda^{N^\lambda} e^{(N^\lambda-1)} \left(\mu_2 N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}}{\sqrt{2\pi} \left(\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \mu_1\right)^{\left(\mu_2 N_2^\lambda/\mu_1 + N_1^\lambda\right)} \sqrt{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda} (1-\rho^\lambda)}.$$

The fifth line follows from Stirling's approximation. The rest is algebra. Note that, $\sqrt{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda}(1-\rho^\lambda) \to \delta$ as $\lambda \to \infty$, and hence

$$C^\lambda \approx \frac{\sqrt{\mu_2}(\lambda)^{N^\lambda} e^{N^\lambda-1} \left(\mu_2 N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}}{\sqrt{2\pi} \left(\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \mu_1\right)^{\left(\mu_2 N_2^\lambda/\mu_1 + N_1^\lambda\right)} \delta}.$$

We now proceed with developing approximation for B^λ.

$$B^\lambda = \sum_{n=N_2^\lambda+1}^{N^\lambda-1} \pi_n^\lambda$$

$$= \sum_{n=N_2^\lambda+1}^{N^\lambda-1} \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \prod_{i=N_2^\lambda+1}^{n} \left(\mu_2 N_2^\lambda + (i-N_2^\lambda)\mu_1\right)}$$

$$\approx \frac{\lambda^{N_2^\lambda} M^\lambda!}{\mu_2^{N_2^\lambda} N_2^\lambda!} \sum_{n=N_2^\lambda+1}^{N^\lambda-1} \frac{\lambda^{n-N_2^\lambda}}{\mu_1^{n-N_2^\lambda} (M^\lambda+n-N_2^\lambda)!}$$

$$= \frac{\lambda^{N_2^\lambda} M^\lambda! \mu_1^{M^\lambda} e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda! \lambda^{M^\lambda}} \sum_{j=M^\lambda+1}^{M^\lambda+N_1^\lambda-1} \frac{\lambda^j e^{-\lambda/\mu_1}}{\mu_1^j j!}.$$

Consider a Poisson random variable with rate λ/μ_1, then due to the central limit theorem, we have,

$$\begin{aligned}\sum_{j=M^\lambda+1}^{M^\lambda+N_1^\lambda-1}\frac{\lambda^j e^{-\lambda/\mu_1}}{\mu_1^j j!} &\approx \Phi\left(\frac{M^\lambda+N_1^\lambda-1-\lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right)-\Phi\left(\frac{M^\lambda+1-\lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right)\\ &\to\Phi(\delta/\sqrt{\mu_1})-\Phi(-\infty)=\Phi(\delta/\sqrt{\mu_1}).\end{aligned}$$

Hence,

$$\begin{aligned}B^\lambda &\approx \frac{M^\lambda!\mu_1^{M^\lambda}e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda}N_2^\lambda!\lambda^{\left(M^\lambda-N_2^\lambda\right)}}\Phi(\delta/\sqrt{\mu_1})\\ &\approx \frac{\sqrt{2\pi\mu_2N_2^\lambda/\mu_1}\left(\mu_2N_2^\lambda/\mu_1\right)^{\mu_2N_2^\lambda/\mu_1}e^{-\mu_2N_2^\lambda/\mu_1}\mu_1^{\mu_2N_2^\lambda/\mu_1}e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda}\sqrt{2\pi N_2^\lambda}\left(N_2^\lambda\right)^{N_2^\lambda}e^{-N_2^\lambda}\lambda^{N_2^\lambda(\mu_2/\mu_1-1)}}\Phi(\delta/\sqrt{\mu_1})\\ &= \frac{\sqrt{\mu_2}\left(\mu_2N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}e^{\lambda/\mu_1}}{\sqrt{\mu_1}\lambda^{N_2^\lambda(\mu_2/\mu_1-1)}e^{N_2^\lambda(\mu_2/\mu_1-1)}}\Phi(\delta/\sqrt{\mu_1}).\end{aligned}$$

Finally, we turn to the approximation of A^λ:

$$\begin{aligned}A^\lambda &= \sum_{n=0}^{N_2^\lambda}\pi_n^\lambda=\sum_{n=0}^{N_2^\lambda}\frac{\lambda^n}{\mu_2^n n!}\\ &= e^{\lambda/\mu_2}\sum_{n=0}^{N_2^\lambda}\frac{\lambda^n}{\mu_2^n n!}e^{-\lambda/\mu_2}\\ &\approx e^{\lambda/\mu_2}\Phi\left(\frac{N_2^\lambda-\lambda/\mu_2}{\sqrt{\lambda/\mu_2}}\right).\end{aligned}$$

Now we examine the ratio $\frac{A^\lambda}{C^\lambda}$.

$$\begin{aligned}\frac{A^\lambda}{C^\lambda} &\approx \frac{e^{\lambda/\mu_2}\ \Phi\left(\frac{N_2^\lambda-\lambda/\mu_2}{\sqrt{\lambda/\mu_2}}\right)}{\frac{\sqrt{\mu_2}(\lambda)^{N^\lambda}e^{N^\lambda-1}\left(\mu_2N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}}{\sqrt{2\pi}\left(\mu_1N_1^\lambda+\mu_2N_2^\lambda-\mu_1\right)^{\mu_2N_2^\lambda/\mu_1+N_1^\lambda}\cdot\delta}}\\ &\approx \frac{\sqrt{2\pi}\cdot\delta}{\sqrt{\mu_2}}\ \frac{e^{\lambda/\mu_2}\left(\lambda+\delta\sqrt{\lambda}\right)^{\frac{1}{\mu_1}\left(\lambda+\delta\sqrt{\lambda}\right)}\Phi\left(-\frac{a_1\sqrt{\lambda}}{\sqrt{\mu_2}}\right)}{(\lambda)^{\frac{\lambda}{\mu}+\sqrt{\lambda}\left(\frac{\delta_1^\lambda}{\mu_1}+\frac{\delta_2^\lambda}{\mu_2}\right)}\ e^{\frac{\lambda}{\mu}+\sqrt{\lambda}\left(\frac{\delta_1^\lambda}{\mu_1}+\frac{\delta_2^\lambda}{\mu_2}\right)-1}\left(a_2\lambda+\delta_2^\lambda\sqrt{\lambda}\right)^{\frac{a_2\lambda+\delta_2^\lambda\sqrt{\lambda}}{\mu_2}\left(\frac{\mu_2}{\mu_1}-1\right)}}\\ &= \frac{\sqrt{2\pi}\cdot\delta}{\sqrt{\mu_2}}\ \frac{\left(1+\frac{\delta}{\sqrt{\lambda}}\right)^{\frac{1}{\mu_1}\left(\lambda+\delta\sqrt{\lambda}\right)}\Phi\left(-\frac{a_1\sqrt{\lambda}}{\sqrt{\mu_2}}\right)}{e^{\frac{\lambda}{\mu}-\frac{\lambda}{\mu_2}-1+\sqrt{\lambda}\left(\frac{\delta_1^\lambda}{\mu_1}+\frac{\delta_2^\lambda}{\mu_2}\right)}\left(a_2+\delta_2^\lambda/\sqrt{\lambda}\right)^{\frac{a_2\lambda+\delta_2^\lambda\sqrt{\lambda}}{\mu_2}\left(\frac{\mu_2}{\mu_1}-1\right)}}.\end{aligned}$$

If $a_2 > 0$ then

$$\frac{A^\lambda}{C^\lambda} \approx \frac{\sqrt{2\pi}\delta}{\sqrt{\mu_2}}\, e^{-\lambda\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right)(a_1+a_2\log a_2)} \cdot \Phi\left(-\frac{a_1\sqrt{\lambda}}{\sqrt{\mu_2}}\right) \to 0, \text{ as } \lambda\to\infty,$$

where the convergence to zero follows from the fact that $a_1 + a_2 \log a_2 > 0$ and $-\frac{a_1\sqrt{\lambda}}{\sqrt{\mu_2}} \to -\infty$ as $\lambda \to \infty$. If $a_2 = 0$ then, by (2.35),

$$\frac{A^\lambda}{C^\lambda} \approx \frac{\sqrt{2\pi}\delta}{\sqrt{\mu_2}}\, e^{-\lambda\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right)\left(1-\frac{\delta_2^\lambda}{\sqrt{\lambda}}+\frac{\delta_2^\lambda}{\sqrt{\lambda}}\log\left(\frac{\delta_2^\lambda}{\sqrt{\lambda}}\right)\right)} \cdot \Phi\left(-\frac{a_1\sqrt{\lambda}}{\sqrt{\mu_2}}\right) \to 0, \text{ as } \lambda\to\infty.$$

Putting it all together, we have

$$\frac{A^\lambda + B^\lambda}{C^\lambda} \approx \frac{\frac{\sqrt{\mu_2}\left(\mu_2 N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}e^{\lambda/\mu_1}}{\sqrt{\mu_1}\lambda^{N_2^\lambda(\mu_2/\mu_1-1)}e^{N_2^\lambda(\mu_2/\mu_1-1)}}\Phi(\delta/\sqrt{\mu_1})}{\frac{\sqrt{\mu_2}\lambda^{N^\lambda}e^{N^\lambda-1}\left(\mu_2 N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}}{\sqrt{2\pi}\left(\mu_1 N_1^\lambda+\mu_2 N_2^\lambda\right)^{\left(\mu_2 N_2^\lambda/\mu_1+N_1^\lambda\right)}\delta}} \to \frac{\delta/\sqrt{\mu_1}\Phi(\delta/\sqrt{\mu_1})}{\phi(\delta/\sqrt{\mu_1})},$$

as $\lambda\to\infty$.

