

**Technical Appendix for:**

**When Promotions Meet Operations:**

**Cross-Selling and Its Effect on Call-Center Performance**

In this technical appendix we provide proofs for the various results stated in the manuscript titled: “When promotions meet operations: Cross-selling and its effect on call-center performance”.

We begin the technical appendix with a formal construction of the sample paths under the  $TP[K]$  control.

**A. Sample-path construction**

The sample path construction follows a strong approximation approach (see for example [7] and [8]). Let  $\mathcal{N}_i(\cdot)$ ,  $i = 1, \dots, 11$ , be independent unit rate Poisson processes. Then, under the  $TP[K]$  control, one can write the system dynamics through the following equations:

$$\begin{aligned}
 Q^\lambda(t) + Z_1^\lambda(t) &= Q^\lambda(0) + Z_1^\lambda(0) + \tilde{\mathcal{N}}_A(t) - \tilde{\mathcal{N}}_{D_1}(t), & (A1) \\
 Z_2^\lambda(t) &= Z_2^\lambda(0) - \tilde{\mathcal{N}}_{D_2}(t) \\
 &+ 1_{\{K^\lambda \geq 0\}} \left[ \mathcal{N}_7 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Q^\lambda(u) \leq K^\lambda\}} du \right) \right. \\
 &+ \left. \mathcal{N}_5 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \right] \\
 &+ 1_{\{K^\lambda < 0\}} \mathcal{N}_8 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \right). & (A2)
 \end{aligned}$$

Here, we use a time change of the processes  $\mathcal{N}_7(\cdot)$  and  $\mathcal{N}_5(\cdot)$  to construct service completions with cross-selling candidates that are followed by actual cross-selling. These two processes are used if the threshold is non-negative. In this case, a service completion is followed by a cross-selling offer whenever the queue is equal to 0 or when it is greater than 0 but smaller than the threshold. We use  $\mathcal{N}_8(\cdot)$  similarly for the cases in which the threshold is negative. In this case, a service completion is followed by a cross-selling offer if the number of idle servers is less than  $K^\lambda$ . The processes  $\tilde{\mathcal{N}}_A(t)$ ,  $\tilde{\mathcal{N}}_{D_1}(t)$  and  $\tilde{\mathcal{N}}_{D_2}(t)$  will be defined shortly.

Also, we have that

$$\begin{aligned}
Z_1^\lambda(t) &= Z_1^\lambda(0) + \mathcal{N}_1 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du \right) \\
&+ \mathcal{N}_3 \left( \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du \right) \\
&- 1_{\{K^\lambda \geq 0\}} \left[ \mathcal{N}_5 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) + \mathcal{N}_6 \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \right] \\
&- 1_{\{K^\lambda \geq 0\}} \mathcal{N}_7 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Q^\lambda(u) \leq K^\lambda\}} du \right) \\
&- 1_{\{K^\lambda < 0\}} \mathcal{N}_8 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \right) \\
&- 1_{\{K^\lambda < 0\}} \left[ \mathcal{N}_9 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{K^\lambda < Y^\lambda(u) - N^\lambda \leq 0\}} du \right) + \mathcal{N}_{10} \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \right]
\end{aligned}$$

Here, in addition to the previous processes, we use a time change of  $\mathcal{N}_1(\cdot)$  to construct the arrivals that occur when some servers are idle. Both processes  $\mathcal{N}_6(\cdot)$  and  $\mathcal{N}_{10}(\cdot)$  are used to construct service completions with customers that are not cross-selling candidates when the queue is empty.  $\mathcal{N}_6(\cdot)$  is used for non-negative thresholds and  $\mathcal{N}_{10}(\cdot)$  is used for negative ones. The process  $\mathcal{N}_9(\cdot)$  is used to construct service completions with cross-selling candidates when the queue is empty (for negative threshold), and, finally, the process  $\mathcal{N}_3(\cdot)$  is used to construct cross-selling completions when the queue is positive. Note that the event epochs, in which there is a service completion that is followed by admission of a customer to service, are not modeled above as these transitions keep  $Z_1^\lambda$  unchanged.

