Technical Appendix for:

Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers

Mor Armony[1] Avishai Mandelbaum[2]

First submitted June 2, 2008; Revised June 12, 2009.

A System Dynamics and Fluid Model

In this section we describe the dynamics of the system in terms of its flow balance equations. We also describe the fluid model associated with the QED regime, and state and prove a fluid-related proposition which is needed for the asymptotic analysis in this regime.

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 customers joining server pool k, up to time t, after being delayed in the queue. The number of arrivals into the queue (excluding direct arrivals to one of the servers) up to time t is denoted by $A_q(t;\pi)$. In addition, let $T_k(t;\pi)$ denote the total time spent serving customers by all N_k servers of pool k up to time t. In particular, $0 \leq T_k(t;\pi) \leq N_k t$. Respectively, let $I_k(t;\pi)$ be the total idle time experienced by servers of pool k up to time t. Also, let $D_k(t)$, $t \geq 0$, be a Poisson process with rate μ_k. Then the number of service completions out of server pool k may be written as $D_k(T_k(t;\pi))$. In addition, let $E(t;\pi)$ represent the total time spent by customers in the queue up to time t, and let $L(t)$, $t \geq 0$, be a Poisson process with rate θ. Then the number of customers who have abandoned up to time t can be written as $L(E(t;\pi))$. The above definitions allow us to write the following *flow balance equations*:

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

$$E(t;\pi) = \int_0^t Q(s;\pi)ds, \tag{A.2}$$

[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.

$$Z_k(t;\pi) = Z_k(0;\pi) + A_k(t;\pi) + B_k(t;\pi) - D_k(T_k(t;\pi)), \;\; k = 1, ..., K, \tag{A.3}$$

$$T_k(t;\pi) = \int_0^t Z_k(s;\pi)ds \tag{A.4}$$

$$Y(t;\pi) = Y(0;\pi) + A(t) - \sum_{k=1}^{K} D_k(T_k(t;\pi)) - L(E(t;\pi)), \tag{A.5}$$

$$A(t) = A_q(t;\pi) + \sum_{k=1}^{K} A_k(t;\pi), \tag{A.6}$$

$$T_k(t;\pi) + I_k(t;\pi) = N_k t. \tag{A.7}$$

Finally, for any work conserving policy π we have the additional three equations:

$$Q(t;\pi) \cdot \left(\sum_{k=1}^{K} (N_k - Z_k(t;\pi)) \right) = 0, \;\; \forall t \geq 0, \tag{A.8}$$

$$\int_0^\infty \sum_{k=1}^{K} (N_k - Z_k(t;\pi)) dA_q(t;\pi) = 0, \tag{A.9}$$

and

$$\sum_{k=1}^{K} \int_0^\infty Q(t;\pi) dI_k(t;\pi) = 0. \tag{A.10}$$

All of the above apply almost surely. In words, (A.8) means that there are customers in queue only when *all* servers are busy. The verbal interpretation of (A.9) is that new arrivals wait in the queue only when all servers are busy. Finally, (A.10) states that servers can only be idle when the queue is empty.

Fluid Scaling: For each $\lambda > 0$, $k = 1, ..., K$, and a fixed sequence of routing policies $\pi^\lambda \in \Pi(\lambda, N^\lambda)$ 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 $\bar{Y}^\lambda(t) = \frac{Y^\lambda(t)}{N^\lambda}$, $\bar{A}^\lambda(t) = \frac{A^\lambda(t)}{N^\lambda}$, $\bar{A}_k^\lambda(t) = \frac{A_k^\lambda(t)}{N^\lambda}$, $\bar{A}_q^\lambda(t) = \frac{A_q^\lambda(t)}{N^\lambda}$, $\bar{B}_k^\lambda(t) = \frac{B_k^\lambda(t)}{N^\lambda}$, $\bar{T}_k^\lambda(t) = \frac{T_k^\lambda(t)}{N^\lambda}$, $\bar{I}_k^\lambda(t) = \frac{I_k^\lambda(t)}{N^\lambda}$, and $\bar{E}^\lambda(t) = \frac{E^\lambda(t)}{N^\lambda}$. Finally, let $\bar{D}_k^\lambda(t) = D_k^\lambda(t) = D_k(t)$ and $\bar{L}^\lambda(t) = L^\lambda(t) = L(t)$. That is, as equalities between processes,

$$(\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda, \bar{E}^\lambda) = (Q^\lambda, Z_k^\lambda, Y^\lambda, A^\lambda, A_k^\lambda, A_q^\lambda, B_k^\lambda, T_k^\lambda, I_k^\lambda, E^\lambda)/N^\lambda,$$

and $(\bar{D}_k^\lambda, \bar{L}^\lambda) = (D_k, L)$. Note that D_k^λ and L^λ need not be divided by N^λ, due to their definitions as Poisson processes with rates μ_k and θ, respectively, which are independent of λ.

Using standard tools of fluid models (see for example [3], Theorem A.1) one can show that if $(\bar{Q}^\lambda(0), \bar{Z}_k^\lambda(0), \; k = 1, ..., K)$ are bounded, then the process $\bar{\mathbb{X}} := (\bar{Q}^\lambda, \bar{Z}_k^\lambda, \bar{Y}^\lambda, \bar{A}^\lambda, \bar{A}_k^\lambda, \bar{A}_q^\lambda, \bar{B}_k^\lambda, \bar{T}_k^\lambda, \bar{I}_k^\lambda, \bar{E}^\lambda, \bar{D}_k^\lambda, \bar{L}^\lambda)$ is pre-compact, as $\lambda \to \infty$, and hence any sequence has a converging subsequence (where the convergence is almost surely, uniformly on compact intervals). 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 $\bar{Q}^{\lambda}(t)$, as $\lambda \to \infty$). Note that, by Theorem A.1 of [3], equations (A.1)-(A.7) imply that the following flow balance equations hold for *any* fluid limit:

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}_q(t) - \sum_{k=1}^{K} \bar{B}_k(t) - \theta \bar{E}(t), \tag{A.11}$$

$$\bar{E}(t) = \int_0^t \bar{Q}(s)ds, \tag{A.12}$$

$$\bar{Z}_k(t) = \bar{Z}_k(0) + \bar{A}_k(t) + \bar{B}_k(t) - \mu_k \bar{T}_k(t), \;\; k = 1, ..., K, \tag{A.13}$$

$$\bar{T}_k(t) = \int_0^t \bar{Z}_k(s)ds \tag{A.14}$$

$$\bar{Y}(t) = \bar{Y}(0) + \mu t - \sum_{k=1}^{K} \mu_k \bar{T}_k(t) - \theta \bar{E}(t), \tag{A.15}$$

$$\mu t = \bar{A}_q(t) + \sum_{k=1}^{K} \bar{A}_k(t), \tag{A.16}$$

$$\bar{T}_k(t) + \bar{I}_k(t) = q_k t. \tag{A.17}$$

Finally, for work conserving policies, conditions (A.8)-(A.10) imply:

$$\bar{Q}(t) \cdot \left(\sum_{k=1}^{K} (q_k - \bar{Z}_k(t)) \right) = 0, \tag{A.18}$$

$$\int_0^{\infty} \sum_{k=1}^{K} (q_k - \bar{Z}_k(t)) d\bar{A}_q(t) = 0, \tag{A.19}$$

and

$$\sum_{k=1}^{K} \int_0^{\infty} \bar{Q}(t) d\bar{I}_k(t) = 0. \tag{A.20}$$

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 A.1 ***(fluid limits)*** *For $\lambda > 0$, let $\pi^{\lambda} \in \Pi(\lambda, N^{\lambda})$ be a sequence of work-conserving policies (omitted from the following notation), and let $\bar{\mathbb{X}}$ be some 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, $\bar{Q}(t) \equiv 0$ and $\bar{Z}_k(t) \equiv q_k,\ k = 1, ..., K$, for all $t \geq 0$.*

Proof of Proposition A.1: 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 Lemma C.1 of [6], 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 (A.18) $\bar{Z}_k(t) = q_k$, for all k. In particular, if f is differentiable at t, then

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

If t is such that $\bar{Y}(t) < 1$, then $\bar{Z}_k(t) < q_k$ for at least one k, and hence, by (A.18), $\bar{Q}(t) = 0$. If f is differentiable at t then,