2. Let $x < 0$. Then $P(X^\lambda(\infty) \le x) = P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) < 0)P(X^\lambda(\infty) < 0) = P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) < 0)(1 - \alpha^\lambda)$. Based on this observation and 1., in order to establish weak convergence for negative values of x, we need to show that

$$P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) < 0) \to P(X(\infty) \le x \mid X(\infty) < 0) = \frac{\Phi\left(\delta/\sqrt{\mu_1} + \sqrt{\mu_1/\mu}x\right)}{\Phi\left(\delta/\sqrt{\mu_1}\right)}.$$

Let $y^\lambda(x) = \left[N^\lambda + x\sqrt{N^\lambda}\right]$, then for all x, we have

$$X^\lambda(\infty) \le x \Leftrightarrow \frac{Y^\lambda(\infty) - N^\lambda}{\sqrt{N^\lambda}} \le x \Leftrightarrow Y^\lambda(\infty) \le y^\lambda(x)\,.$$

Also, note that for all $x < 0$, we have $y^\lambda(x) \ge N_2^\lambda$ for all λ large enough. Hence, for such λ, we can write:

$$\begin{aligned} P(X^\lambda(\infty) \le x \mid X^\lambda(\infty) < 0) &= \frac{\sum_{n=0}^{y^\lambda(x)} \pi_n^\lambda}{\sum_{n=0}^{N^\lambda-1} \pi_n^\lambda} \\ &= \left[1 + \frac{\sum_{n=y^\lambda(x)+1}^{N^\lambda-1} \pi_n^\lambda}{\sum_{n=0}^{N_2^\lambda} \pi_n^\lambda + \sum_{n=N_2^\lambda+1}^{y^\lambda(x)} \pi_n^\lambda}\right]^{-1}. \end{aligned}$$

Let

$$\begin{aligned} A_-^\lambda &= \sum_{n=0}^{N_2^\lambda} \pi_n^\lambda \\ B_-^\lambda &= \sum_{n=y^\lambda(x)+1}^{N^\lambda-1} \pi_n^\lambda \\ C_-^\lambda &= \sum_{n=N_2^\lambda+1}^{y^\lambda(x)} \pi_n^\lambda. \end{aligned}$$

We have already shown that

$$A_-^\lambda = A^\lambda \approx e^{\lambda/\mu_2}\, \Phi\left(\frac{N_2^\lambda - \lambda/\mu_2}{\sqrt{\lambda/\mu_2}}\right).$$

Consider B_-^λ next.

$$\begin{aligned} B_-^\lambda &= \sum_{n=y^\lambda(x)+1}^{N^\lambda-1} \pi_n^\lambda \\ &= \sum_{n=y^\lambda(x)+1}^{N^\lambda-1} \frac{\lambda^n}{\mu_2^{N_2^\lambda} N_2^\lambda! \prod_{i=N_2^\lambda+1}^{n} \left(\mu_2 N_2^\lambda + \left(i - N_2^\lambda\right)\mu_1\right)} \\ &\approx \frac{(\lambda)^{N_2^\lambda} M^\lambda!\, \mu_1^{M^\lambda} e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda!\, \lambda^{M^\lambda}} \sum_{j=M^\lambda+y^\lambda(x)-N_2^\lambda+1}^{M^\lambda+N^\lambda-1-N_2^\lambda} \frac{\lambda^j\, e^{-\lambda/\mu_1}}{\mu_1^j\, j!}, \end{aligned}$$

where $M^\lambda = \left[\frac{\mu_2 N_2^\lambda}{\mu_1}\right]$. From the CLT we have

$$\begin{aligned} \sum_{j=M^\lambda+y^\lambda(x)-N_2^\lambda+1}^{M^\lambda+N^\lambda-1-N_2^\lambda} \frac{\lambda^j\, e^{-\lambda/\mu_1}}{\mu_1^j\, j!} &\approx \Phi\left(\frac{M^\lambda + N^\lambda - 1 - N_2^\lambda - \lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right) \\ &- \Phi\left(\frac{M^\lambda + y^\lambda(x) - N_2^\lambda + 1 - \lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right) \\ &\to \Phi\left(\delta/\sqrt{\mu_1}\right) - \Phi\left(\delta/\sqrt{\mu_1} + \sqrt{\mu_1/\mu}\, x\right). \end{aligned}$$

Therefore,

$$B_-^\lambda \approx \frac{\lambda^{N_2^\lambda} M^\lambda!\, \mu_1^{M^\lambda} e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda!\, \lambda^{M^\lambda}} \left(\Phi\left(\delta/\sqrt{\mu_1}\right) - \Phi\left(\delta/\sqrt{\mu_1} + \sqrt{\mu_1/\mu}\, x\right)\right).$$

We thus have,

$$\begin{aligned}
\frac{A_-^\lambda}{B_-^\lambda} &\approx \frac{e^{\lambda/\mu_2}\ \Phi\left(\frac{N_2^\lambda-\lambda/\mu_2}{\sqrt{\lambda/\mu_2}}\right)}{\frac{(\lambda)^{N_2^\lambda} M^\lambda!\ \mu_1^{M^\lambda} e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda!\ (\lambda)^{M^\lambda}}\ \left(\Phi\left(\delta/\sqrt{\mu_1}\right) - \Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}x\right)\right)} \\
&\approx \frac{(\lambda)^{N_2^\lambda(\mu_2/\mu_1-1)}\ e^{N_2^\lambda(\mu_2/\mu_1-1)-\lambda(1/\mu_1-1/\mu_2)}\ \Phi\left(-a_1\sqrt{\lambda}/\sqrt{\mu_2}+\delta_2/\sqrt{\mu_2}\right)}{\left(\mu_2 N_2^\lambda\right)^{N_2^\lambda(\mu_2/\mu_1-1)}\left(\Phi\left(\delta/\sqrt{\mu_1}\right) - \Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}x\right)\right)} \\
&= \frac{\Phi\left(-a_1\sqrt{\lambda}/\sqrt{\mu_2}+\delta_2/\sqrt{\mu_2}\right)}{\Phi\left(\delta/\sqrt{\mu_1}\right) - \Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}x\right)}\ e^{-\{\lambda[a_1+a_2\log(a_2+o(1))]+o(\lambda)\}(1/\mu_1-1/\mu_2)} \\
&\to 0, \text{ as } \lambda\to\infty.
\end{aligned}$$

We now turn to approximating C_-^λ.

$$\begin{aligned}
C_-^\lambda &= \sum_{n=N_2^\lambda+1}^{y^\lambda(x)} \pi_n^\lambda \\
&\approx \frac{\lambda^{N_2^\lambda} M^\lambda!\ \mu_1^{M^\lambda} e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda!\ \lambda^{M^\lambda}} \sum_{j=M^\lambda+1}^{M^\lambda+y^\lambda(x)-N_2^\lambda} \frac{\lambda^j\ e^{-\lambda/\mu_1}}{\mu_1^j\ j!}\ .
\end{aligned}$$

From the CLT we have,

$$\begin{aligned}
\sum_{j=M^\lambda+1}^{M^\lambda+y^\lambda(x)-N_2^\lambda} \frac{\lambda^j\ e^{-\lambda/\mu_1}}{\mu_1^j\ j!} &\approx \Phi\left(\frac{M^\lambda+y^\lambda(x)-N_2^\lambda-\lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right) \\
&- \Phi\left(\frac{M^\lambda+1-\lambda/\mu_1}{\sqrt{\lambda/\mu_1}}\right) \\
&\to \Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}\,x\right).
\end{aligned}$$

In particular,

$$C_-^\lambda \approx \frac{\lambda^{N_2^\lambda} M^\lambda!\ \mu_1^{M^\lambda}\ e^{\lambda/\mu_1}}{\mu_2^{N_2^\lambda} N_2^\lambda!\ \lambda^{M^\lambda}}\ \cdot\ \Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}\,x\right).$$

Finally,

$$\left[1+\frac{B_-^\lambda}{A_-^\lambda+C_-^\lambda}\right]^{-1} \rightarrow \left[1+\frac{\Phi\left(\delta/\sqrt{\mu}\right)-\Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}\,x\right)}{\Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}\,x\right)}\right]^{-1}$$
$$= \frac{\Phi\left(\delta/\sqrt{\mu_1}+\sqrt{\mu_1/\mu}\,x\right)}{\Phi\left(\delta/\sqrt{\mu_1}\right)}.$$

3. We now turn to approximating $P(X^\lambda(\infty)\le x)$ for $x\ge 0$. Clearly,

$$\begin{aligned} P(X^\lambda(\infty)\le x) &= P(X^\lambda(\infty)\le x \mid X^\lambda(\infty)\ge 0)\, P(X^\lambda(\infty)\ge 0) \\ &= P(X^\lambda(\infty)\le x \mid X^\lambda(\infty)\ge 0)\alpha^\lambda. \end{aligned}$$

Based on 1., it is hence sufficient to show that

$$\begin{aligned} P(X^\lambda(\infty)\le x \mid X^\lambda(\infty)\ge 0) &\rightarrow P(X(\infty)\le x \mid X(\infty)\ge 0) \\ &= 1-e^{-\frac{\delta}{\sqrt{\mu}}x}, \text{ as } \lambda\rightarrow\infty \text{ for all } x\ge 0. \end{aligned}$$

As before, if $y^\lambda(x)=\left[N^\lambda+x\sqrt{N^\lambda}\right]$, then

$$P(X^\lambda(\infty)\le x)=P(Y^\lambda(\infty)\le y^\lambda(x)).$$

Note that the Markov process, Y^λ, restricted to values above N^λ has a stationary distribution:

$$P\left(Y^\lambda(\infty)=n \mid Y^\lambda(\infty)\ge N^\lambda\right)=\left(1-\rho^\lambda\right)\left(\rho^\lambda\right)^{(n-N^\lambda)}, \quad n\ge N^\lambda.$$

Specifically, for $x\ge 0$,

$$\begin{aligned} P(X^\lambda(\infty)\le x \mid X^\lambda(\infty)\ge 0) &= \sum_{n=N^\lambda}^{y^\lambda(x)}\left(1-\rho^\lambda\right)\left(\rho^\lambda\right)^{(n-N^\lambda)} \\ &= 1-\left(\rho^\lambda\right)^{\left[\sqrt{N^\lambda}x\right]+1}\rightarrow 1-e^{-\frac{\delta x}{\sqrt{\mu}}}, \text{ as } \lambda\rightarrow\infty. \end{aligned}$$

∎

Remark 3.10 *Note that Proposition 3.6 also implies the weak convergence of the stationary distribution of $\vec{X}^\lambda$ to $\vec{X}$, which are both K dimensional processes. This is due to the state-space collapse that holds, in fact, for all $\lambda\ge 1$ (see Remark 3.2), 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 3.3 that if $X_k^\lambda(0)\Rightarrow X_k(0)$ for all $k=1,\ldots,K$, then $X^\lambda(\cdot)\Rightarrow X(\cdot)$ for $0\le t<\infty$. Our goal is to show that this convergence also prevails at $t=\infty$. This result is stated in the next proposition.