We construct the aggregate-arrival process,  $\tilde{\mathcal{N}}_A(t)$ , and the process of cross-selling completions,  $\tilde{\mathcal{N}}_{D_2}(t)$ , as follows:

$$\begin{aligned}
\tilde{\mathcal{N}}_A(t) &:= \mathcal{N}_1 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du \right) + \mathcal{N}_2 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) \geq N^\lambda\}} du \right). \\
\tilde{\mathcal{N}}_{D_2}(t) &:= \mathcal{N}_3 \left( \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du \right) + \mathcal{N}_4 \left( \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right).
\end{aligned}$$

Here, we use  $\mathcal{N}_2(\cdot)$  to construct arrivals when all servers are busy and  $\mathcal{N}_4(\cdot)$  to construct cross-

selling completions when the queue is empty. Finally, we have

$$\begin{aligned}
\tilde{\mathcal{N}}_{D_1}(t) &= 1_{\{K^\lambda \geq 0\}} \mathcal{N}_5 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u)=0\}} du \right) \\
&+ 1_{\{K^\lambda \geq 0\}} \mathcal{N}_6 \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u)=0\}} du \right) \\
&+ 1_{\{K^\lambda \geq 0\}} \mathcal{N}_7 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Q^\lambda(u) \leq K^\lambda\}} du \right) \\
&+ 1_{\{K^\lambda < 0\}} \mathcal{N}_8 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \right) \\
&+ 1_{\{K^\lambda < 0\}} \mathcal{N}_9 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{K^\lambda < Y^\lambda(u) - N^\lambda \leq 0\}} du \right) \\
&+ 1_{\{K^\lambda < 0\}} \mathcal{N}_{10} \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u)=0\}} du \right) \\
&+ \mathcal{N}_{11} \left( \int_0^t \hat{\lambda}(u) du \right), \tag{A3}
\end{aligned}$$

where the rate function  $\hat{\lambda}(t)$  is set to satisfy that the sum of the instantaneous rates of all the processes in (A3) equals  $\mu_s Z_1^\lambda(t)$  at time  $t$ .

This construction follows by noting that all input and output processes in the system can be generated through thinning of Poisson processes. By Lemma 9.4 in [7], there exists a probability space  $(\Omega, \mathbb{F}, P)$ , a filtration  $\{\mathcal{F}_t\}_{t \geq 0}$  and an 11-dimensional Brownian Motion  $(B_1(\cdot), \dots, B_{11}(\cdot))$  such that the random variable

$$\mathcal{E}_i := \sup_{t \geq 0} \frac{\mathcal{N}_i(t) - t - B_i(t)}{\log(2 \vee t)}, \tag{A4}$$

has a moment generating function in a neighborhood of the origin and in particular, there exist constants  $c_1, c_2$  and  $\Gamma$ , such that, for all  $i = 1, \dots, 11$ , and all  $x \geq 0$ ,

$$P\{\mathcal{E}_i \geq \Gamma + x\} \leq c_1 e^{-c_2 x}. \tag{A5}$$

We let  $\mathcal{E} := \sum_{i=1}^{11} \mathcal{E}_i$ .

As we consider only cases in which  $N^\lambda \leq R + \frac{\lambda p}{\mu_{cs}}$ , the value of the time change at time  $t$  in each of the equations (A1)-(A3) is bounded by  $c\lambda t$  for some positive constant  $c$ .