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

∎

B Proofs

Proof of Proposition 3.1: Due to (A.4) below which relates between the abandonment probability and the expected queue length, minimizing $P_\pi(ab)$ is equivalent to minimizing $EQ(\infty;\pi)$. We show that $EQ(\infty;\pi)$ under any policy π which is not necessarily FCFS is equal to $EQ(\infty;\pi')$, where π' is a corresponding FCFS policy. We prove this using a construction of the policy π' and a sample path coupling. Consider the system under a particular sample path ω and the policy π. Construct a policy π' with a sample path ω' as follows: The arrival times under both ω and ω' are the same. Every time the policy π serves a tagged customer which is *not* at the head of the line, the policy π' leaves this customer in line, and instead serves the head-of-line (HOL) customer. The service time of this HOL customer under ω' is set equal to the service time of the tagged customer under ω. Similarly, the time to abandon from that moment on of the tagged customer under ω' is set equal to the time to abandon of the HOL customer under ω. Since the time to abandon distribution is exponential one can couple those two systems and get the same steady-state expected queue length. Also, by construction, π' is a FCFS policy. ∎

Proof of Proposition 3.2: We prove the Proposition using sample-path coupling arguments. Consider two coupled systems both with the same initial conditions, and the same sequence of arrivals. System 1 operates under an arbitrary policy $\pi \in \Pi_p$ while System 2 operates under FSF$_p$. For all $t \geq 0$ and $i = 1, 2$, let $Q^i(t)$, $Y^i(t)$, and $Ab^i(t)$ be the queue length at time t, the head-count at this time, and the total number of

abandonment up to this time in System i, respectively. We claim that the two systems can be coupled such that the following three properties hold almost surely for all $t \geq 0$:

$$Ab^1(t) \geq Ab^2(t), \tag{A.1}$$

$$Q^2(t) - Q^1(t) \leq Ab^1(t) - Ab^2(t), \tag{A.2}$$

and

$$Y^2(t) - Y^1(t) \leq Ab^1(t) - Ab^2(t). \tag{A.3}$$

Establishing property (A.1) will complete the proof of the proposition. Let $t_0 = 0$. We define the set of path-dependent time points $0 < t_1 < t_2 < ...$ and corresponding state transitions, inductively. For $n \geq 1$, suppose that $0 < t_1 < t_2 < ... < t_n$ have been determined. Let t_{n+1} be the time of the first transition in either system, after time t_n, and let $i = 1$ if the transition is in system 1 and $i = 2$, otherwise.

- If $Ab^1(t_n) = Ab^2(t_n)$ and the transition at time t_{n+1} corresponds to an abandonment in system 2, then we impose an abandonment in system 1 at the same time.
- Otherwise, if $Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ or $Y^2(t_n) - Y^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ and the transition at time t_{n+1} corresponds to a service completion in system 1, then we impose a service completion in system 2 at the same time.
- Otherwise, the relevant transitions occur as follows: arrivals occur into both systems simultaneously, while departures and abandonment occur in system i only.

We prove (A.1)-(A.3) by induction on $t_n,\ n = 0, 1, 2, ...$ At time $t_0 = 0$ both systems are assumed to have the same state and therefore properties (A.1)-(A.3) are trivially satisfied. Suppose that these properties are satisfied for all $t \leq t_n$. We need to establish that they are also satisfied at $t_n < t \leq t_{n+1}$. Clearly, it suffices to prove that they are satisfied at $t = t_{n+1}$. We verify the three properties as follows:

- **Verification of (A.1):** This property might be violated only if $Ab^1(t_n) = Ab^2(t_n)$ and at time t_{n+1}, there is an abandonment from system 2 and not in system 1. But, by the construction of out coupling, any such transition in system 2 will be accompanied by a transition in system 1. This coupling is valid only if $Q^2(t_n) \leq Q^1(t_n)$, which holds due to (A.2) and the equality in (A.1).
- **Verification of (A.2):** This property might be violated if $Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$, and one or more of the following occurs: a) $Q^1(t_n) > 0$ and $Q^2(t_n) = 0$, b) $Q_1(t_n) > 0$, and there

is a service completion in system 1 and not in system 2, or c) there is an arrival into both systems that enters service in system 1 and joins the queue in system 2. Case a) cannot occur, because, by (A.1), $Q^2(t_n) \geq Q^1(t_n) > 0$. Case b) may be contradicted by our construction of the coupling. This coupling is valid only if $\sum_{k=1}^K \mu_k Z_k^1(t_n) \leq \sum_{k=1}^K \mu_k Z_k^2(t_n)$. But, since $Q^2(t_n) \geq Q^1(t_n) > 0$, we have that, by the work-conservation properties of FSF$_p$, all servers are busy in system 2 at time t_n, which implies that $\sum_{k=1}^K \mu_k Z_k^2(t_n) = \sum_{k=1}^K \mu_k N_k$. Finally, c) implies that, at time t_n, all the servers are busy in system 2 and some servers are idle in system 1. Th! erefore, $Y^2(t_n) - Y^1(t_n) > Q^2(t_n) + N - (Q^1(t_n) + N) = Q^2(t_n) - Q^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ which violates (A.3).

- **Verification of (A.3):** The latter might be violated if $Y^2(t_n) - Y^1(t_n) = Ab^1(t_n) - Ab^2(t_n)$ and a service completion occurs in system 1 only. But, this cannot occur due to our coupling construction. This coupling is valid only if $\sum_{k=1}^K \mu_k Z_k^1(t_n) \leq \sum_{k=1}^K \mu_k Z_k^2(t_n)$. But, due to (A.1), $Y^2(t_n) \geq Y^1(t_n)$. In particular, due to the work-conserving nature of FSF$_p$, there are more busy servers in system 2 than in system 1. Now, due to the fast server first property of FSF$_p$ this also implies that $\sum_{k=1}^K Z_k^2(t_n)\mu_k \geq \sum_{k=1}^K Z_k^1(t_n)\mu_k$.

∎

Proof of Corollary 3.1: Notice that, in steady-state, the following balance equation holds for any policy $\pi \in \Pi_p$ (see also equation (2) in [4]):

$$\theta \cdot E[Q(\infty;\pi)] = \lambda \cdot P_\pi(ab). \tag{A.4}$$

The left-hand-side corresponds to the rate of abandonment from the system, and the right-hand-side describes the rate of arrival of customers who will eventually abandon. From Little's law and (A.4) we also obtain a relationship between the expected waiting time and probability of abandonment in steady-state:

$$\theta \cdot E[W(\infty;\pi)] = P_\pi(ab). \tag{A.5}$$

Proposition 3.2 together with the relationships (A.4) and (A.5) completes the proof. ∎

Proof of Proposition 3.3: The proof is shown for $K = 2$. The general case follows similarly. We first show that $Q_A \leq Q_B$ and that $Z_A - (N_1 + N_2) \leq Z_B - N_B$, almost surely. Consider two coupled systems, A and B, both with the same initial conditions (all servers are busy and no customers in queue) and the same sequence of arrivals. We will show that the two systems can be coupled such that the following two properties hold, almost surely, for all $t \geq 0$:

1. $Q_A(t) \leq Q_B(t)$, and

2. $Z_A(t) - (N_1 + N_2) \leq Z_B(t) - N_B$,

Let $t_0 = 0$. We define the set of path-dependent time points $0 < t_1 < t_2 < ...$ and corresponding state transitions, inductively. For $n \geq 1$, suppose that $0 < t_1 < t_2 < ... < t_n$ have been determined. Let t_{n+1} be the time of the first transition in either system, after time t_n, and let $i = A$ if the transition is in system A and $i = B$, otherwise.

- If $Q_A(t_n) = Q_B(t_n) > 0$ and the transition at time t_{n+1} corresponds to a service completion or an abandonment in system B, then we impose a service completion or an abandonment in system A, at the same time, respectively.

- Otherwise, if $Z_A(t_n) - (N_1 + N_2) = Z_B(t_n) - N_B$, $Q_B(t_n) = 0$, and the transition at time t_{n+1} corresponds to a service completion in system B, then we impose a service completion in system A, at the same time.

- Otherwise, the relevant transitions occur as follows: arrivals occur into both systems simultaneously, while departures and abandonment occur in system i only.