Proposition 3.7 ***(Convergence of the non-preemptive process in steady-state)*** *Suppose that conditions (2.14) and (2.33) 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 (3.32), as $\lambda \to \infty$.*

Proof: The proof is based on Ethier and Kurtz [19, Theorem 9.10 and Remark 9.11, p. 244]. According to [19] and based on our Propositions 3.2 and 3.3, 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 $\vec{X}^\lambda$, it is sufficient to establish that the state $(0,0)$ is positive recurrent, due to the irreducibility of the process (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 $ET_{(0,0)} \leq ET^P_{(0,0)}$, where $T^P_{(0,0)}$ is the equivalent of $T_{(0,0)}$ under FSF$_P$. The finiteness of $ET^P_{(0,0)}$ is known due to the existence of the stationary distribution of $\vec{X}$ under FSF$_P$ (which can be obtained from (3.34)). In particular,

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

   Recall the definition of $\tilde{\Pi}_P$ (given in Section 3.2) 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 $E\tilde{T}_{(0,0)} = ET^P_{(0,0)} < \infty$. We will show that $ET_{(0,0)} \leq E\tilde{T}_{(0,0)}$. The latter is true due to the following observations:

   a) The processes $\vec{X}(\cdot; \tilde{\pi})$ and $\vec{X}(\cdot; \text{FSF})$ both have state spaces which are subsets of $S = \mathbb{R}^2_- \cup (\mathbb{R}_+ \times \{0\}) \stackrel{\Delta}{=} S_- \cup S_+$ (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

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

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 $\vec{X}^\lambda$ 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 3.6, $\eta^\lambda(\cdot)$ 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 (see Figure 3.1 for illustration). From Proposition 3.1, we have $\nu^\lambda(K^+) \geq \eta^\lambda(K^+) \geq \eta^\lambda(K_0) \geq 1 - \tilde{\epsilon}$. This is almost what we need, except for the fact that K^+ is not compact, because it is not bounded from above.

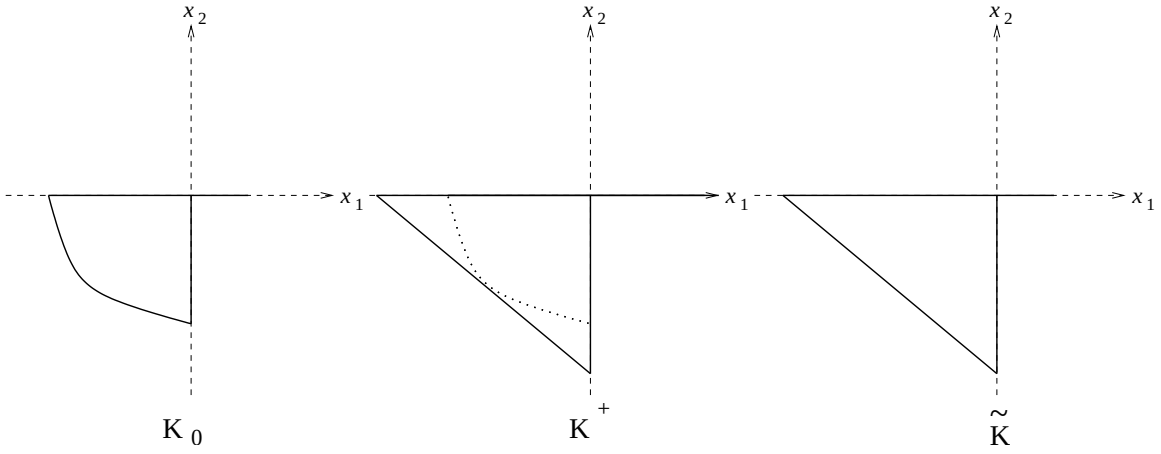

Figure 3.1: Illustration of the tightness proof.

Let $\bar{K}^+ = \{(x_1, 0) \in S \mid x_1 \geq 0, (x_1, 0) \notin K_0$, and $\exists (y_1, y_2) \in K_0$ with $y_1 + y_2 \leq x_1\}$. Then, $\bar{K}^+ \subseteq K^+$, and $K^+ \backslash \bar{K}^+$ 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 $\vec{X}^\lambda$ restricted to $\mathbb{R}_+ \times \{0\}$ are the same for both FSF$_P$ and FSF. Hence, $\nu^\lambda(K \mid \mathbb{R}_+ \times \{0\}) = \eta^\lambda(K \mid \mathbb{R}_+ \times \{0\})$ for all λ and any measurable set K. Specifically,

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

Here, α^λ_{NP} and α^λ_P are the steady-state probabilities of waiting for the λ system, under FSF and FSF$_P$, respectively. The first inequality follows from the fact that $\bar{K}^+ \cap K_0 = \phi$. The second inequality is due to Proposition 3.6, and particularly, the fact that $\alpha^\lambda_P \to \alpha$ as $\lambda \to \infty$. Finally, let $\tilde{K} = K^+ \setminus \bar{K}^+$, then $\tilde{K}$ is compact and

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

∎

Proof of Theorem 3.1: 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 $X^\lambda(\infty;\pi^\lambda)$, $\hat{X}^\lambda_0(\infty;\{\pi^\lambda\}) := Q^\lambda(\infty;\{\pi^\lambda\})/\sqrt{N^\lambda}$ and $\hat{W}^\lambda(\infty;\pi^\lambda)$, respectively, exist as $\lambda\to\infty$.

We prove the theorem in four steps:

1. First we show asymptotic optimality of FSF in terms of $X^\lambda(\infty)$, as $\lambda\to\infty$.

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 the (i) the steady-state 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 $\lambda\to\infty$. Recall that both (i) and (ii) were established in Proposition 3.7. The latter together with Proposition 3.6 also established that $X(\infty;\text{FSF}) = X(\infty;\text{FSF}_P) = \lim_{\lambda\to\infty} X^\lambda(\infty;\text{FSF}_P)$. 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^λ, as $\lambda\to\infty$.

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

$$P\left(\hat{X}_0(\infty;\text{FSF}) > q\right) \leq P\left(\hat{X}_0(\infty;\{\pi^\lambda\}) > q\right), \tag{3.35}$$

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 $\lambda > 0$.

3. For any sequence of policies $\{\pi^\lambda\}$, for which the steady state of $X^\lambda(\cdot;\pi^\lambda)$ exists for all λ, let $\tilde{\alpha}^\lambda = P(X^\lambda(\infty;\pi^\lambda) \geq N^\lambda)$, be the probability of having at least N^λ customers in the system. For work conserving policies $\tilde{\alpha}^\lambda = \alpha^\lambda = P^\lambda(wait > 0)$. Suppose that $X(\infty;\{\pi^\lambda\}) = \lim_{\lambda\to\infty} X^\lambda(\infty;\pi^\lambda)$ exists. Then 1. implies that

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

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

$$P\left(\hat{W}(\infty;\text{FSF}) > w\right) \leq P\left(\hat{W}(\infty;\{\pi^\lambda\}) > w\right). \tag{3.36}$$

To prove (3.36) 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 $\lambda\to\infty$ exists.

(iii) $\hat{W}(\infty;\text{FSF}) \overset{st}{\leq} \hat{W}(\infty;\{\pi^\lambda\})$.

(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 $\lambda\to\infty$, recall that by (3.5),

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

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

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

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

(iii) To show that $\hat{W}(\infty;\text{FSF}) \overset{st}{\leq} \hat{W}(\infty;\{\pi^\lambda\})$, 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

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

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

∎

Remark 3.11 *Note that the latter proof 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 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 3.2 shows a diagram of the asymptotic optimality relationships between the four entities included in the proof of Theorem 3.1.*

The next lemma establishes a simple relationship between the steady-state queue length and waiting time distributions for work conserving policies. This relationship is of the same form as the one shown in Proposition 3.4 for the transient limits of the queue length and waiting time processes.

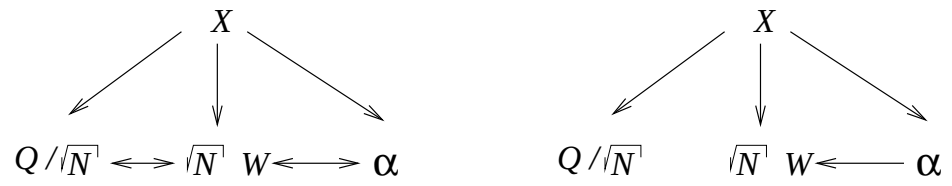

Figure 3.2: Asymptotic optimality relationships for work conserving policies (left) and in general (right).

Lemma 3.4 *Suppose that conditions (2.14) and (2.33) hold, and consider a sequence of policies* $\{\pi^\lambda\} \subseteq \Pi$, $\lambda > 0$. *Let* X, $\hat{X}_0$, *and* $\hat{W}$, *be the weak limits as* $\lambda \to \infty$ *for the steady state of the processes* $X^\lambda(\cdot;\pi^\lambda)$, $Q^\lambda(\cdot;\pi^\lambda)/\sqrt{N^\lambda}$, *and* $\sqrt{N^\lambda}W^\lambda(\cdot;\pi^\lambda)$, *respectively. Then,*

$$\hat{W} \stackrel{st}{\geq} \frac{[X]^+}{\mu}, \tag{3.37}$$

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

$$\hat{W} \stackrel{\mathcal{D}}{=} \frac{[X]^+}{\mu} = \frac{\hat{X}_0}{\mu}. \tag{3.38}$$

Proof: We prove (3.38). The relationship (3.37) 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

$$W^\lambda \stackrel{\mathcal{D}}{=} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} T_i^\lambda, \;\; \lambda > 0, \tag{3.39}$$

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 $\frac{[Y^\lambda - N^\lambda + 1]^+}{\sqrt{N^\lambda}} \Rightarrow [X]^+$. Let Y^λ, and X be versions of the original random variables such that the latter convergence is almost surely. For samples paths such that $Y^\lambda - N^\lambda + 1 \to \infty$ we have:

$$\sqrt{N^\lambda}W^\lambda \stackrel{\mathcal{D}}{=} \sqrt{N^\lambda} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} T_i^\lambda = \frac{[Y^\lambda - N_r + 1]^+}{\sqrt{N^\lambda}} \frac{1}{[Y^\lambda - N^\lambda + 1]^+} \sum_{i=1}^{[Y^\lambda - N^\lambda + 1]^+} N^\lambda T_i^\lambda \to \frac{[X]^+}{\mu},$$

almost surely, as $\lambda \to \infty$. 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, $\lim_{\lambda\to\infty} \frac{[Y^\lambda - N^\lambda + 1]^+}{\sqrt{N^\lambda}} = [X]^+ = 0$. 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

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

∎

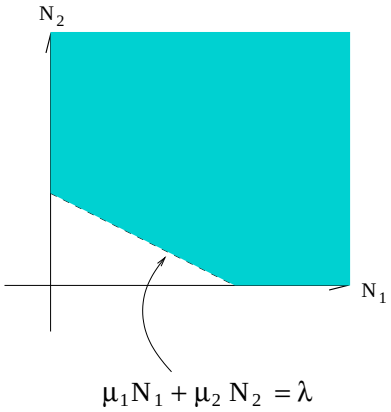

Figure 4.1: The Stability Region for $K = 2$.

4 Asymptotic Feasibility

In this section, we wish to characterize the feasible region for the staffing problem (2.3). As was noted before, characterizing this region exactly for fixed $\lambda, \mu_1, \ldots, \mu_K$, and α seems difficult. Instead, we characterize this region asymptotically, for large values of λ (i.e. as $\lambda \to \infty$).