We now use (A4) to express the dynamics as follows:

$$\begin{aligned} Q^\lambda(t) &= Q^\lambda(0) + Z_1^\lambda(0) + \lambda t - \mu_s \int_0^t Z_1^\lambda(u) du - Z_1^\lambda(t) + M_{Z,Q}^\lambda(t) \\ &\quad + O(\log(2 \vee c\lambda t)), \end{aligned} \quad (\text{A6})$$

$$\begin{aligned} Z_2^\lambda(t) &= Z_2^\lambda(0) - \mu_{cs} \int_0^t Z_2^\lambda(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du \\ &\quad - p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda > K^\lambda\}} du + M_{Z_2}^\lambda(t) + O(\log(2 \vee c\lambda t)), \end{aligned} \quad (\text{A7})$$

and

$$\begin{aligned} Z_1^\lambda(t) &= Z_1^\lambda(0) + \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du - \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \\ &\quad - p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \\ &\quad + \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du + M_{Z_1}^\lambda(t) + O(\log(2 \vee c\lambda t)) \end{aligned} \quad (\text{A8})$$

Here,  $M_{Z,Q}^\lambda(\cdot)$ ,  $M_{Z_1}^\lambda(\cdot)$  and  $M_{Z_2}^\lambda(\cdot)$  are sums of time changed Brownian motions. For example, if  $K^\lambda > 0$ ,

$$\begin{aligned} M_{Z_1}^\lambda(t) &= B_1 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du \right) \\ &\quad + B_3 \left( \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du \right) \\ &\quad - B_5 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) - B_6 \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \\ &\quad + B_7 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \right), \end{aligned}$$

where  $B_i(\cdot)$ ,  $i \in \{1, 3, 5, 6, 7\}$ , are standard Brownian motions.

Using the Brownian motion strong law of large numbers (see problem 2.9.3. in [6]) and the bound  $c\lambda t$  on the time change values, we have that, uniformly on compact sets,

$$\left( \frac{M_{Z,Q}^\lambda(t)}{\lambda}, \frac{M_{Z_2}^\lambda(t)}{\lambda}, \frac{M_{Z_1}^\lambda(t)}{\lambda} \right) \rightarrow (0, 0, 0), \text{ as } \lambda \rightarrow \infty. \quad (\text{A9})$$

Put

$$\begin{aligned}
T_1^\lambda(t) &:= p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda > K^\lambda\}} du, & T_2^\lambda(t) &:= \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du, \\
T_3^\lambda(t) &:= \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du, & T_4^\lambda(t) &:= \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du, \\
T_5^\lambda(t) &:= p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du,
\end{aligned}$$

and re-write (A6)-(A8) as follows:

$$\begin{aligned}
Q^\lambda(t) &= Q^\lambda(0) + Z_1^\lambda(0) + \lambda t - \mu_s \int_0^t Z_1^\lambda(u) du - Z_1^\lambda(t) + M_{Z,Q}^\lambda(t) \\
&\quad + O(\log(2 \vee c\lambda t)), \tag{A10}
\end{aligned}$$

$$\begin{aligned}
Z_2^\lambda(t) &= Z_2^\lambda(0) - \mu_{cs} \int_0^t Z_2^\lambda(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du - T_1^\lambda(t) + M_{Z_2}^\lambda(t) \\
&\quad + O(\log(2 \vee c\lambda t)), \tag{A11}
\end{aligned}$$

$$\begin{aligned}
Z_1^\lambda(t) &= Z_1^\lambda(0) + T_2^\lambda(t) - \mu_s \int_0^t Z_1^\lambda(u) du + T_3^\lambda(t) + T_4^\lambda(t) - T_5^\lambda(t) \\
&\quad + M_{Z_1}^\lambda(t) + O(2 \vee \log(c\lambda t)). \tag{A12}
\end{aligned}$$