We prove 1. and 2. by induction on $t_n,\ n = 0, 1, 2, ...$ At time $t_0 = 0$ both systems are assumed to have all servers busy and no queue. Therefore properties 1. and 2. are trivially satisfied. Suppose that these properties are satisfied for all $t \leq t_n$. We need to establish that they are also satisfied at $t_n < t \leq t_{n+1}$. Clearly, it suffices to prove that they are satisfied at $t = t_{n+1}$. We verify the two properties as follows:

- **Verification of 1.:** This property might be violated only if $Q_A(t_n) = Q_B(t_n) > 0$ and at time t_{n+1}, there is a service completion or an abandonment in system B and not in system A. But, by the construction of out coupling, any such transition in system B will be accompanied by a similar transition in system A. The coupling with respect to service completions is valid because, due to work-conservation, the total service rate in system B at time t_n is $N_B\mu_2$ which is less than or equal to $N_1\mu_1 + N_2\mu_2$, the total service rate in system A. The coupling with respect to abandonment is valid because both queue lengths are equal at time t_n.

- **Verification of 2.:** This property might be violated if $Z_A(t_n) - (N_1 + N_2) = Z_B(t_n) - N_B$, $Q_B(t_n) = 0$, and there is a service completion in system B, but not in system A. This cannot occur due to the construction of our coupling. This coupling is valid only if the total service rate in system B is less

than or equal to the total service rate in system A. The latter is true because if $Z_B = N_B - m$ where $m \leq N_B$, then $Z_A = N_1 + N_2 - m$. Also, the total service rate in system A is minimal when the idle servers are the faster ones. In other words, the total service rate in system A is greater than or equal to

$$x := \begin{cases} N_1\mu_1 + (N_2 - m)\mu_2, & \text{if } m \leq N_2, \\ (N_1 - (m - N_2))\mu_1, & \text{if } m > N_2. \end{cases}$$

In either case, $x \geq (N_B - m)\mu_2$.

The comparison between systems A and C is analogous. The details are omitted. ∎

Proof of Proposition 4.1: The proof follows directly from Theorem 3.1 of [5] (our model satisfies assumptions C-1 (pool dependent service rate) and C-3 (the graph that connects server pools to customer classes is a tree) of that theorem). ∎

Proof of Remark 4.1: The proof follows directly from Theorem 3.1 of [5] (our model satisfies assumptions C-1 and C-3 of that theorem). ∎

Proof of Proposition 4.2: The proof follows from [2]. Note that the process $X(\cdot)$, restricted to $[0, \infty)$, is a an O-U process with infinitesimal drift $-\delta\sqrt{\mu} - \theta x$ and variance 2μ. Hence, according to [2, (18.33)], its steady-state density, conditional on $X(\infty) \geq 0$, is normal with mean $-\delta/\sqrt{\mu}/\theta$ and variance μ/θ, conditioned on having non-negative values only (see [2, (18.28)]). 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. 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 [2, p. 471]). Hence, α may be solved for by a smooth fit, namely, by equating the limits of $f(\cdot)$ at 0 from both left and right. ∎

Proof of Proposition 4.3: By Corollary 4.2 of [5] 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 $\lambda > 0$. First note that under FSF$_p$ the total number in the system Y^λ is a Birth and Death process with birth rates $\lambda(y) = \lambda$ and death rates $\mu^\lambda(y)$ as given in (3.1). Due to abandonment, 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. Finally, since $\vec{X}^\lambda$ is easily obtained as a one-to-one function of its sum X^λ, the existence of a steady-state distribution for $\vec{X}^\lambda$ has been established.

   To show the existence of a stationary distribution of $\vec{X}^\lambda$ under the non-preemptive policy FSF one can use the stationarity of the process with respect to FSF$_p$ and the dominance of FSF$_p$ over FSF which was established in Proposition 6.1, noting that FSF is work-conserving. The details are omitted as the proof is identical to the proof of part 1. of Proposition 4.6 in [1].

2. Tightness of $\vec{X}^\lambda(\infty)$, $0 < \lambda < \infty$, is established in two stages. First, we show that $\vec{X}^\lambda(\infty)$ is tight under FSF$_p$. We then conclude that this sequence is also tight under FSF.

   Tightness under FSF$_p$: Suppose that the policy FSF$_p$ is used (to be omitted from the notation for brevity). We start by establishing the tightness of $X^\lambda(\infty) = \sum_{k=1}^K X_k^\lambda(\infty)$. Assume, without loss of generality, that $K = 2$. Along the lines of Proposition 3.3 define two related sequences of systems. One is sequence B which is a sequence of $M/M/N_B^\lambda + M$ systems with N_B^λ servers all working with rate μ_2, where $N_B^\lambda = \left\lfloor \frac{N_1^\lambda \mu_1 + N_2^\lambda \mu_2}{\mu_2} \right\rfloor$. Similarly, define the sequence C to be a sequence of $M/M/N_C^\lambda + M$ systems with N_C^λ servers all working with rate μ_1, where $N_C^\lambda = \left\lceil \frac{N_1^\lambda \mu_1 + N_2^\lambda \mu_2}{\mu_1} \right\rceil$. The sequences B and C both have the same sequence of arrival rates λ as the original system, and the same abandonment rate of θ. Then, according to Proposition 3.3, for every fixed λ, $X^\lambda(\infty)$ is stochastically dominated from above by $Q_B^\lambda(\infty)/\sqrt{N^\lambda}$ and is stochastically dominated from below by $\frac{Z_C^\lambda(\infty) - N_C^\lambda}{\sqrt{N^\lambda}}$. Tightness of $X^\lambda(\infty)$ now follows from [4, Theorem 2], and the facts that $N_B^\lambda = \lambda/\mu_2 + \frac{\delta}{\sqrt{\mu_2}}\sqrt{\lambda/\mu_2} + o(\sqrt{\lambda})$, $N_C^\lambda = \lambda/\mu_1 + \frac{\delta}{\sqrt{\mu_1}}\sqrt{\lambda/\mu_1} + o(\sqrt{\lambda})$, and that both $\sqrt{N_B^\lambda/N^\lambda}$ and $\sqrt{N_C^\lambda/N^\lambda}$ have finite limits.

   Now that we have established that $X^\lambda(\infty)$ is tight, we proceed by showing that $\vec{X}^\lambda(\infty)$ is tight (again for $K = 2$, without loss of generality). Note that under FSF$_p$, $Q^\lambda + Z_1^\lambda = \left[Y^\lambda - N_2^\lambda\right]^+$ and $Z_2^\lambda = \min\{Y^\lambda, N_2^\lambda\}$. Therefore, as long as $Y^\lambda(\infty) \geq N_2^\lambda$, $X_1^\lambda(\infty) = X^\lambda(\infty)$ and $X_2^\lambda(\infty) = 0$. But $Y^\lambda(\infty) \geq N_2^\lambda$ is equivalent to $X^\lambda(\infty) \geq -\frac{N_1^\lambda}{\sqrt{N^\lambda}}$, whose probability goes to 1 as $\lambda \to \infty$ by tightness of $X^\lambda(\infty)$. Therefore, the vector $\vec{X}^\lambda(\infty)$ is tight.

   Tightness under FSF: To establish the tightness of $\vec{X}^\lambda(\infty)$ under FSF, we can use a proof which is essentially identical the proof of part 2. in [1, Proposition 4.6]. All that is missing is to establish that

the steady-state probability that all the servers are busy under FSF$_p$ goes to a non-zero limit as $\lambda\to\infty$. Since we have already established that under FSF$_p$, $X^\lambda(\infty)$ weakly converges to $X(\infty)$, it is left to show that the probability that $X(\infty)$ is non-negative is non-zero. But this probability is equal to α (in the statement of Proposition 4.2) which is clearly positive. ∎

Proof of Corollary 4.2: The proof of the corollary follows from Corollary 4.2 in [5] and the proof of Proposition 4.3. ∎

We are now finally in a position to prove the asymptotic optimality of FSF as stated in Theorem 4.1.