We begin by recalling that a necessary and sufficient condition for stability is $\sum_{k=1}^{K} \mu_k N_k > \lambda$ (See Figure 4.1). This observation gives us a superset of the feasible region, as it is impossible to have a steady state waiting probability which is less than 1, if the system is unstable. Proposition 4.1 characterizes the asymptotic feasible region as a subset of the stability region. Although in principle, this region could be very complicated, it turns out to have a simple *linear* form. The linearity of the feasible region is not surprising in view of the fact that, under FSF the limiting waiting probability depends on the overall service capacity (as long as the slowest server pool is non-negligible). In particular, the limiting waiting probability does not depend on the individual capacities of the different server pools. Note that the overall service capacity is a linear function of the number of servers in each pool. Hence the linearity of the asymptotically feasible region. The asymptotically feasible region is illustrated in Figure 4.2.

Proposition 4.1 (Asymptotic Feasible Region - Square-Root Safety Capacity) *Let $0 < \alpha < 1$ and $\mu_1 < \mu_2 < \cdots < \mu_K$ be fixed, and consider a sequence of systems indexed by the arrival rate $\lambda > 0$, which is growing to infinity, and N_k^λ servers in pool k, $k = 1, \ldots, K$. Let $N^\lambda = \sum_{k=1}^{K} N_k^\lambda$ be the total number of servers in system λ, and suppose that*

$$\liminf_{\lambda \to \infty} \frac{N_1^\lambda}{N^\lambda} > 0. \tag{4.1}$$

Then, there exists a sequence $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N^\lambda})\}$ of non-preemptive policies, under which

$$\limsup_{\lambda \to \infty} P_{\pi^\lambda}(\textit{wait} > 0) \leq \alpha \tag{4.2}$$

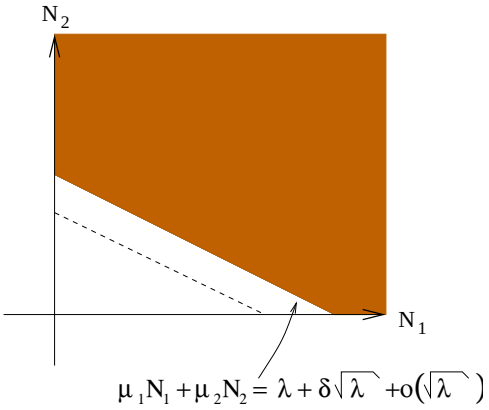

Figure 4.2: The Asymptotically Feasible Region for $K = 2$.

if and only if

$$\mu_1 N_1^\lambda + ... + \mu_K N_K^\lambda \geq \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right), \tag{4.3}$$

where $0 < \delta < \infty$ *satisfies*

$$\alpha \stackrel{\Delta}{=} \alpha(\delta/\sqrt{\mu_1}) = \left[1 + \frac{\left(\delta/\sqrt{\mu_1}\right)\Phi\left(\delta/\sqrt{\mu_1}\right)}{\phi\left(\delta/\sqrt{\mu_1}\right)}\right]^{-1}. \tag{4.4}$$

In addition, $\delta = 0$ *if and only if* $\alpha = 1$*, and* $\delta = \infty$ *(i.e. (4.3) holds for all* $\delta < \infty$*) if and only if* $\alpha = 0$*.*

Remark 4.1 *Notice that if one is interested in determining the total capacity needed for the system in order to obtain a target waiting probability* α*, then by (4.3) a safety capacity of* $\delta\sqrt{\lambda}$ *is needed beyond the minimal capacity of* λ *(hence, the term "square-root safety capacity"). According to (4.4 the value of* δ *is determined from the model parameters solely based on* α *and* μ_1*, which is the service rate of the slowest servers. The other, faster, service rates do not play a role at this stage. As will be shown later (Proposition 5.1) those faster service rates are needed in order to determine how to distribute this total capacity among the server pools in order to minimize staffing costs.*

Proof: We prove the proposition for $K = 2$. The general case follows similarly. Fix $0 < \delta < \infty$, and suppose that (4.3) holds for all λ. Let $a_k = \liminf_{\lambda\to\infty} \frac{\mu_k N_k^\lambda}{\lambda}$, $k = 1, 2$. Clearly, $a_1 + a_2 \geq 1$ and $a_1 > 0$. Suppose first that $a_1 + a_2 > 1$. In this case, we can obtain (4.2) with $\alpha = \alpha(\delta/\sqrt{\mu_1})$ by choosing to use only a subset of each server pool of size $\tilde{N}_k^\lambda = \frac{(a_k/(a_1+a_2))\lambda + (\delta/2)\sqrt{\lambda}}{\mu_k}$, $k = 1, 2$, and apply the policy FSF. Proposition 3.7 then confirms that (4.2) is satisfied. Now, suppose that $a_1 + a_2 = 1$, and without loss of generality, let $a_k = \lim_{\lambda\to\infty} \frac{\mu_k N_k^\lambda}{\lambda}$. Let $\tilde{\delta} = \liminf_{\lambda\to\infty} \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}}$ (again, without loss of generality, assume that $\tilde{\delta} = \lim_{\lambda\to\infty} \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}}$). Clearly, $\tilde{\delta} \geq \delta$, and possibly, $\tilde{\delta} = \infty$. If $\tilde{\delta} > \delta$, then one is able to obtain (4.2) by using FSF with respect to a subset of

each server pool of size $\tilde{N}_k^\lambda = \frac{\mu_k N_k^\lambda - (\Delta^\lambda/2)\sqrt{\lambda}}{\mu_k}$, where $\Delta^\lambda := \frac{\mu_1 N_1^\lambda + \mu_2 N_2^\lambda - \lambda}{\sqrt{\lambda}} - \delta$. Finally, if $\tilde{\delta} = \delta$, then (4.2) holds if FSF is used from Proposition 3.7.

To complete the proof, we need to examine the cases where $\delta = 0$ and $\delta = \infty$. Suppose first that $\delta = 0$, and assume, by contradiction, that there exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\}$, such that $\limsup_{\lambda\to\infty} P_{\pi^\lambda}(\text{wait} > 0) = \alpha < 1$. Let $0 < \delta_0 < \infty$ be such that $\alpha(\delta_0/\sqrt{\mu_1}) = \alpha$; such δ_0 exists due to the continuity of $\alpha(\cdot)$ and the fact that $\lim_{\delta\to\infty} \alpha(\delta/\sqrt{\mu_1}) = 0$ and $\lim_{\delta\to 0} \alpha(\delta/\sqrt{\mu_1}) = 1$. Consider another sequence of systems with server pools of size $\tilde{N}_k^\lambda = N_k^\lambda + (\delta_0/4)\sqrt{\lambda}/\mu_k$, $k = 1, 2$. Clearly, (4.3) holds for the new sequence, with $\delta = \delta_0/2$. Now, according to Proposition 3.7, if FSF is used with the new sequence of systems, $\lim_{\lambda\to\infty} P^\lambda_{\text{FSF}}(\text{wait} > 0) = \alpha(\delta_0/2\sqrt{\mu_1}) > \alpha$. However, $\{\pi^\lambda\}$ is assumed to obtain a waiting probability of α (asymptotically, over a subsequence) by using only a subset of the servers. This is a contradiction to the asymptotic optimality of FSF. Finally, if $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$ for all $\delta < \infty$, then by using FSF on a subset of the servers, one can obtain that $\limsup_{\lambda\to\infty} P^\lambda(\text{wait} > 0) \leq \alpha$, for all $0 < \alpha < 1$. Letting $\alpha \to 0$ establishes the desired result. ∎

Corollary 4.1 *Let $0 < \alpha < 1$ and $\mu_1 < \mu_2 < \cdots < \mu_K$ be fixed, and consider a sequence of systems indexed by the arrival rate λ, which is growing to infinity, and N_k^λ servers in pool k, $k = 1, \ldots, K$. Let $N^\lambda = \sum_{k=1}^K N_k^\lambda$ be the total number of servers in system λ, and suppose that (4.1) holds. Then, there exists a sequence $\{\pi^\lambda\}$ of non-preemptive policies, under which*

$$\lim_{\lambda\to\infty} P_{\pi^\lambda}(\textit{wait} > 0) = \alpha \tag{4.5}$$

if and only if

$$\mu_1 N_1^\lambda + ... + \mu_K N_K^\lambda = \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right), \tag{4.6}$$

where $0 < \delta < \infty$ satisfies $\alpha = \alpha(\delta/\sqrt{\mu_1})$ is given in (4.4).

In addition, $\delta = 0$ if and only if $\alpha = 1$, and $\delta = \infty$ (i.e. (4.3) holds for all $\delta < \infty$) if and only if $\alpha = 0$.

Proof: The proof follows immediately from Proposition 4.1. ∎

Remark 4.2 ***(Feasibility in the single server type system)*** *Consider, in comparison to our system, the sequence of systems described in Remark 3.8, with a single customer class and a single server pool, instead of K types. Suppose that all these servers have service rate $\mu = \sum_{k=1}^K \gamma_k \mu_k$, for some arbitrary weights $\gamma_1, ..., \gamma_K \in [0, 1]$, with $\sum_{k=1}^K \gamma_k = 1$. In addition, suppose that the sequence of arrival rates, $\{\lambda\}$ is identical for both models. For this single server type model, Halfin and Whitt [30] showed that if the number of servers in the λ system is N^λ, then for $\alpha \in (0, 1)$,*

$$\lim_{\lambda\to\infty} P^\lambda(wait > 0) = \alpha,$$

if and only if

$$N^\lambda = \lambda/\mu + \beta\sqrt{\lambda/\mu} + o(\sqrt{\lambda}),$$

where $0 < \beta < \infty$ *satisfies* $\alpha = \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} = \alpha(\beta)$, *(recall (4.4)). The condition on* N^λ *can be equivalently written as*

$$\mu N^\lambda = \lambda + \beta\sqrt{\mu}\sqrt{\lambda} + o(\sqrt{\lambda}) = \lambda + \delta\sqrt{\mu/\mu_1}\sqrt{\lambda} + o(\sqrt{\lambda}),$$

where δ *satisfies* $\alpha = \alpha(\delta/\sqrt{\mu_1})$. *Comparing this with (4.6), we conclude that the multiple server-type system requires less total service capacity than the one required by the single server-type with mean service rate, if both aim at achieving the same limiting steady-state waiting probability. Specifically, suppose that the single server-type system is compared with a* K *server-types system with* $N_k^\lambda = q_k(N_1^\lambda + ... + N_K^\lambda)$, *with* $q_k = \gamma_k = (a_k/\mu_k)\mu$, *for* $k = 1, ..., K$. *Then, assuming that* $q_1 > 0$ *(hence, condition (4.1) of Proposition 4.1 is satisfied), one can see that the multi-type systems requires overall fewer servers than the single-type system to achieve the same limiting steady-state waiting probability.*

5 Asymptotically Optimal Staffing

In this section, we study the staffing problem (2.3). Recall that exact optimality is difficult to obtain, and hence, we present asymptotically optimal solutions. Our previous results already identify, under certain conditions, an asymptotically optimal policy (FSF) and the asymptotic feasible region given in (4.3). It is now left to find the asymptotically optimal staffing rule that minimizes the staffing costs among all the vectors $\vec{N} = (N_1, \ldots, N_K)$, which belong to the feasible region. For the remainder of this section, consider a fixed target waiting probability $0 < \alpha < 1$.