**Notational conventions and organization of the appendix.** Under a  $TP[K]$  policy, the process  $(\Xi^\lambda(t), t \geq 0)$  defined by

$$\Xi^\lambda(t) := (Q^\lambda(t), Z_2^\lambda(t), Z_1^\lambda(t)),$$

is a Markov process. We denote  $\mathcal{X}$  as the state-space of  $\Xi^\lambda(t)$  and use the notation  $\xi$  for a general element in  $\mathcal{X}$ . For a given  $\xi \in \mathcal{X}$  we let  $q(\xi)$ ,  $z_2(\xi)$  and  $z_1(\xi)$  be its corresponding coordinates. We use  $P_\xi\{\cdot\} = P_\xi\{\cdot | \Xi^\lambda(0) = \xi\}$  and  $E_\xi[\cdot] := E_\xi[\cdot | \Xi^\lambda(0) = \xi]$ . If  $\nu^\lambda$  is the steady-state distribution of  $\Xi^\lambda(t)$ <sup>1</sup>, we let  $P_{\nu^\lambda}\{\cdot\}$  be the probability with respect to an initial condition that is distributed according to  $\nu^\lambda$ .  $E_{\nu^\lambda}[\cdot]$  is the corresponding expectation. Finally, for  $x \in \mathbb{R}$ , we let  $x^- := \max\{0, -x\}$  and  $x^+ := \max\{0, x\}$ .

The rest of this appendix is organized as follows. Each of the sections B, C, D is dedicated to the proof of Theorem 3.1 under one of the conditions 1, 2 and 3, respectively. §E is dedicated to the proofs of the results in §4 of the main paper. The proofs of some auxiliary results are relegated to §F. Finally, §G provides some numerical examples that were omitted from §5 of the paper for space considerations.

---

<sup>1</sup>In Lemma F.1 we prove the existence of a steady-state distribution under  $TP[K]$  assuming  $N^\lambda > \lambda/\mu_s$ .

## B. Asymptotic optimality under Condition 1

Asymptotic optimality under Condition 1 is proved in Corollary B.1. The main step is the following theorem.

**Theorem B.1** *Consider a sequence of systems such that: (a) the  $\lambda^{\text{th}}$  system uses  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{\lambda})$  agents for some  $0 < \beta \leq \frac{\underline{\mu}_s}{\mu_{cs}}$  and, (b) the  $\lambda^{\text{th}}$  system uses  $TP[K^\lambda]$  for control with a non-negative sequence  $\{K^\lambda\}_{\lambda \geq 0}$  that satisfies  $K^\lambda/\sqrt{R} \rightarrow \varrho \geq 0$  as  $\lambda \rightarrow \infty$ . Then,*

$$\frac{E[(Q^\lambda - K^\lambda)^+]}{\sqrt{R}} \rightarrow 0 \text{ as } \lambda \rightarrow \infty, \quad (\text{A13})$$

and

$$\frac{E[I^\lambda]}{N^\lambda - R} \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \quad (\text{A14})$$

The main challenge in proving Theorem B.1 arises from the focus on the steady-state behavior rather than on the behavior on compact intervals. Most of the section is dedicated to the proof of this result. We first use Theorem B.1 to prove the asymptotic optimality result for this section:

**Corollary B.1** *Suppose that Assumptions 3.1 and 3.2 hold. If, in addition,  $N_2^\lambda - R \gg N_1^\lambda - R$ , then the following is asymptotically optimal in the sense of Definition 3.2:*

- **Staffing:** Staff with  $N_2^\lambda$  agents.
- **Control:** Use  $TP[K^\lambda]$  with  $K^\lambda \in [0, \lceil \lambda \bar{W}^\lambda \rceil]$ .

**Proof:** By Little's law:

$$\frac{E[W^\lambda]}{\bar{W}^\lambda} = \frac{E[Q^\lambda]}{\lambda \bar{W}^\lambda} \leq \frac{K^\lambda + E[(Q^\lambda - K^\lambda)^+]}{\lambda \bar{W}^\lambda}. \quad (\text{A15})$$

Equation (A13) now implies that

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1, \quad (\text{A16})$$

and in particular that  $TP[K^\lambda]$  is asymptotically feasible.