Proof of Theorem 4.1: Let $\{\pi^\lambda\}_{\lambda>0} \subseteq \Pi$ be a sequence of policies, and suppose that the steady-state distributions of $Q^\lambda(\cdot;\pi^\lambda)$, $V^\lambda(\cdot;\pi^\lambda)$ and $P^\lambda_{\pi^\lambda}(ab,\cdot)$ exist for all $\lambda > 0$ (here $P^\lambda_{\pi^\lambda}(ab,t)$ is defined as the probability of abandonment for a virtual customer who arrives at time t.) In addition, for $\lambda > 0$, define $\hat{Q}^\lambda(\infty;\pi^\lambda) := Q^\lambda(\infty;\pi^\lambda)/\sqrt{N^\lambda}$, $\hat{W}^\lambda(\infty;\pi^\lambda) := \sqrt{N^\lambda}W^\lambda(\infty;\pi^\lambda)$, and $\hat{P}^\lambda_{\{\pi^\lambda\}}(ab) := \sqrt{N^\lambda}P^\lambda_{\pi^\lambda}(ab)$.

We prove the theorem in three steps:

1. First we show asymptotic optimality of FSF$_p$ in terms of minimizing $\limsup_{\lambda\to\infty} E\hat{Q}^\lambda(\infty)$, as $\lambda\to\infty$.
2. The asymptotic optimality of FSF in terms of minimizing $\limsup_{\lambda\to\infty} E\hat{Q}^\lambda(\infty)$ as $\lambda\to\infty$ is shown next.
3. We conclude by showing the asymptotic optimality of FSF with respect to both $E\hat{W}^\lambda(\infty)$, $\hat{P}^\lambda(ab)$, and $\sqrt{\lambda}P^\lambda(ab)$ as $\lambda\to\infty$.

Step 1. In Corollary 3.1 we have shown that FSF$_p$ minimizes $E[Q^\lambda(\infty)]$ for every fixed λ. Therefore, we can conclude that

$$\limsup_{\lambda\to\infty} E\hat{Q}^\lambda(\infty;\text{FSF}_p) \le \liminf_{\lambda\to\infty} E\hat{Q}^\lambda(\infty;\pi^\lambda) \tag{A.6}$$

Step 2. In light of 1. it is sufficient to show that $\lim_{\lambda\to\infty} E\hat{Q}^\lambda(\infty;\text{FSF}) = \lim_{\lambda\to\infty} E\hat{Q}^\lambda(\infty;\text{FSF}_p)$, in order to establish the asymptotic optimality of FSF with respect to $E\hat{Q}^\lambda(\infty)$ as $\lambda\to\infty$. From Proposition 4.3 and the continuous mapping theorem it follows that $\hat{Q}^\lambda(\infty)$ converges weakly to $[X(\infty)]^+$ under both FSF and FSF$_p$. In turn, corollary 4.2, shows that $\lim_{\lambda\to\infty} E\hat{Q}^\lambda(\infty) = E[X(\infty)]^+$ under both these policies.

Step 3. The asymptotic optimality of FSF with respect to $E\hat{W}^{\lambda}(\infty)$ and $\hat{P}^{\lambda}(ab)$ follows from Little's law and the relationship (A.4). Finally, the asymptotic optimality of FSF with respect to $\sqrt{\lambda}P^{\lambda}(ab)$ follows from (4.6).

∎

Proof of Corollary 4.1: The proof of this corollary is included in the proof of Theorem 4.1. ∎

Proof of Proposition 5.1: For $0 < \Delta < 1$ the proof follows directly from the discussion of Section 4.1.

For $\Delta = 1$, suppose that $\sum_{k=1}^{K} \mu_k N_k^{\lambda} = o(\lambda)$, and that, by contradiction, there exists a sequence of policies $\{\pi^{\lambda}\}$ such that $\limsup_{\lambda\to\infty} P_{\pi^{\lambda}}^{\lambda}(ab) = 1 - \epsilon < 1$. In particular, this implies that with a staffing vector $\vec{N}'$ that satisfies $\sum_{k=1}^{K} \mu_k N_k'^{\lambda} = \lambda\epsilon/2 + o(\lambda)$ one can obtain $\limsup_{\lambda\to\infty} P_{\pi^{\lambda}}^{\lambda}(ab) = 1 - \epsilon$ (by only using the servers in the original staffing vector). This is a contradiction to the result of this proposition with respect to $\Delta = 1 - \epsilon/2$.

Finally, for $\Delta = 0$ we wish to establish that if $\sum_{k=1}^{K} \mu_k N_k^{\lambda} \geq \lambda + o(\lambda)$ then there exists a sequence of policies $\{\pi^{\lambda}\}$ such that $\lim_{\lambda\to\infty} P_{\pi^{\lambda}}^{\lambda}(ab) = 0$. This can be done by using the first part of this proposition (for $0 < \Delta < 1$) to establish that there exists a sequence $\{\Delta^{\lambda}\}$ with $\lim_{\lambda\to\infty} \Delta^{\lambda} = 0$ and a sequence of policies $\{\pi^{\lambda}\}$ such that $P_{\pi^{\lambda}}^{\lambda}(ab) \leq \Delta^{\lambda}$ for all λ large enough. ∎

Lemma B.1 *The function $\Delta_{\mu_1,\theta}(\cdot) := \Delta(\delta, \mu_1, \theta)$ defined in (5.8) is continuous and monotonically decreasing in δ. Moreover, $\lim_{\delta\to\infty} \Delta(\delta, \mu_1, \theta) = 0$ and $\lim_{\delta\to-\infty} \Delta(\delta, \mu_1, \theta) = \infty$.*

The proof of Lemma B.1 follows, in a straightforward manner, from the proof of Theorem 4.1 in [32].

Proof of Proposition 5.2: We prove the proposition for $K = 2$. The general case follows similarly. Fix $-\infty < \delta < \infty$, and suppose that (5.7) holds. 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 (5.6) with $\Delta{:=}\Delta(\delta, \mu_1, \theta)$ 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. Corollary 4.2 then confirms that (5.6) 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 (5.6) by using FSF with respect to a subset of each server pool of size $\tilde{N}_k^{\lambda} = \frac{\mu_k N_k^{\lambda} - (d^{\lambda}/2)\sqrt{\lambda}}{\mu_k}$, where $d^{\lambda} := \frac{\mu_1 N_1^{\lambda}+\mu_2 N_2^{\lambda}-\lambda}{\sqrt{\lambda}} - \delta$. Finally, if $\tilde{\delta} = \delta$, then (5.6) holds if FSF is used by Corollary 4.2.

Now suppose (5.6) holds for some $0 < \Delta_0 < \infty$, and let $-\infty < \delta_0 < \infty$ be such that $\Delta(\delta_0, \mu_1, \theta) = \Delta_0$ (such δ exists due to Lemma B.1). Assume by contradiction that (5.7) is violated with respect to $\delta = \delta_0$. Then if (5.7) holds with respect $\delta = \delta_1$ for some $-\infty < \delta_1 < \delta_0$, then by the monotonicity of $\Delta(\delta)$ and Corollary 4.2, FSF will satisfy $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}(ab) = \Delta_1$ where $\Delta_1 > \Delta_0$, which contradicts the asymptotic optimality of FSF (Theorem 4.1). Finally, if (5.7) is violated with respect to any $\delta > -\infty$ then the case $\delta = -\infty$ applies. This case is dealt with next.

To complete the proof, we need to examine the cases where $\delta = -\infty$ and $\delta = \infty$. Suppose first that $\delta = -\infty$, and assume, by contradiction, that there exists a sequence of policies $\{\pi^{\lambda}\}$, with $\pi^{\lambda} \in \Pi(\lambda, \vec{N}^{\lambda})$, such that $\limsup_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\pi^{\lambda}}(\text{ab}) = \Delta < \infty$. Let $-\infty < \delta_0 < \infty$ be such that $\Delta(\delta, \mu_1, \theta) = \Delta$ (δ_0 exists due to Lemma B.1). Consider another sequence of systems with server pools of size $\tilde{N}^{\lambda}_k = N^{\lambda}_k + (\delta_0/4)\sqrt{\lambda}/\mu_k$, $k = 1, 2$. Clearly, (5.7) holds for the new sequence, with $\delta = \delta_0/2$. Now, according to Corollary 4.2, if FSF is used with the new sequence of systems, then $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\text{FSF}}(\text{ab}) = \Delta(\delta_0/2, \mu_1, \theta) > \Delta$. However, $\{\pi^{\lambda}\}$ is assumed to obtain a scaled abandonment probability of Δ (asymptotically, over a subsequence) by using only a subset of the servers $(\tilde{N}^{\lambda}_1, \tilde{N}^{\lambda}_2)$. This is a contradiction to the asymptotic optimality of FSF (Theorem 4.1). Finally, if $\mu_1 N^{\lambda}_1 + \mu_2 N^{\lambda}_2 \geq \lambda + \delta\sqrt{\lambda} + o(\sqrt{\lambda})$ for all $\delta < \infty$, then by using FSF with a subset of the servers, one can obtain that $\limsup_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}(\text{ab}) = \Delta$, for all $0 < \Delta < \infty$. ∎