Consider a cost function $C(\vec{N}) = C_1(N_1) + \cdots + C_K(N_K)$ which is increasing and strictly convex in all its arguments, and such that $C(\vec{N}) \to \infty$, as $\|\vec{N}\| \to \infty$. Because of the characterization of the feasible region given in (4.3), it is expected that the staffing cost will be at least of the order of $C(\lambda \cdot \vec{e})$, where $\vec{e}$ is a vector of 1's of dimension K. In addition, it is expected that differences between staffing costs of two different staffing vectors which are close to the efficient frontier of the feasible region, will be of the order of $C(\sqrt{\lambda} \cdot \vec{e})$. Hence, in order to establish a meaningful form of asymptotic optimality, one needs to compare *normalized* staffing costs that measure the difference between the actual staffing costs and a basic cost of order $C(\lambda \cdot \vec{e})$, which is a lower bound on the staffing cost.

To get such a lower bound, consider the following related problem:

$$\begin{aligned} \text{minimize} \quad & C_1(N_1) + C_2(N_2) + ...C_K(N_K) \\ \text{subject to} \quad & \mu_1 N_1 + \mu_2 N_2 + \cdots + \mu_K N_K \geq \lambda \\ & N_1, N_2, \ldots, N_K \geq 0, \end{aligned} \tag{5.1}$$

that is, we seek to minimize the staffing cost within the closure of the stability region. This problem, if accompanied by integral constraints, is a special case of a set covering problem. Without the integral constraints, its optimal solution, $\vec{N}^*$ satisfies

$$\frac{C_k'(N_k^*)}{\mu_k} = \frac{C_j'(N_j^*)}{\mu_j}, \quad j,k = 1,2,...,K, \tag{5.2}$$

and

$$\mu_1 N_1 + \mu_2 N_2 + \cdots + \mu_K N_K = \lambda. \tag{5.3}$$

Let $\underline{C}$ be the optimal cost obtained by solving (5.1). Then clearly, $\underline{C}$ is a lower bound on the solution of (2.3) because the stability region is a superset with respect to the feasible region. In addition, $\underline{C}$ is of order $C(\lambda \cdot \vec{e})$, and hence, it naturally serves as the right normalizing factor.

Definition: Consider a sequence of staffing vectors $\left\{\tilde{N}^\lambda\right\}_{\lambda>0}$. Then, $\left\{\tilde{N}^\lambda\right\}_{\lambda>0}$ is an *asymptotically optimal staffing* sequence if (i) it is asymptotically feasible, and (ii) its associated limiting staffing cost is minimal among all asymptotically feasible staffing vectors. More precisely, if $\left\{\vec{N}^\lambda\right\}_{\lambda>0}$ is another asymptotically feasible sequence of staffing vectors (that is, there exists a sequence of routing policies $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda)\} \subseteq \Pi$, such that $\limsup_{\lambda\to\infty} P_{\pi^\lambda}(wait > 0) \leq \alpha$), then $\lim_{\lambda\to\infty} \frac{C^\lambda(\tilde{N}^\lambda) - \underline{C}^\lambda}{C^\lambda(\vec{N}^\lambda) - \underline{C}^\lambda} \leq 1$.

Remark 5.1 (A "practical" definition of asymptotic optimal staffing) *The following definition is equivalent to the definition of asymptotic optimal staffing given above. In our proofs we use this definition, as it is easier to verify its validity. Suppose that* $\left\{\vec{N}^{*\lambda}\right\}_{\lambda>0}$ *is a sequence of optimal solutions of (2.3) with respect to sequences of arrival rates* $\{\lambda\}$ *and staffing cost functions* $\left\{C_1^\lambda(\cdot), ..., C_K^\lambda(\cdot)\right\}$. *Let* $\left\{\tilde{N}^\lambda\right\}_{\lambda>0}$ *be another sequence of staffing vectors. Then,* $\left\{\tilde{N}^\lambda\right\}_{\lambda>0}$ *is an* asymptotically optimal staffing sequence *if when used to staff the system,*

a. There exists a sequence of policies $\{\pi^\lambda = \pi^\lambda(\lambda, \tilde{N}^\lambda)\} \subseteq \Pi$ *such that* $\limsup_{\lambda\to\infty} P_{\pi^\lambda}($*wait* $> 0) \leq \alpha$, *and*

b. $\lim_{\lambda\to\infty} \frac{C^\lambda(\tilde{N}^\lambda) - \underline{C}^\lambda}{C^\lambda(\vec{N}^{*\lambda}) - \underline{C}^\lambda} = 1.$

We now investigate homogeneous cost functions of the form $C^\lambda(\vec{N}) \equiv C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, where $1 < p < \infty$, and $c_k > 0$ for $k = 1, \ldots, K$. Let $\delta > 0$ be such that $\alpha = \alpha(\delta/\sqrt{\mu_1})$, and let $\vec{M}^{*\lambda}$ be an optimal solution of the problem (5.1) with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Note that the vector $\vec{M}^{*\lambda}$ is not necessarily all integers, and let $\tilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil := (\lceil M_1^{*\lambda} \rceil, \ldots, \lceil M_K^{*\lambda} \rceil)$, that is $\tilde{N}^\lambda$ is obtained from $\vec{M}^{*\lambda}$ by rounding off its elements to the closest integers above. We claim that $\tilde{N}^\lambda$ is an asymptotically optimal staffing vector.

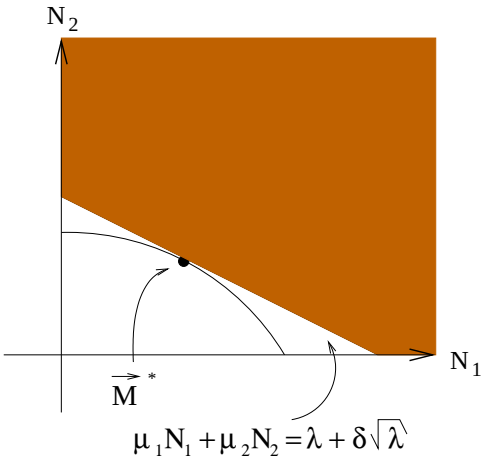

Figure 5.1: Asymptotic Cost Optimization for $K = 2$.

Proposition 5.1 ***(Asymptotically optimal staffing)*** *Consider a fixed target waiting probability of $\alpha \in (0,1)$. Suppose that $C(\vec{N}) = c_1 N_1^p + \cdots + c_K N_K^p$, and for $\lambda > 0$ consider the staffing vector $\tilde{N}^\lambda = \lceil \vec{M}^{*\lambda} \rceil$, where $\vec{M}^{*\lambda} = (M_1^{*\lambda}, \ldots, M_K^{*\lambda})$ is an optimal solution to (5.1), with the right hand side λ replaced by $\lambda + \delta\sqrt{\lambda}$. Here δ satisfies $\alpha = \alpha(\delta/\sqrt{\mu_1})$ (see (4.4)), and*

$$\vec{M}^{*\lambda} = (\lambda + \delta\sqrt{\lambda}) \frac{\left((\mu_1/c_1)^{1/(p-1)}, (\mu_2/c_2)^{1/(p-1)}, ..., (\mu_K/c_K)^{1/(p-1)}\right)}{\sum_{k=1}^K (\mu_k^p/c_k)^{1/(p-1)}}, \quad \lambda > 0. \tag{5.4}$$

Then $\left\{\tilde{N}^\lambda\right\}_{\lambda>0}$ is an asymptotically optimal staffing sequence.

Proof: We prove the proposition for the case $K = 2$. The general case follows similarly. Let $\vec{M}^{*\lambda}$ be the non-negative vector on the half-plain $\mu_1 M_1 + \mu_2 M_2 \geq \lambda + \delta\sqrt{\lambda}$ that minimizes the staffing cost $C(M)$, $\lambda > 0$. Clearly, $\mu_1 M_1^{*\lambda} + \mu_2 M_2^{*\lambda} = \lambda + \delta\sqrt{\lambda}$. Let $\tilde{N}_k^\lambda = \lceil M_k^{*\lambda} \rceil$, $k = 1, 2$. We prove that $\lim_{\lambda\to\infty} \frac{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda}{C(\vec{N}^{*\lambda}) - \underline{C}^\lambda} = 1$. The asymptotic optimality of $\tilde{N}^\lambda$ then easily follows. The outline of the proof is as follows:

1. We solve for $\underline{C}^\lambda$, $\vec{M}^{*\lambda}$, and $C(\vec{M}^{*\lambda})$ for all $\lambda > 0$, and show that $\lceil \vec{M}^{*\lambda} \rceil$ satisfies the conditions of Proposition 4.1. Solving for $\vec{M}^{*\lambda}$ is illustrated in Figure 5.1.

2. Assuming first that $\liminf_{\lambda\to\infty} \frac{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda}{C(\vec{N}^{*\lambda}) - \underline{C}^\lambda} < 1$, we show that there exist $\lambda_0 > 0$ and a vector $\vec{L}^{\lambda_0}$ such that if it is used to staff the λ_0 system, then

   a. there exists a policy $\pi = \pi(\lambda_0, \vec{L}^{\lambda_0}) \in \Pi$ such that $P_\pi(\text{wait} > 0) < \alpha$, and

   b. $C(\vec{L}^{\lambda_0}) < C(\vec{N}^{*\lambda_0})$.

   This, of course, contradicts the optimality of $\vec{N}^{*\lambda_0}$.

3. If $\limsup_{\lambda\to\infty} \frac{C(\vec{M}^{*\lambda})-\underline{C}^{\lambda}}{C(\vec{N}^{*\lambda})-\underline{C}^{\lambda}} > 1$, then we establish (see Lemmas 5.1 and 5.2), with respect to the sequence of vectors $\left\{\vec{N}^{*\lambda}\right\}_{\lambda>0}^{\infty} = \left\{(N_1^{*\lambda}, N_2^{*\lambda})\right\}_{\lambda>0}$, that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$, for all λ, along the subsequence.