To establish optimality, recall that  $E[Z_2^\lambda] = N_2^\lambda - E[Z_1^\lambda] - E[I^\lambda]$ . Since, by Little's law  $E[Z_1^\lambda] = \lambda/\mu_s := R$ , (A14) implies that

$$\frac{\mu_{cs}E[Z_2^\lambda]}{\mu_{cs}(N_2^\lambda - R)} \rightarrow 1 \text{ as } \lambda \rightarrow \infty. \quad (\text{A17})$$

For each  $\lambda$ ,  $r\mu_{cs}(N_2^\lambda - R) - (C^\lambda(N_2^\lambda) - C^\lambda(R))$  constitutes an upper bound for the (centered) optimal value of the cross-selling problem given by  $V^*(\lambda) := \sup_{\pi \in \Pi, N \in \mathbb{Z}_+} V^\lambda(\pi, N)$ ; see formulation (6) and the discussion below it as well as Definition 3.2.

Equation (A17) now implies that the upper bound is asymptotically achieved since, then,

$$\frac{r\mu_{cs}E[Z_2^\lambda] - (C^\lambda(N_2^\lambda) - C^\lambda(R))}{r\mu_{cs}(N_2^\lambda - R) - (C^\lambda(N_2^\lambda) - C^\lambda(R))} \rightarrow 1 \text{ as } \lambda \rightarrow \infty.$$

Here we used the fact that, for three sequences  $\{a^\lambda\}$ ,  $\{b^\lambda\}$  and  $\{c^\lambda\}$  such that  $a^\lambda \rightarrow \infty$  and  $a^\lambda/b^\lambda \rightarrow 1$  as  $\lambda \rightarrow \infty$ , we also have that  $(a^\lambda + c^\lambda)/(b^\lambda + c^\lambda) \rightarrow 1$  as  $\lambda \rightarrow \infty$ .

Hence, the sequence of pairs  $\{(N_2^\lambda, TP[K^\lambda])\}$  asymptotically achieves the upper bound and is, in particular, asymptotically optimal. ■

We proceed now to prove Theorem B.1. We prove the theorem in two steps. Proposition B.1 covers (A13) and Corollary B.2 covers (A14).

**Proposition B.1** *Under the conditions of Theorem B.1,*

$$\frac{E[(Q^\lambda - K^\lambda)^+]}{\sqrt{\lambda}} \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \quad (\text{A18})$$

We first provide an informal outline of the proof:

1. **A constrained Lyapunov function argument:** We start by examining the behavior of the Markov process  $(\Xi^\lambda(t), t \geq 0)$  initialized, at time 0, within the set

$$\mathcal{A}_\epsilon^\lambda := \left\{ \xi \in \mathcal{X} : \left( z_1(\xi) - \frac{\lambda}{\mu_s} \right)^- \leq \epsilon\lambda \right\}. \quad (\text{A19})$$

We will show that, if  $\Xi^\lambda(0) \in \mathcal{A}_\epsilon^\lambda$  and  $Q^\lambda(0) > M$  for some constant  $M > 0$  large enough, then the process  $(Q^\lambda(t), t \geq 0)$ , decreases at a certain rate. In other words, we require a negative drift condition, reminiscent of the standard conditions used in Lyapunov function arguments; see e.g. [3]. In contrast to the standard requirement, which is imposed on all  $\xi \in \mathcal{X}$ , we impose this requirement only for  $\xi \in \mathcal{A}_\epsilon^\lambda$ . Hence the term *Constrained Lyapunov function*.

The (constrained) negative drift condition will allow us to obtain bounds on the stationary queue length. This bound will depend, however, on the behavior of the queue length when the process  $\Xi^\lambda(t)$  is not inside the set  $\mathcal{A}_\epsilon^\lambda$ .