Proof of Proposition 5.3: Let $\vec{M}^{*\lambda}$ be the non-negative vector on the half-plain $\mu_1 M_1 + \mu_2 M_2 + ... + \mu_K M_K \geq \lambda(1 - \Delta)$ that minimizes the staffing cost $C(M)$, $\lambda > 0$. Clearly, $\mu_1 M^{*\lambda}_1 + \mu_2 M^{*\lambda}_2 + ... + \mu_K M^{*\lambda}_K = \lambda(1 - \Delta)$. Let $\tilde{N}^{\lambda}_k = \lceil M^{*\lambda}_k \rceil$, $k = 1, ..., K$. We prove (5.14), which also implies the validity of (5.15). The outline of the proof is as follows: We solve for $\vec{M}^{*\lambda}$, and $C(\vec{M}^{*\lambda})$ for all $\lambda > 0$, and then assume by contradiction that $\limsup_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^p} > 0$, and without loss of generality assume that

$$\lim_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^p} = \epsilon > 0. \tag{A.7}$$

1. Assuming first that $C(\vec{M}^{*\lambda_n}) < C(\vec{N}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.7) implies that for all n large enough there exists a staffing vector $\vec{L}^{\lambda_n}$ which is feasible for the problem (5.12), but $C(\vec{L}^{\lambda_n}) < C(\vec{N}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{N}^{*\lambda_n}$.

2. Assuming now that $C(\vec{N}^{*\lambda_n}) < (\vec{M}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.7) implies that for all n large enough there exists a staffing vector $\vec{L}^{\lambda_n}$ which is feasible for the problem SP($\lambda(1 - \Delta)$), but $C(\vec{L}^{\lambda_n}) < C(\vec{M}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{M}^{*\lambda_n}$.

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

$$\begin{array}{lrcl} \text{minimize} & c_1 M_1^p + c_2 M_2^p + ... + c_K M_K^p & & \\ \text{subject to} & \mu_1 M_1 + \mu_2 M_2 + ... \mu_K M_K & \geq & \lambda(1-\Delta) \\ & M_1, M_2, ..., M_K & \geq & 0\,. \end{array} \tag{A.8}$$

The solution to (A.8) is as given in (5.13), and the corresponding optimal cost is:

$$C(\vec{M}^{*\lambda}) = (\lambda(1-\Delta))^p \xi, \tag{A.9}$$

where $\xi = \min\{C(\vec{x}) \mid \mu_1 x_1 + ... + \mu_K x_K \geq 1\}$.

1. Assume that (A.7) holds and that, without loss of generality, $C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda})$ for all λ. By (A.7), we have that for all λ large enough

$$C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda}) \geq \lambda^p \frac{\epsilon}{2}\,.$$

Let $\vec{M}^{\lambda}$ be the solution of SP$(\lambda(1-\Delta+\eta))$, where $0 < \eta < \Delta$ is such that $(1-\Delta+\eta)^p - (1-\Delta)^p < \epsilon/(4\xi)$. Then,

$$\frac{C(\vec{M}^{\lambda}) - C(\vec{M}^{*\lambda})}{\lambda^p} = \xi\left((1-\Delta+\eta)^p - (1-\Delta)^p\right) < \frac{\epsilon}{4}.$$

In particular, $C(\vec{M}^{\lambda}) < \lambda^p \frac{\epsilon}{4} + C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda}) - \lambda^p \frac{\epsilon}{4}$ for all λ large enough. Let $\vec{L}^{\lambda}$ be such that $L_k^{\lambda} = \lceil M_k^{\lambda} \rceil$, $k = 1, 2, ..., K$, for all λ. Then, for all λ large enough, we also have that

$$C(\vec{L}^{\lambda}) \leq C(\vec{N}^{*\lambda}) - \lambda^p \frac{\epsilon}{8} < C(\vec{N}^{*\lambda})\,. \tag{A.10}$$

Now, note that by the results of Section 4.1 we have that, when staffing the λ system with $\vec{L}^{\lambda}$, and using any work-conserving policy

$$\lim_{\lambda \to \infty} P^{\lambda}(ab) = \Delta - \eta < \Delta.$$

In particular, $P^{\lambda}(ab) \leq \Delta$ for all λ large enough, which implies that $\vec{L}^{\lambda}$ is a feasible solution of (5.12), which by (A.10) is a contradiction to the optimality of $\vec{N}^{*\lambda}$.

2. Assume that (A.7) holds and that, without loss of generality, $C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda})$ for all λ. By (A.7), we have that for all λ large enough

$$C(\vec{M}^{*\lambda}) - C(\vec{N}^{*\lambda}) \geq \lambda^p \frac{\epsilon}{2}\,.$$

By the optimality of $\vec{M}^{*\lambda}$, we have that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} + ... + \mu_K N_K^{*\lambda} < \lambda(1-\Delta)$ for all λ large enough. By the feasibility of $\vec{N}^{*\lambda}$, we have that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} + ... + \mu_K N_K^{*\lambda} \geq \lambda(1-\Delta) + o(\lambda)$.

Let $f^\lambda := \lambda(1-\Delta)-(\mu_1 N_1^*+\mu_2 N_2^*+...+\mu_K N_K^*)$. Then, $f^\lambda > 0$ and $f^\lambda = o(\lambda)$. Let $\vec{L}^\lambda = b^\lambda \cdot \vec{N}^{*\lambda}$, where $b^\lambda := \frac{\lambda(1-\Delta)}{\lambda(1-\Delta)-f^\lambda} > 1$. Then, one can verify that $\mu_1 L_1^\lambda + \mu_2 L_2^\lambda + ... + \mu_K L_K^\lambda = \lambda(1-\Delta)$. In particular,

$$C(\vec{L}^\lambda) \geq C(\vec{M}^{*\lambda})\,. \tag{A.11}$$

Note that $(b^\lambda)^p = \frac{1}{\left(1-\frac{f^\lambda}{\lambda(1-\Delta)}\right)^p} = \frac{1}{1-p\,\frac{f^\lambda}{\lambda(1-\Delta)}+o(1/\lambda)} = 1 + p\,\frac{f^\lambda}{\lambda(1-\Delta)} + o(1/\lambda)$. We now have that

$$\begin{aligned}\frac{C(\vec{L}^{*\lambda}) - C(N^{*\lambda})}{\lambda^p} &= \frac{C(\vec{N}^{*\lambda})((b^\lambda)^p - 1)}{\lambda^p} = \frac{C(\vec{N}^{*\lambda})\left(p\,\frac{f^\lambda}{\lambda(1-\Delta)} + o\left(1/\lambda\right)\right)}{\lambda^p} \\ \leq\ & \frac{C(M^{*\lambda})\left(p\,\frac{f^\lambda}{\lambda(1-\Delta)} + o\left(1/\lambda\right)\right)}{\lambda^p} = \frac{\xi\left(\lambda(1-\Delta)\right)^p\left(p\,\frac{f^\lambda}{\lambda(1-\Delta)} + o\left(1/\lambda\right)\right)}{\lambda^p} \to 0 \ \text{ as } \ \lambda\to\infty\,.\end{aligned}$$

In particular,

$$C(\vec{L}^{*\lambda}) \leq \lambda^p\,\frac{\epsilon}{4} + C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda}) - \lambda^p\,\frac{\epsilon}{4} < C(\vec{M}^{*\lambda}),$$

for all λ large enough. This is in contradiction to (A.11). ∎

Proof of Proposition 5.4: Let $\vec{M}^{*\lambda}$ be the non-negative vector on the half-plain $\mu_1 M_1 + \mu_2 M_2 + ... + \mu_K M_K \geq \lambda(1-\Delta)$ that minimizes the staffing cost $C(M)$, $\lambda > 0$. Clearly, $\mu_1 M_1^{*\lambda} + \mu_2 M_2^{*\lambda} + ... + \mu_K M_K^{*\lambda} = \lambda(1-\Delta)$. Let $\tilde{N}_k^\lambda = \lceil M_k^{*\lambda} \rceil$, $k = 1, ..., K$. We prove that $||\vec{N}^{*\lambda} - \vec{M}^{*\lambda}|| = o(\lambda)$, which automatically shows that $||\vec{N}^{*\lambda} - \tilde{N}^\lambda|| = o(\lambda)$.