   The case $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} \geq \lambda + \delta\sqrt{\lambda}$ can be ruled out by the optimality of $\vec{M}^{*\lambda}$. In particular, $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$. In this case, we find another vector $\vec{L}^{\lambda}$, such that

   a. $\mu_1 L_1^{\lambda} + \mu_2 L_2^{\lambda} = \lambda + \delta\sqrt{\lambda}$, for all λ along the subsequence, and

   b. $\lim_{\lambda\to\infty} \frac{C(\vec{L}^{\lambda})-\underline{C}^{\lambda}}{C(\vec{N}^{*\lambda})-\underline{C}^{\lambda}} = 1$, for the same sequence that attains the limsup.

   But this contradicts the optimality of $\vec{M}^{*\lambda}$.

   We now turn to the details of the three steps of the proof.

1. To find $\underline{C}^{\lambda}$, one needs to solve the problem (5.1). Simple constrained optimization obtains:

$$\underline{C}^{\lambda} = \frac{\lambda^p c_1 c_2}{\left((\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)}\right)^{p-1}} \stackrel{\Delta}{=} \lambda^p \xi. \tag{5.5}$$

   Similarly, to find $\vec{M}^{*\lambda}$ and $C(\vec{M}^{*\lambda})$ one needs to solve the problem:

$$\begin{array}{lrcl} \text{minimize} & c_1 M_1^p + c_2 M_2^p & & \\ \text{subject to} & \mu_1 M_1 + \mu_2 M_2 & \geq & \lambda + \delta\sqrt{\lambda} \\ & M_1, M_2 & \geq & 0\,. \end{array} \tag{5.6}$$

   The solution to (5.6) is given in (5.4), and for $K = 2$ it satisfies

$$\left(M_1^{*\lambda}, M_2^{*\lambda}\right) = (\lambda + \delta\sqrt{\lambda}) \cdot \frac{\left((\mu_1 c_2)^{1/(p-1)}, (\mu_2 c_1)^{1/(p-1)}\right)}{(\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)}},$$

   and

$$C(\vec{M}^{*\lambda}) = (\lambda + \delta\sqrt{\lambda})^p \xi. \tag{5.7}$$

   In particular, $\lceil \vec{M}^{*\lambda} \rceil$ satisfies condition (4.1) of Proposition 4.1, because

$$\frac{M_1^{*\lambda}}{M_1^{*\lambda} + M_2^{*\lambda}} \equiv \frac{(\mu_1 c_2)^{1/(p-1)}}{(\mu_1 c_2)^{1/(p-1)} + (\mu_2 c_1)^{1/(p-1)}} > 0.$$

2. Suppose that $\liminf_{\lambda\to\infty} \frac{C(\vec{M}^{*\lambda})-\underline{C}^{\lambda}}{C(\vec{N}^{*\lambda})-\underline{C}^{\lambda}} < 1$. Without loss of generality, assume that $\lim_{\lambda\to\infty} \frac{C(\vec{M}^{*\lambda})-\underline{C}^{\lambda}}{C(\vec{N}^{*\lambda})-\underline{C}^{\lambda}} < 1$. This implies that there exists $\Delta > 0$ such that $\frac{C(\vec{N}^{*\lambda})-\underline{C}^{\lambda}}{C(\vec{M}^{*\lambda})-\underline{C}^{\lambda}} \geq 1 + \Delta$, for all λ large enough, or, $C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda}) \geq \Delta(C(\vec{M}^{*\lambda}) - \underline{C}^{\lambda})$, for all λ large enough.

Let $\epsilon = \frac{\Delta\delta}{2}$, and let $\tilde{M}^\lambda$ be the optimal solution of (5.6) with $\delta+\epsilon$ replacing δ. Note that $C(\tilde{M}^\lambda) \geq C(\vec{M}^{*\lambda})$ and that

$$\begin{aligned}
\frac{C(\tilde{M}^\lambda) - C(\vec{M}^{*\lambda})}{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda} &= \frac{(\lambda + (\delta+\epsilon)\sqrt{\lambda})^p - (\lambda+\delta\sqrt{\lambda})^p}{(\lambda+\delta\sqrt{\lambda})^p - \lambda^p} \\
&= \frac{(1 + (\delta+\epsilon)/\sqrt{\lambda})^p - (1+\delta/\sqrt{\lambda})^p}{(1+\delta/\sqrt{\lambda})^p - 1} \\
&= \frac{1 + p(\delta+\epsilon)/\sqrt{\lambda} + o(1/\sqrt{\lambda}) - 1 - p\delta/\sqrt{\lambda} + o(1/\sqrt{\lambda})}{1 + p\delta/\sqrt{\lambda} + o(1/\sqrt{\lambda}) - 1} \\
&= \frac{p\epsilon + o(1)}{p\delta + o(1)} \leq \frac{3\Delta}{4}\,, \text{ for all } \lambda \text{ large enough.}
\end{aligned}$$

Specifically, for all λ large enough, we have,

$$\begin{aligned}
C(\tilde{M}^\lambda) - C(\vec{M}^{*\lambda}) &\leq \frac{3\Delta}{4}(C(\vec{M}^{*\lambda}) - \underline{C}^\lambda) \\
&< \Delta(C(\vec{M}^{*\lambda}) - \underline{C}^\lambda) \\
&\leq C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})\,.
\end{aligned}$$

Let $\vec{L}^\lambda = \lceil \tilde{M}^\lambda \rceil$, then for all λ large enough $C(\vec{L}^\lambda) < C(\vec{N}^{*\lambda})$, and $\vec{L}^\lambda$ satisfies the conditions of Proposition 4.1, with $\delta+\epsilon$ replacing δ. Hence, under staffing of $\vec{L}^\lambda$, $P^\lambda_{\text{FSF}}(\text{wait} > 0) \to \alpha((\delta+\epsilon)/\sqrt{\mu_1}) < \alpha$. In particular, for all λ large enough, under staffing of $\vec{L}^\lambda$, we have $P^\lambda_{\text{FSF}}(\text{wait} > 0) < \alpha$. This is a contradiction to the optimality of $\vec{N}^{*\lambda}$.

Before we turn to step 3 of the proof, we state and prove two lemmas.

Lemma 5.1 *Suppose that for all $\lambda > 0$, $\vec{N}^{*\lambda}$ is the optimal solution of (2.3) and $\liminf_{\lambda\to\infty} \frac{N_1^{*\lambda}}{N_1^{*\lambda}+N_2^{*\lambda}} > 0$. Then, $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} = \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$.*

Proof: By contradiction, assume that either there exists a subsequence $\{\lambda_j\}$ for which $\mu_1 N_1^{*\lambda_j} + \mu_2 N_2^{*\lambda_j} < \lambda_j + \delta\sqrt{\lambda_j} + o(\sqrt{\lambda_j})$, or there exists $\tilde{\epsilon} > 0$ such that $\mu_1 N_1^{*\lambda_j} + \mu_2 N_2^{*\lambda_j} \geq \lambda_j + (\delta + \tilde{\epsilon})\sqrt{\lambda_j} + o(\sqrt{\lambda_j})$. In the first case, by Proposition 4.1, $\limsup_{j\to\infty} P^{\lambda_j}_{\pi^{\lambda_j}}(\text{wait} > 0) > \alpha$, for all $\pi^{\lambda_j} \in \Pi$, which is a contradiction to the feasibility of $\vec{N}^{*\lambda_j}$, for some large values of j. In the second case, let $\vec{N}^{\lambda_j} = \vec{N}^{*\lambda_j} - \vec{e}$ (where $\vec{e}$ is a vector of 1's). Then $C(\vec{N}^{\lambda_j}) < C(\vec{N}^{*\lambda_j})$, and by Proposition 4.1, there exists a sequence of policies $\{\pi^{\lambda_j} = \pi^{\lambda_j}(\lambda_j, \vec{N}^{\lambda_j})\} \subseteq \Pi$ under which $\limsup_{j\to\infty} P_{\pi^{\lambda_j}}(\text{wait} > 0) < \alpha$. This is a contradiction to the optimality of $\vec{N}^{*\lambda_j}$ for all large j. ∎

Lemma 5.2 *Suppose that for a sequence $\vec{N}^\lambda$ of staffing vectors, such that there exists a sequence $\{\pi^\lambda = \pi^\lambda(\lambda, \vec{N}^\lambda\} \subseteq \Pi$ of policies under which $P_{\pi^\lambda}(\text{wait} > 0) \leq \alpha$, for all $\lambda > 0$. Suppose, in addition, that $\liminf_{\lambda\to\infty} \frac{N_1^\lambda}{N_1^\lambda+N_2^\lambda} = 0$. Then, $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$, where $\alpha = \alpha(\delta/\sqrt{\mu_1})$.*

Proof: Suppose that (without loss of generality), $\lim_{\lambda\to\infty}\frac{N_1^\lambda}{N_1^\lambda+N_2^\lambda}=0$. First note that, from stability, there exists $0\le\tilde\delta<\infty$ such that $\mu_1N_1^\lambda+\mu_2N_2^\lambda\ge\lambda+\tilde\delta\sqrt{\lambda}+o(\sqrt{\lambda})$, as $\lambda\to\infty$. In this case, we can show that for $\delta_2=\lim_{\lambda\to\infty}\delta_2^\lambda$, $-\infty\le\delta_2\le\tilde\delta$ (consider δ_2 as a partial limit of δ_2^λ, if the limit does not exist), then under FSF$_P$,

$$\begin{aligned}
&\frac{1}{P^\lambda(\text{wait}>0)}-1\\
\to\ &\frac{1}{\alpha(\tilde\delta/\sqrt{\mu_1})}-1+\sqrt{2\pi}\tilde\delta e^{\frac{1}{2}\tilde\delta^2/\mu_1}\left(\frac{1}{\sqrt{\mu_2}}e^{-\frac{1}{2}\delta_2^2\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right)}\Phi\left(\delta_2/\sqrt{\mu_2}\right)-\frac{1}{\sqrt{\mu_1}}\Phi\left(\delta_2/\sqrt{\mu_1}\right)\right),\ \text{as }\lambda\to\infty,\\
\le\ &\frac{1}{\alpha(\tilde\delta/\sqrt{\mu_1})}-1.
\end{aligned}$$

In particular, $\lim_{\lambda\to\infty}P^\lambda(\text{wait}>0)\ge\alpha(\tilde\delta/\sqrt{\mu_1})$. Therefore, from the optimality of FSF$_P$ in Π_P (see Proposition 3.1), we also have, $\limsup_{\lambda\to\infty}P_{\pi^\lambda}(\text{wait}>0)\ge\alpha(\tilde\delta/\sqrt{\mu_1})$. But we assumed that $P_{\pi^\lambda}(\text{wait}>0)\le\alpha(\delta/\sqrt{\mu_1})$ for all λ. Hence, $\alpha(\delta/\mu_1)\ge\alpha(\tilde\delta/\mu_1)$, or equivalently, $\delta\le\tilde\delta$. Finally, the latter implies that $\mu_1N_1^\lambda+\mu_2N_2^\lambda\ge\lambda+\delta\sqrt{\lambda}+o(\sqrt{\lambda})$. ∎

We now return to the third step of the proof of Proposition 5.1.