2. **Bounding the behavior outside of  $\mathcal{A}_\epsilon^\lambda$ :** We will bound the behavior of the queue length outside of the set  $\mathcal{A}_\epsilon^\lambda$  by: (a) showing that the stationary distribution is, in a sense, concentrated in  $\mathcal{A}_\epsilon^\lambda$  for all  $\lambda$  large enough—see Lemma B.2, and (b) establishing a fluid scale bound on the stationary queue length—see Lemma B.3. This fluid scale bound is weaker than the diffusion-scale one in (A13), but together with step (a) will allow us to bound the stationary queue length on states that are not in  $\mathcal{A}_\epsilon^\lambda$ .

**Proof of Proposition B.1:** We prove the result for  $K^\lambda \equiv 0$ . The proof is extended to arbitrary  $K^\lambda > 0$  by replacing  $Q^\lambda(\cdot)$  everywhere with  $(Q^\lambda(\cdot) - K^\lambda)^+$ .

We first establish bounds on the behavior of the queue length assuming that  $\Xi^\lambda(0) \in \mathcal{A}_\epsilon^\lambda$  with  $\mathcal{A}_\epsilon^\lambda$  as in (A19). Fix constants  $\Theta, T > 0$ , assume that  $Q^\lambda(0) > 2\Theta$  and define the random time<sup>2</sup>

$$\tau^\lambda = \inf\{t \geq 0 : Q^\lambda(t) \leq Q^\lambda(0) - 3\Theta/2\} \wedge \frac{T}{\lambda}.$$

Define the set

$$\Omega^*(\delta, \lambda, T, \Theta) := \left\{ \omega \in \Omega : 11 \cdot \max_{i=1, \dots, 11} \sup_{0 \leq t \leq T} B_i(c\lambda t) + \mathcal{E}_i \log(2 \vee c\lambda t) - \delta\lambda t \leq \Theta \right\}. \quad (\text{A20})$$

We will be using the set  $\Omega^*(\cdot, \cdot, \cdot, \cdot)$  in various proofs in this appendix and set the parameters  $\delta, \lambda, T, \Theta$  according to the need of the specific proof. In any of these cases, we will choose the parameters in a way that will guarantee that the set of sample paths that we consider has sufficiently

---

<sup>2</sup>Note that  $\tau^\lambda$  is a random time but not necessarily a stopping time. Our arguments are sample-path arguments so that whether or not  $\tau^\lambda$  is a stopping time is immaterial for our proofs.

large probability; see Lemma F.2. In the current proof it is important that we use  $T/\lambda$  instead of  $T$ . That is we focus on a small time interval and characterize the behavior of the queue length there. To that end, plugging equation (A8) into equation (A6), and using the fact that  $Z_2^\lambda(t) = N^\lambda - Z_1^\lambda(t)$  whenever  $Q^\lambda(t) > 0$ , we have that on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$

$$Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \lambda(t \wedge \tau^\lambda) - \mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du - \mu_{cs} \int_0^{t \wedge \tau^\lambda} (N^\lambda - Z_1^\lambda(u)) du + \delta \lambda t + \Theta/2. \quad (\text{A21})$$

The following lemma provides a handle on the process  $Z_1^\lambda(t)$  that, in turn, allows us to characterize the behavior of  $Q^\lambda(t)$ .

**Lemma B.1** *Suppose the conditions of Theorem B.1 hold and fix  $\delta > 0$  small enough. Assume that  $\Xi^\lambda(0) \in \mathcal{A}_\epsilon^\lambda$ . Then, for all  $\epsilon > 0$ , there exist  $\lambda^0(\epsilon)$  (independent of the initial conditions) such that, for all  $\omega \in \Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$ ,*

$$\sup_{0 \leq t \leq T/\lambda} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \leq 2\epsilon\lambda, \quad (\text{A22})$$

for all  $\lambda \geq \lambda^0(\epsilon)$ .