By contradiction, suppose that, without loss of generality, $\lim_{\lambda\to\infty} \frac{||\vec{N}^{*\lambda} - \vec{M}^{*\lambda}||}{\lambda} = \epsilon > 0$. Let $\vec{x}$ be the optimal solution to the problem $\min\{C(\vec{x}) \mid \mu_1 x_1 + ... + \mu_K x_K = 1\}$. Then from homogeneity, $\vec{M}^{*\lambda} = \lambda(1-\Delta)\vec{x}$. Now, for any subsequence $\{\lambda_n\}$ with $\lim_{n\to\infty} \frac{N_k^{*\lambda_n}}{\lambda(1-\Delta)} = y_k$ we have that $\vec{x} \neq \vec{y}$. The latter follows from the contradicting assumption due to the fact that $\sum_{m=1}^K \mu_m N_m^{*\lambda_n} - \sum_{m=1}^K \mu_m M_m^{*\lambda_n} = o(\lambda)$ (by Proposition 5.1). This implies that

$$\begin{aligned}\tfrac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^p} &= \tfrac{\left|\sum_{k=1}^K c_k[(N_k^{*\lambda})^p - (M_k^{*\lambda})^p]\right|}{\lambda^p} \\ &\to \left|(1-\Delta)^p \textstyle\sum_{k=1}^K c_k(x_k^p - y_k^p)\right| > 0,\end{aligned} \tag{A.12}$$

where the convergence is as $\lambda \to \infty$ and the last inequality is due to the strict convexity of $C(\cdot)$ and the fact that $\sum_{k=1}^K \mu_k x_k = \sum_{k=1}^K \mu_k y_k = 1$. ∎

Proof of Proposition 5.5: 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 (5.17), which also implies the validity of (5.18). The outline of the proof is as follows:

1. We solve for $\vec{M}^{*\lambda}$, and $C(\vec{M}^{*\lambda})$ for all $\lambda > 0$, and show that $\lceil \vec{M}^{*\lambda} \rceil$ satisfies the conditions of Proposition 5.2. Now assume by contradiction that $\limsup_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} > 0$, and without loss of generality assume that

$$\lim_{\lambda\to\infty} \frac{|C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda})|}{\lambda^{p-1/2}} = \epsilon > 0. \tag{A.13}$$

2. Assuming first that $C(\vec{M}^{*\lambda_n}) \leq C(\vec{N}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.13) implies that for all n large enough there exists a staffing vector $\vec{L}^{\lambda_n}$ which is feasible for the problem (5.11), but $C(\vec{L}^{\lambda_n}) < C(\vec{N}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{N}^{*\lambda_n}$.

3. Assuming now that $C(\vec{N}^{*\lambda_n}) \leq C(\vec{M}^{*\lambda_n})$ on a subsequence $\{\lambda_n\}$, we show that (A.13) implies that for all n large enough there exists a staffing vector $\vec{L}^{\lambda_n}$ which is feasible for the problem SP($\lambda+\delta\sqrt{\lambda}$), but $C(\vec{L}^{\lambda_n}) < C(\vec{M}^{*\lambda_n})$, which is a contradiction to the optimality of $\vec{M}^{*\lambda_n}$.

We now proceed with the details of steps 1-3.

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

$$\begin{array}{lrcl} \text{minimize} & c_1 M_1^p + c_2 M_2^p & & \\ \text{subject to} & \mu_1 M_1 + \mu_2 M_2 & \geq & \lambda + \delta\sqrt{\lambda} \\ & M_1, M_2 & \geq & 0\,. \end{array} \tag{A.14}$$

The solution to (A.14) is given in (5.16), 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

$$C(\vec{M}^{*\lambda}) = \frac{(\lambda + \delta\sqrt{\lambda})^p c_1 c_2}{\left((\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)}\right)^{p-1}} \stackrel{\Delta}{=} (\lambda + \delta\sqrt{\lambda})^p \xi. \tag{A.15}$$

In particular, $\lceil \vec{M}^{*\lambda} \rceil$ satisfies condition (5.5) of Proposition 5.2, 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. Assume that (A.13) holds and that, without loss of generality, $C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda})$ for all λ. By (A.13), we have that for all λ large enough

$$C(\vec{N}^{*\lambda}) - C(\vec{M}^{*\lambda}) \geq \lambda^{p-\frac{1}{2}} \frac{\epsilon}{2}\,.$$

Let $\vec{M}^\lambda$ be the solution of SP$(\lambda + (\delta+\eta)\sqrt{\lambda})$, with $\eta = \frac{\epsilon}{8\xi p}$. Then,

$$\frac{C(\vec{M}^\lambda) - C(\vec{M}^{*\lambda})}{\lambda^{p-\frac{1}{2}}} = \xi\lambda^p \frac{\left(1+\frac{\delta+\eta}{\sqrt{\lambda}}\right)^p - \left(1+\frac{\delta}{\sqrt{\lambda}}\right)^p}{\lambda^{p-\frac{1}{2}}}$$

$$= \xi\lambda^p \frac{1 + p(\delta+\eta)/\sqrt{\lambda} - 1 - p\,\delta/\sqrt{\lambda} + o(1/\sqrt{\lambda})}{\lambda^{p-\frac{1}{2}}} = \xi[p\eta + o(1)] < \frac{\epsilon}{4}\,,$$

for all λ large enough.

In particular, $C(\vec{M}^\lambda) < \lambda^{p-\frac{1}{2}}\,\frac{\epsilon}{4} + C(\vec{M}^{*\lambda}) \leq C(\vec{N}^{*\lambda}) - \lambda^{p-\frac{1}{2}}\,\frac{\epsilon}{4}$ for all λ. Let $\vec{L}^\lambda$ be such that $L_k^\lambda = \lceil M_k^\lambda \rceil$, $k = 1, 2$, for all λ. Then, for all λ large enough, we also have that

$$C(\vec{L}^\lambda) \leq C(\vec{N}^{*\lambda}) - \lambda^{p-\frac{1}{2}}\,\frac{\epsilon}{8} < C(\vec{N}^{*\lambda})\,. \tag{A.16}$$

Now, note that by Corollary 4.2, we have that, when staffing the λ system with $\vec{L}^\lambda$,

$$\lim_{\lambda\to\infty} \sqrt{\lambda}\, P^\lambda_{\rm FSF}(ab) = \Delta(\delta+\eta, \mu, \theta) < \Delta(\delta, \mu, \theta)\,.$$

In particular, $\sqrt{\lambda}\, P^\lambda_{\rm FSF}(ab) \leq \Delta(\delta, \mu, \theta)$ for all λ large enough, which implies that $\vec{L}^\lambda$ is a feasible solution of (5.11), which by (A.16) 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 B.2 *Suppose that for all $\lambda > 0$, $\vec{N}^{*\lambda}$ is an optimal solution of (5.11) 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 5.2, $\limsup_{j\to\infty} \sqrt{\lambda_j} P^{\lambda_j}_{\pi^{\lambda_j}}(ab) > \Delta$, 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 5.2, there exists a sequence of policies $\{\pi^{\lambda_j}\}$, with $\pi^\lambda \in \Pi(\lambda_j, \vec{N}^{\lambda_j})$ under which $\limsup_{j\to\infty} \sqrt{\lambda} P^{\lambda_j}_{\pi^{\lambda_j}}(ab) < \Delta$. This is a contradiction to the optimality of $\vec{N}^{*\lambda_j}$ for all large j. ∎

Lemma B.3 *Let $\vec{N}^\lambda$ be a sequence of staffing vectors satisfying $\mu_1 N_1^\lambda + \mu_2 N_2^\lambda < \lambda + \delta\sqrt{\lambda}$ for some $-\infty < \delta < \infty$, and $\lim_{\lambda\to\infty} \frac{N_1^\lambda}{N_1^\lambda + N_2^\lambda} = 0$. Suppose that there exists a sequence of policies $\{\pi^\lambda\}$, with $\pi^\lambda \in \Pi(\lambda, \vec{N}^\lambda)$ such that $\lim_{\lambda\to\infty} \sqrt{\lambda} P^\lambda_{\pi^\lambda}(ab) = \Delta$, where $\Delta = \Delta(\delta, \mu_1, \theta)$. Then, $\vec{N}^\lambda$ satisfies*

$$\mu_1 N_1^\lambda + \mu_2 N_2^\lambda \geq \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right)\,. \tag{A.17}$$

Proof: Let $\delta_1 = \lim_{n\to\infty} \frac{\mu_1 N_1^{\lambda_n}}{\sqrt{\lambda_n}},\ 0 \le \delta_1 \le \infty$ and let $\delta_2 = \lim_{n\to\infty} \frac{\mu_2 N_2^{\lambda_n} - \lambda_n}{\sqrt{\lambda_n}},\ -\infty \le \delta_2 \le \infty$, where $\{\lambda_n\}$ is a subsequence along which these limits are well defined. Without loss of generality, assume that $\{\lambda_n\} \equiv \{\lambda\}$. We show that if the policy FSF_p is used to process the system (which by Proposition 3.2 implies that $\limsup_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) \le \Delta$), then (A.17) is satisfied.