3. Suppose that $\limsup_{\lambda\to\infty}\frac{C(\vec M^{*\lambda})-\underline{C}^\lambda}{C(\vec N^{*\lambda})-\underline{C}^\lambda}>1$. Without loss of generality, suppose that $\lim_{\lambda\to\infty}\frac{C(\vec M^{*\lambda})-\underline{C}^\lambda}{C(\vec N^{*\lambda})-\underline{C}^\lambda}>1$. Due to the definition of $\vec M^{*\lambda}$ it follows that $\mu_1N_1^{*\lambda}+\mu_2N_2^{*\lambda}<\lambda+\delta\sqrt{\lambda}$ for all λ large enough. From Lemmas 5.1 and 5.2, we know that $\mu_1N_1^{*\lambda}+\mu_2N_2^{*\lambda}\ge\lambda+\delta\sqrt{\lambda}+o(\sqrt{\lambda})$. Let $f^\lambda(\lambda)=\lambda+\delta\sqrt{\lambda}-(\mu_1N_1^{*\lambda}+\mu_2N_2^{*\lambda})$. Then, $f^\lambda(\lambda)=o(\sqrt{\lambda})\ge0$. Let $c^\lambda:=\frac{\lambda+\delta\sqrt{\lambda}}{\mu_1N_1^{*\lambda}+\mu_2N_2^{*\lambda}}=\left(1-\frac{f^\lambda(\lambda)}{\lambda+\delta\sqrt{\lambda}}\right)^{-1}$, and consider the vector $\vec L^\lambda=c^\lambda\cdot\vec N^{*\lambda}$. Note that $\mu_1L_1^\lambda+\mu_2L_2^\lambda=\lambda+\delta\sqrt{\lambda}$, and that $C(\vec L^\lambda)\ge C(\vec N^{*\lambda})$. Hence, we have

$$\begin{aligned}
1\ &\ge\ \frac{C(\vec N^{*\lambda})-\underline{C}^\lambda}{C(\vec L^\lambda)-\underline{C}^\lambda}=\frac{\frac{1}{(c^\lambda)^p}C(\vec L^\lambda)-\underline{C}^\lambda}{C(\vec L^\lambda)-\underline{C}^\lambda}=1-\frac{\left(1-\left(1-f^\lambda(\lambda)/(\lambda+\delta\sqrt{\lambda})\right)^p\right)C(\vec L^\lambda)}{C(\vec L^\lambda)-\underline{C}^\lambda}\\
&=\ 1-\frac{\left(pf^\lambda(\lambda)/(\lambda+\delta\lambda)+o(1/\lambda)\right)C^\lambda(\vec L)}{C(\vec L^\lambda)-\underline{C}^\lambda}\approx1-\frac{pf^\lambda(\lambda)/(\lambda+\delta\lambda)}{1-\underline{C}^\lambda/C(\vec L^\lambda)}\\
&\ge\ 1-\frac{pf^\lambda(\lambda)/(\lambda+\delta\lambda)}{1-\underline{C}^\lambda/C(\vec M^{*\lambda})}=1-\frac{pf^\lambda(\lambda)/(\lambda+\delta\lambda)}{1-\left(1/(1+\delta/\sqrt{\lambda})\right)^p}\\
&=\ 1-\frac{pf^\lambda(\lambda)/(\sqrt{\lambda}+\delta)}{\sqrt{\lambda}\left(1-\left(1/(1+\delta/\sqrt{\lambda})\right)^p\right)}\ge1-\frac{pf^\lambda(\lambda)/(\sqrt{\lambda}+\delta)}{\delta/2},\qquad\text{for all }\lambda\text{ large enough}\\
&\to\ 1,\qquad\text{as }\lambda\to\infty.
\end{aligned}$$

But the latter implies that, in particular, $C(\vec L^\lambda)<C(\vec M^{*\lambda})$ for all λ large enough, which is a contradiction to the optimality of $\vec M^{*\lambda}$. ∎

Example 5.1 ***Quadratic Cost Functions:*** *Consider the staffing problem:*

$$\begin{array}{ll} \textit{minimize} & c_1 N_1^2 + c_2 N_2^2 + ... + c_K N_K^2 \\ \textit{subject to} & P_\pi(wait > 0) \leq \alpha, \ \textit{for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & N_1, N_2, ..., N_K \in \mathbb{Z}_+, \end{array} \tag{5.8}$$

with a fixed target waiting probability $0 < \alpha < 1$. To emphasize the dependence of the staffing level on the arrival rate λ, we denote our proposed solution by $\vec{N}^\lambda$. To determine the total capacity needed to satisfy the waiting probability constraint, Proposition 4.1 suggests that

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + ... + \mu_K N_k^\lambda \geq \lambda + \delta\sqrt{\lambda} + o(\lambda), \ \textit{ as } \lambda \to \infty,$$

where δ satisfies $\alpha = \alpha(\delta/\sqrt{\mu_1})$. That is, the total capacity required to achieve the target waiting probability depends asymptotically on the service rates through the service rate μ_1 of the slow servers only. To determine the actually staffing level, one needs to take into account the actual individual service rates. By Proposition 5.1 the proposed staffing vector $\vec{N}^\lambda$ which satisfies:

$$\frac{N_k^\lambda}{N_j^\lambda} = \frac{c_j/\mu_j}{c_k/\mu_k}, \ \ k, j = 1, 2, ..., K, \tag{5.9}$$

and

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda + ... + \mu_K N_k^\lambda = \lambda + \delta\sqrt{\lambda},$$

is asymptotically optimal among all asymptotically feasible vectors. The verbal interpretation of (5.9) is that when the staffing cost is quadratic, then staffing levels for individual server pools are inversely proportional to the ratio c_k/μ_k. This rule is intuitive as it implies that when the cost per unit of service rate is high, the staffing level should be low. Note that the ratio c/μ is not to be confused with the quantity $c\mu$ often used (in different contexts) to determine routing rules when holding costs is associated with waiting customers.

Extensions

Arrival Rate Dependent Homogeneous Cost Functions: Suppose that, instead of the fixed staffing cost function considered in Proposition 5.1, a cost function which is dependent on the arrival rate. We capture this dependence through the superscript λ. Particularly, consider, for $\lambda > 0$, the staffing cost function is $C^\lambda(\vec{N}) = c_1^\lambda N_1^{p^\lambda} + c_2^\lambda N_2^{p^\lambda} + ... + c_K^\lambda N_K^{p^\lambda}$. For $\lambda > 0$ and $k = 1, ..., K$, assume that $c_k^\lambda > 0$, $\liminf_{\lambda\to\infty} c_k^\lambda > 0$, $p^\lambda > 1$, $\liminf_{\lambda\to\infty} p^\lambda > 1$, and $\limsup_{\lambda\to\infty} p^\lambda < \infty$.

In this case, one can verify that the sequence of staffing vectors $\lceil \vec{M}^{*\lambda} \rceil$ proposed in (5.4) - with superscripts λ accompanying c_k and p, is asymptotically optimal staffing. Specifically, let

$$\vec{M}^{*\lambda} = (\lambda + \delta\sqrt{\lambda}) \frac{\left((\mu_1/c_1^\lambda)^{1/(p^\lambda - 1)}, (\mu_2/c_2^\lambda)^{1/(p^\lambda - 1)}, ..., (\mu_K/c_K^\lambda)^{1/(p^\lambda - 1)} \right)}{\sum_{k=1}^K \left(\mu_k^{p^\lambda}/c_k^\lambda \right)^{1/(p^\lambda - 1)}}, \ \ \lambda \geq 1, \tag{5.10}$$

then one can show that $\lceil \vec{M}^{*\lambda} \rceil$ is asymptotically optimal.

Linear Cost Functions with Constraints: In many practical situations one is interested in determining staffing levels to minimize linear staffing costs. This is the case where the staffing costs are associated, for example, with salaries of the servers. However, the linear cost case has not been included in our discussion so far. To illustrate why this case is problematic within our framework, consider the following example: suppose that one is interested in solving the staffing problem

$$\begin{array}{ll} \text{minimize} & c_1N_1 + c_2N_2 + ... + c_KN_K \\ \text{subject to} & P_\pi(wait > 0) \leq \alpha, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & N_1, N_2, ..., N_K \in \mathbb{Z}_+, \end{array} \tag{5.11}$$

for some fixed value of $0 < \alpha < 1$. If one, instead, solved the deterministic problem:

$$\begin{array}{ll} \text{minimize} & c_1N_1 + c_2N_2 + ... + c_KN_K \\ \text{subject to} & \mu_1N_1 + \mu_2N_2 + ... + \mu_KN_K \geq \lambda + \delta\sqrt{\lambda}, \\ & N_1, N_2, ..., N_K \geq 0, \end{array} \tag{5.12}$$

then any optimal solution $\vec{N}^\lambda$ will satisfy:

$$N_k^\lambda > 0 \text{ only if } \frac{c_k}{\mu_k} = \min_{j=1,...,K}\left\{\frac{c_j}{\mu_j}\right\}, \quad k = 1, 2, ..., K.$$

In particular, if $\frac{c_1}{\mu_1} \neq \min_j\left\{\frac{c_j}{\mu_j}\right\}$, then $N_1^\lambda = 0$ for all λ. The problem in this case is that one is no longer guaranteed that the proposed staffing vector is asymptotically *feasible* because condition (4.1) of Proposition 4.1 is not satisfied. In fact, one can show that when $N_1^\lambda = 0$ one needs higher overall capacity level in order to get the same limiting waiting probability.

Note that if $\frac{c_1}{\mu_1} = \min_j\left\{\frac{c_j}{\mu_j}\right\}$, then one can choose N_1^λ to be non-negligible relatively to the other server pools, and then the proposed solution is indeed asymptotically optimal (the proof follows through similarly to the proof of Proposition 5.1). However, even in the case that $\frac{c_1}{\mu_1} \neq \min_j\left\{\frac{c_j}{\mu_j}\right\}$, there are scenarios where are theory can provide useful solutions. Consider the following staffing problem with linear staffing costs and additional linear constraints:

$$\begin{array}{ll} \text{minimize} & c_1N_1 + c_2N_2 + ... + c_KN_K \\ \text{subject to} & P_\pi(wait > 0) \leq \alpha, \text{ for some } \pi = \pi(\lambda, \vec{N}) \in \Pi, \\ & A\vec{N} \geq b \\ , & N_1, N_2, ..., N_K \in \mathbb{Z}_+, \end{array} \tag{5.13}$$

where A is a $i \times K$ matrix, and b is an $i-$ dimensional matrix for some $i \geq 1$. Examples for such additional constraints can include $N_1/(N_1 + ... + N_K) \geq p$ for some $0 < p < 1$, or $l_k \leq N_k/(N_1 + ... + N_K) \leq u_k$ for some $0 \leq l_k \leq u_k \leq 1$. The first example can result from a case where servers of pool 1 are trainees, and one wants to make sure that they get the experience the need. The second set of constraints can result out of given proportions of servers types in the particular server population.

For the problem (5.13) we claim that if the set of problems

$$\begin{array}{ll} \text{minimize} & c_1 N_1 + c_2 N_2 + ... + c_K N_K \\ \text{subject to} & \mu_1 N_1 + \mu_2 N_2 + ... + \mu_K N_K \geq \lambda + \delta\sqrt{\lambda}, \\ & A\vec{N} \geq b \\ , & N_1, N_2, ..., N_K \in \mathbb{Z}_+, \end{array} \tag{5.14}$$

has a sequence of solutions $\vec{N}^\lambda$ which satisfy $\liminf_{\lambda\to\infty} N_1^\lambda/(N_1^\lambda + ... + N_K^\lambda)$ then the proposed sequence is an asymptotically optimal staffing. The proof follows similarly to the proof of Proposition 5.1.

References

[1] Aksin, O.Z. and Karaesmen A.F. (2002), Designing Flexibility: Characterizing the Value of Cross-Training Practices, working paper, INSEAD.

[2] M. Armony and N. Bambos (2001), Queueing Dynamics and Maximal Throughput Scheduling in Switched Processing Systems, Technical Report SU NETLAB-2001-09/01, Engineering Library, Stanford University, Stanford, CA 94305; September 29, 2001, to appear in *Queueing Systems*.

[3] Armony, M. and Maglaras, C. (2003), On customer contact centers with a call-back option: customer decisions, routing rules and system design, *Oper. Res.*, to appear.

[4] Armony, M. and Maglaras, C. (2003), Contact centers with a call-back option and real-time delay information, *Oper. Res.*, to appear.

[5] Atar R. (2003), Treelike parallel server stations in heavy traffic, preprint.

[6] Atar R. (2003), Asymptotically optimal policies for treelike parallel server stations in heavy traffic, preprint.

[7] R. Atar, A. Mandelbaum, and M. Reiman (2002), Scheduling a multi-class queue with many i.i.d. servers: asymptotic optimality in heavy traffic, preprint, available at `http://iew3.technion.ac.il/serveng/References/references.html`.

[8] N. Bambos and J. Walrand (1993), Scheduling and stability aspects of a general class of parallel processing systems, *Advances in Applied Probability*, **25**, pp. 176–202.

[9] S.L. Bell and R.J. Williams (2001), 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**, pp. 608–649.

[10] Borst, S., Mandelbaum, A. and Reiman, M. (2003), Dimensioning large call centers, *Oper. Res.*, to appear.

[11] S.C. Borst and P. Seri (2000), Robust algorithms for sharing agents with multiple skills, Working Paper.

[12] Bramson, M. (1997), State space collapse with applications to heavy-traffic limits for multiclass queueing networks, *Queueing Systems*, **30**,pp 89–148.

[13] A. Brandt and M. Brandt (1999), On a two-queue priority system with impatience and its application to a call center, *Methodology and Computing in Applied Probablity*, **1**, pp. 191-210.

[14] Browne, S. and Whitt, W. (1995), Piecewise-linear diffusion processes, *Advances in Queueing. Theory, Methods, and Open Problems*, Dshalalow, J.H. (editor), CRC Press, Chapter 18, pp 463–480.

[15] H. Chen and D. Yao (2001), *Fundamentals of queueing networks : performance, asymptotics, and optimization*, Springer, New-York.

[16] Dai, J. G. (1999), Stability of fluid and stochastic processing networks. *MaPhySto*, 9.

[17] de Véricourt, F. and Zhou, Y.-P. (2003), Managing response time and service quality in a call allocation problem, preprint.

[18] A.K. Erlang (1948), On the rational determination of the number of circuits, *The life and works of A.K. Erlang.* E. Brockmeyer, H.L. Halstrom, A. Jensen, eds. Copenhagen: the Copenhafen Telephone Company.

[19] Ethier, S.N. and Kurtz, T.G. (1985), *Markov Processes, Characterization and Convergence*, John Wiley & Sons.

[20] A. Federgruen and H. Groenevelt (1988), M/G/c systems with multiple customer classes: characterization and achievable performance unrder nonpreemptive priority rules, *Management Science*, **34**, pp. 1121-1138.

[21] P. Fleming, A. Stolyar and B. Simon (1994), Heavy traffic limit for a mobile phone system loss model, *Proceedings of 2nd Int'l Conf. on Telecomm. Syst. Mod. and Analysis*, Nashville, TN.

[22] N. Gans, G. Koole and A. Mandelbaum (2003), Telephone call centers: tutorial, review and research prospects, *Manufacturing & Service Operations Management*, **5**:2, pp. 79–141.

[23] N. Gans and G. van Ryzin (1997), Optimal control of a multiclass, flexible queue system, *Operations Research*, **45**, pp. 677–693.

[24] N. Gans and Y.-P. Zhou (1999), Managing learning and turnover in employee staffing, to appear in *Operations Research.*

[25] N. Gans and Y.-P. Zhou (2002), A call-routing problem with service-level constraints, to appear in *Operations Research.*

[26] Garnett, O., Mandelbaum, A. and Reiman, M. (2002), Designing a call center with impatient customers, *Manufacturing & Service Operations Management*, **4**:3, pp. 208–227.

[27] Glazebrook, K. and Niño-Mora, J. (2001), Parallel scheduling of multiclass M/M/m queues: approximate and heavy-traffic optimization of achievable performance, *Operations Research*, **49**:4, pp. 609-623.

[28] Glynn, P.W. (1990), Diffusion Approximations, *Stochastic Models, Handbooks in OR & MS*, D. Heyman and M. Sobel (editors), North-Holland, **2**, pp. 145-198.

[29] Gurvich, I. (2004), Design and Control of the M/M/N Queue with Multi-Type Customers and Many Servers, *Masters Thesis*, Tehcnion Institute of Technology, Israel.

[30] Halfin, S. and Whitt, W. (1981), Heavy-traffic limits for queues with many exponential servers, *Oper. Res.*, **29**(3), pp. 567–588.

[31] Harrison, J.M. (1998), Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies, *Annals of applied probability*, **8**, pp. 822-848.

[32] J.M. Harrison and A. Zeevi (2003), Dynamic scheduling of a multiclass queue in the Halfin and Whitt heavy traffic regime, to appear in *Operations Research*.

[33] J.M. Harrison and A. Zeevi (2003), A method for staffing large call centers using stochastic fluid models, preprint.

[34] D.L. Jagerman (1974), Some properties of the Erlang loss function, *Bell Systems Technical Journal*, **53**:3, pp. 525–551.

[35] P. Jelenkovic, A. Mandelbaum, and P. Momcilovicb (2002), The GI/D/N queue in the QED regime, Preprint, available at
`http://iew3.technion.ac.il/serveng/References/references.html`.

[36] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt (1996), Server staffing to meet time-varying demand, *Management Science*, **42**, pp. 1383–1394.

[37] Kella, O. and U. Yechiali (1985), Waiting times in the nonpreemptive priority M/M/c queue, *Stochastic Models* **1**:2, pp. 257-262.

[38] Lipster, R.Sh. and Shiryaev, A.N. (1989), *Theory of Martingales*, Kluwer, Amsterdam.

[39] Luh, H.P. and Viniotis, I. (2002), Threshold Control Policies for Heterogeneous Server Systems, *Math Meth Oper Res*, **55**, pp 121-142.

[40] C. Maglaras and A. Zeevi (2003), Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions, To appear in *Management Science*.

[41] A. Mandelbaum and A.L. Stolyar (2003), Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule, preprint, available at
`http://iew3.technion.ac.il/serveng/References/references.html`.

[42] W.A. Massey and R.B. Wallace (2002), An optimal design of the M/M/C/K queue for call centers, Working Paper, Princeton University.

[43] M. Perry and A. Nilsson (1992), Performance modeling of automatic call distributors: assignable grade of service staffing, In *XIV International Switching Symposium*, pp. 294–298.

[44] Puhalskii, A. (1994), On the invariance principle for the first passage time, *Mathematics of Operations Research*, **19**:4, pp. 946-954.

[45] Puhalskii, A.A and Reiman, M.I. (2000), The Multiclass GI/PH/N Queue in the Halfin-Whitt Regime, *Advances in Applied Probability*, **32**, pp. 564-595.

[46] Reiman, M.I. (1984), Some Diffusion Approximations with State Space Collapse, *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (editors), Springer-Verlag, pp. 209-240.

[47] Rykov, V.V. (2001), Monotone Control of Queueing Systems with Heterogeneous Servers, *Queueing Systems*, **37**, pp 391-403.

[48] R.A. Shumsky (2000), Approximation and analysis of a queueing system with flexible and specialized servers, Working Paper, University of Rochester.

[49] D.A. Stanford and W.K. Grassmann (2000), Bilingual server call centres, In *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.), Fields Institute Communications **28**, pp. 31–48.

[50] S. Stolyar (2004), Optimal routing in output-queues flexible server systems, preprint.

[51] Sze, D.Y. (1984), A queueing model for telephone operator staffing, *Operations Research* **32**, pp. 229–249.

[52] Y.-Ch. Teh and A.R. Ward (2002), Critical thresholds for dynamic routing in queueing networks, *Queueing Systems*, **42**, pp. 297–316.

[53] W. Whitt (1984), Heavy traffic approximations for service systems with blocking, *AT&T Bell Lab. Tech. Journal* **63**, pp. 689–708.

[54] W. Whitt (1992), Understanding the efficiency of multi-server service systems, *Management Science*, **38**, pp. 708–723.

[55] W. Whitt (2002), A diffusion approximation for the G/GI/n/m queue, preprint, available at `http://www.research.att.com/~wow/`.

[56] W. Whitt (2002), Heavy-traffic limits for the $G/H_2^*/n/m$ queue, prepring available at `http://www.research.att.com/~wow/`.

[57] Yahalom T. and Mandelbaum, A. (2004), Optimal Scheduling of a Multi-Server Multi-Class Non-Preemptive Queueing System, Preprint.