We consider three different cases with respect to the value of δ_1: (a) $0 < \delta_1 < \infty$, (b) $\delta_1 = 0$ and (c) $\delta_1 = \infty$, and note that since $\delta_1 + \delta_2 < \delta$, we have $\delta_2 < \infty$. Denote $\tilde{\delta} := \delta_1 + \delta_2$.

Case (a) ($\mathbf{0 < \delta_1 < \infty}$): First suppose that $\delta_2 > -\infty$. In this case, we can show, using Stone's criterion and the birth-and-death representation of the total number in the system given in (3.1), that the scaled process $X^{\lambda}(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\tilde{\delta}\sqrt{\mu_2} - \theta x & x \ge 0 \\ -\tilde{\delta}\sqrt{\mu_2} - \mu_1 x & -\frac{\delta_1}{\mu_1}\sqrt{\mu_2} \le x < 0 \\ -\mu_2\left(\frac{\delta_1}{\mu_1} + \frac{\delta_2}{\mu_2}\right)\sqrt{\mu_2} - \mu_2 x & x < -\frac{\delta_1}{\mu_1}\sqrt{\mu_2} \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$.

Let $\underline{X}$ be another diffusion process with the same infinitesimal variance and with infinitesimal drift

$$\underline{m}(x) = \begin{cases} -\tilde{\delta}\sqrt{\mu_2} - \theta x & x \ge 0 \\ -\tilde{\delta}\sqrt{\mu_2} - \mu_1 x & x < 0\,. \end{cases}$$

Then, clearly $m(x) \ge \underline{m}(x)$ for all x, and therefore, by Proposition 18.5 of [2] we have that $X(\infty) \overset{st}{\ge} \underline{X}(\infty)$. In particular, $\frac{\theta}{\sqrt{\mu_2}}\, EX^{+}(\infty) \ge \frac{\theta}{\sqrt{\mu_2}}\, E\underline{X}^{+}(\infty) = \Delta(\tilde{\delta}, \mu_1, \theta)$. Therefore, by establishing that $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \frac{\theta}{\sqrt{\mu_2}} EX^{+}(\infty)$, the proof is complete for this case (recalling that Δ is a decreasing function of δ (see Lemma B.1). Note that this latter limit holds due to the tightness and uniform integrability results established in the proofs of Proposition 4.3 (step 2) and of Theorem 4.1 (step 2).

Next, consider the case $\delta_2 = -\infty$. We claim that in this case, $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \infty$. To see this, we first note that if $-\infty < \delta_2 < \infty$, then by [2] along with the tightness and uniform integrability results, we have that

$$\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \tilde{\Delta}(\delta_1, \delta_2, \mu_1, \mu_2, \theta) := \sqrt{\theta}\, \alpha_1 \left[h\left(\frac{\delta_1 + \delta_2}{\sqrt{\theta}}\right) - \frac{(\delta_1 + \delta_2)}{\sqrt{\theta}} \right],$$

where $\alpha_1 := \alpha_1(\delta_1, \delta_2, \mu_1, \mu_2, \theta) = \left[1 + \sqrt{\frac{\theta}{\mu_1}} h\left(\frac{\tilde{\delta}}{\sqrt{\theta}}\right) \frac{\Phi(\tilde{\delta}/\sqrt{\mu_1}) - \Phi(\delta_2/\sqrt{\mu_1})}{\phi(\tilde{\delta}/\sqrt{\mu_1})} + \sqrt{\frac{\theta}{\mu_2}} \frac{h(\tilde{\delta}/\theta)}{h(-\delta_2/\sqrt{\mu_2})}\right]^{-1}$, with $\tilde{\delta} := \delta_1 + \delta_2$. Note that $\tilde{\Delta}$ is continuous in δ_2 and that $\lim_{\delta_2\to\infty} \tilde{\Delta} = 0$ and $\lim_{\delta_2\to-\infty} \tilde{\Delta} = \infty$ (these limits may be obtained by a successive application of L'Hôpital's rule). Therefore, by an

argument analogous to the proof of Proposition 5.2 for the case $\delta = -\infty$, one can show that indeed if $\delta_2 = -\infty$ then $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \infty$. This finally leads to a contradiction due to the optimality of FSF_p.

Case (b) ($\boldsymbol{\delta_1 = 0}$): In this case, if $\delta_2 > -\infty$, then one can show that the scaled process $X^{\lambda}(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\delta_2\sqrt{\mu_2} - \theta x & x \geq 0 \\ -\delta_2\sqrt{\mu_2} - \mu_2 x & x < 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$. Consider another diffusion process $\underline{X}$ with the same infinitesimal variance and with infinitesimal drift equal to

$$\underline{m}(x) = \begin{cases} -\delta_2\sqrt{\mu_2} - \theta x & x \geq 0 \\ -\delta_2\sqrt{\mu_2} - \mu_1 x & x < 0\,. \end{cases}$$

Then, clearly, $m(x) \geq \underline{m}(x)$ for all x, and hence, by analogous arguments to the ones used in case (a), we have that

$$\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) \geq \Delta(\delta_2, \mu_1, \theta)\,,$$

which implies by Lemma B.1 that $\delta_2 \geq \delta$, and in turn, that $\mu_1 N_1^{\lambda} + \mu_2 N_2^{\lambda} \geq \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right)$. The case $\delta_2 = -\infty$ may be analyzed analogously to $\delta_2 = -\infty$ in case (a) to show that if $\delta_2 = -\infty$ then $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \infty$, which leads to a contradiction.

Case (c) ($\boldsymbol{\delta_1 = \infty}$): In this case, since $\delta_1 + \delta_2 := \lim_{\lambda\to\infty} \frac{\mu_1 N_1 + \mu_2 N_2 - \lambda}{\sqrt{\lambda}} \leq \delta$, we have that necessarily, $\delta_2 = -\infty$. Assume first that $\delta_1 + \delta_2 = \tilde{\delta} > -\infty$. Then, in this case, the scaled process $X^{\lambda}(t)$ weakly converges to a diffusion process $X(t)$ with infinitesimal drift

$$m(x) = \begin{cases} -\tilde{\delta}\sqrt{\mu_2} - \theta x & x \geq 0 \\ -\tilde{\delta}\sqrt{\mu_2} - \mu_1 x & x < 0 \end{cases}$$

and infinitesimal variance $\sigma^2(x) = 2\mu_2$. This process has the same law as the diffusion process defined through (4.19) and (4.20) with $\delta = \tilde{\delta}$ and $\mu = \mu_2$. In particular, one can show that $\lim_{\lambda\to\infty} \sqrt{\lambda} P^{\lambda}_{\mathrm{FSF}_p}(ab) = \Delta(\tilde{\delta}, \mu_1, \theta)$, which, in turn, implies that $\tilde{\delta} \geq \delta$, so that $\mu_1 N_1 + \mu_2 N_2 \geq \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right)$. Finally, if $\delta_1 + \delta_2 = -\infty$, one can obtain a contradiction in the same way that was done for cases (a) and (b). ∎

We now return to step 3. in the proof of Proposition 5.5.

3. Assume that (A.13) holds and that, without loss of generality, $C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda})$ for all λ. By (A.13), we have that for all λ large enough

$$C(\vec{M}^{*\lambda}) - C(\vec{N}^{*\lambda}) \geq \lambda^{p-\frac{1}{2}} \frac{\epsilon}{2}\,.$$

By the optimality of $\vec{M}^{*\lambda}$, we have that $\mu_1 N_1^{*\lambda}+\mu_2 N_2^{*\lambda} < \lambda+\delta\sqrt{\lambda}$ for all λ large enough. Therefore, by Lemmas B.2 and B.3, we have that $\mu_1 N_1^{*\lambda} + \mu_2 N_2^{*\lambda} \geq \lambda + \delta\sqrt{\lambda} + o\left(\sqrt{\lambda}\right)$.

Let $f^\lambda := \lambda + \delta\sqrt{\lambda} - (\mu_1 N_1^* + \mu_2 N_2^*)$. Then, $f^\lambda > 0$ and $f^\lambda = o\left(\sqrt{\lambda}\right)$. Let $\vec{L}^\lambda = b^\lambda \cdot \vec{N}^{*\lambda}$, where $b^\lambda := \frac{\lambda+\delta\sqrt{\lambda}}{\lambda+\delta\sqrt{\lambda}-f^\lambda} = \left(1 - \frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}}\right)^{-1}$. Then, one can verify that $\mu_1 L_1^\lambda + \mu_2 L_2^\lambda = \lambda + \delta\sqrt{\lambda}$. In particular,

$$C(\vec{L}^\lambda) \geq C(\vec{M}^{*\lambda})\,. \tag{A.18}$$

Note that $(b^\lambda)^p = \frac{1}{\left(1-\frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}}\right)^p} = \frac{1}{1-p\,\frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}}+o(1/\sqrt{\lambda})} = 1 + p\,\frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}} + o(1/\sqrt{\lambda})$. We now have that

$$\begin{aligned}
\frac{C(\vec{L}^{*\lambda}) - C(N^{*\lambda})}{\lambda^{p-\frac{1}{2}}} &= \frac{C(\vec{N}^{*\lambda})((b^\lambda)^p - 1)}{\lambda^{p-\frac{1}{2}}} = \frac{C(\vec{N}^{*\lambda})\left(p\,\frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} \\
&\leq \frac{C(M^{*\lambda})\left(p\,\frac{f^\lambda}{\lambda+\delta\sqrt{\lambda}} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} = \frac{\xi\left(\lambda+\delta\sqrt{\lambda}\right)^p\left(p\frac{1}{\sqrt{\lambda}}\frac{f^\lambda}{\sqrt{\lambda}+\delta} + o\left(1/\sqrt{\lambda}\right)\right)}{\lambda^{p-\frac{1}{2}}} \\
&= \xi\left(p\left(1+\frac{\delta}{\sqrt{\lambda}}\right)^p\frac{f^\lambda}{\sqrt{\lambda}+\delta} + o\left(1\right)\right) \to 0 \;\text{ as }\; \lambda\to\infty\,.
\end{aligned}$$

In particular,

$$C(\vec{L}^{*\lambda}) \leq \lambda^{p-\frac{1}{2}}\,\frac{\epsilon}{4} + C(\vec{N}^{*\lambda}) \leq C(\vec{M}^{*\lambda}) - \lambda^{p-\frac{1}{2}}\,\frac{\epsilon}{4} < C(\vec{M}^{*\lambda}),$$

for all λ large enough. This is in contradiction to (A.18). ∎

Proof of Proposition 5.6: 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 $\lim_{\lambda\to\infty} \frac{C(\vec{N}^{*\lambda}) - \underline{C}^\lambda}{C(\vec{M}^{*\lambda}) - \underline{C}^\lambda} = 1$. The asymptotic optimality of $\tilde{N}^\lambda$ then easily follows. The outline of the proof is as follows:

1. We solve for $\underline{C}^\lambda$, and explicitly express $C(\vec{M}^{*\lambda}) - \underline{C}^\lambda$.
2. The asymptotic optimality then follows easily from Proposition 5.5.

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

$$\underline{C}^\lambda = \frac{\lambda^p c_1 c_2}{\left((\mu_1^p c_2)^{1/(p-1)} + (\mu_2^p c_1)^{1/(p-1)}\right)^{p-1}} = \lambda^p \xi. \tag{A.19}$$

If follows that

$$C(\vec{M}^{*\lambda}) - \underline{C}^\lambda = \xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2}).$$

2. By proposition 5.5, we have that $C(\vec{N}^{*\lambda}) = C(\vec{M}^{*\lambda}) + f^{\lambda}$, where $f^{\lambda} = o(\lambda^{p-1/2})$. Therefore,

$$\frac{C(\vec{N}^{*\lambda}) - \underline{C}^{\lambda}}{C(\vec{M}^{*\lambda}) - \underline{C}^{\lambda}} = \frac{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2}) + f^{\lambda}}{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2})} = 1 + \frac{f^{\lambda}}{\xi p \delta \lambda^{p-1/2} + o(\lambda^{p-1/2})} \to 1,$$

as $\lambda \to \infty$, provided that $\delta \neq 0$.

∎

Proof of Proposition 6.1: The proof is analogous to the proof of Proposition 3.1 in [2]. The details are omitted. ∎

Proof of Corollary 6.1: The proof follows from Proposition 6.1 and the work conservation of FSF$_p$. For the queue length, the proof directly follows from the relationship $Q(t; \text{FSF}_p) = [Y(t; \text{FSF}_p) - N]^+$ and $Q(t; \pi) = [Y(t; \pi) - N]^+$ for all $t \geq 0$ and for a general FCFS policy in Π_p.

For the virtual waiting time, we have that, for $Y := Y(\infty; \pi)$ for some $\pi \in \Pi_p$ which is FCFS and work-conserving,

$$V(\infty; \pi) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{[Y-N+1]^+} T_i,$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution, and $T_i \sim \exp(\sum_{k=1}^{K} \mu_k N_k + (i-1)\theta)$, and all T_i's are independent. If π is FCFS, and not work-conserving, but $\theta \leq \mu_1$, then we have that

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

Since $\sum_{i=1}^{[Y-N+1]^+} T_i$ is an increasing function of Y, the stochastic dominance of FSF$_p$ with respect to Y, implies that FSF$_p$ also stochastically minimizes the steady-state virtual waiting time.

Finally, the stochastic dominance of FSF$_p$ with respect to the actual steady-state waiting time, follows from the relationship $W(\infty) = V(\infty) \wedge \tau$, where $\tau \sim \exp(\theta)$ is independents of $V(\infty)$. ∎

Proof of Proposition 6.2: The first step in the proof is to show that if the staffing vector satisfies (6.3) and under any work-conserving policy, we have that (6.4) is satisfied with $T = G^{-1}(\Delta_1)$ and $\alpha = (1 - G(T)) \cdot \left(\delta_1 / \sqrt{g(T)}\right)$. To see that, consider the three systems of Proposition 3.3 with $N_B = \lfloor \frac{1}{\mu_K} \sum_{k=1}^{K} \mu_k N_k \rfloor$, and $N_C = \lceil \frac{1}{\mu_1} \sum_{k=1}^{K} \mu_k N_k \rceil$. By [32, Remark 4.5] we have that for both systems B and C, $\lim_{\lambda \to \infty} P(W^{\lambda} > T) = (1 - G(T))\bar{\Phi}(\delta / g(T))$. The proof then follows from the fact that $W_C^{\lambda} \stackrel{st}{\leq} W_A^{\lambda} \stackrel{st}{\leq} W_B^{\lambda}$, where $W_A^{\lambda}, W_B^{\lambda}$, and W_C^{λ} are the steady-state waiting times for systems A, B and C, respectively. The latter is a straightforward consequence of Proposition 3.3.

The rest of the proof is analogous to the proofs of Propositions 5.2 and 5.1. Details are omitted. ∎

References

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

[2] 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.

[3] J. G. Dai and T. Tezcan (2005), State space collapse in many-server diffusion limits of parallel server systems. Working paper.

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

[5] Gurvich, I. and Whitt, W. (2007) Queue-and-Idleness-Ratio Controls in Many-Server Service Systems. *Mathematics of Operations Research*, forthcoming.

[6] T. Tezcan and J.G. Dai (2006) Dynamic control of N-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research*, forthcoming.