Using Lemma B.1 we have that on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$  and for all  $t \leq \tau^\lambda$ ,

$$\begin{aligned} \lambda - \mu_s Z_1^\lambda(t) - \mu_{cs} Z_2^\lambda(t) &= \mu_s \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^- - \mu_s \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^+ - \mu_{cs} \left( \frac{1+\beta}{\mu_s} - Z_1^\lambda(t) \right) \\ &\leq \mu_s \epsilon - \underline{\mu} \left( \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^+ + \frac{1+\beta}{\mu_s} - Z_1^\lambda(t) \right) \\ &\leq \mu_s \epsilon \lambda - \underline{\mu} \frac{\beta}{\mu_s} \lambda, \end{aligned} \quad (\text{A23})$$

where  $\underline{\mu} = \mu_s \wedge \mu_{cs}$ . Hence,

$$Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \left( \delta \lambda + \mu_s \epsilon \lambda - \underline{\mu} \frac{\beta}{\mu_s} \lambda \right) (t \wedge \tau^\lambda) + \Theta/2. \quad (\text{A24})$$

Let  $\eta := -(\mu_s \epsilon + \delta - \underline{\mu} \frac{\beta}{\mu_s})$ , choose  $\epsilon$  and  $\delta$  small enough so that  $\eta > 0$  and let  $t^* := 2\Theta/\eta$ . Then, as  $\tau^\lambda$  is the first time that  $Q^\lambda(t)$  goes below  $Q^\lambda(0) - 3\Theta/2$  we must have that  $\tau^\lambda \leq t^*/\lambda$  on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$ .

The arguments that lead to equation (A24) can be modified to show that, on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$  the queue will remain below  $Q^\lambda(0) - \Theta$  after it reaches  $Q^\lambda(0) - 3\Theta/2$  for the first time. That is,

that  $Q^\lambda(t) \leq Q^\lambda(0) - \Theta$  for all  $t^*/\lambda \leq t \leq T/\lambda$ . In particular, on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta/2)$ ,

$$\sup_{\xi \in \mathcal{A}_\epsilon^\lambda: q(\xi) > 2\Theta} \frac{Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2}{q(\xi)} \leq \sup_{\xi \in \mathcal{A}_\epsilon^\lambda: q(\xi) > 2\Theta} \frac{(q(\xi) - \Theta)^2 - q(\xi)^2}{q(\xi)} \leq -\Theta, \quad (\text{A25})$$

where  $\mathcal{A}_\epsilon^\lambda$  is as in (A19). Since  $Q^\lambda(t) \leq Q^\lambda(0) + A^\lambda(t)$ , it is also the case that

$$(Q^\lambda(t))^2 - (Q^\lambda(0))^2 \leq 2Q^\lambda(0)A^\lambda(t) + (A^\lambda(t))^2, \quad (\text{A26})$$

so that, removing the condition that  $\xi \in \mathcal{A}_\epsilon^\lambda$ , we have that

$$\sup_{q(\xi) > 2\Theta} \frac{E_\xi[Q^\lambda(2t^*/\lambda)^2] - q(\xi)^2}{q(\xi)} \leq -\Theta + 2E \left[ (\Theta + 2A^\lambda(2t^*/\lambda) + (A^\lambda(2t^*/\lambda))^2) 1_{(\Omega^*(\delta, \lambda, T/\lambda, \Theta/2))^c} \right]. \quad (\text{A27})$$

In Lemma F.2 we show that  $P \{(\Omega^*(\delta, \lambda, T/\lambda, \Theta/2))^c\} \leq c_5 e^{-c_6(\Theta/2 - \Gamma)/\log(2\nu cT)}$ , for some positive constants  $c_5$  and  $c_6$  and all  $\lambda$  large enough. We note that  $E[2A^\lambda(2t^*/\lambda) + (A^\lambda(2t^*/\lambda))^2] \leq c_7$  for some constant  $c_7$  and all  $\lambda$ . Then, applying the Cauchy-Schwartz inequality, and re-choosing  $\Theta$  large enough, we have that

$$\sup_{\xi \in \mathcal{A}_\epsilon^\lambda: q(\xi) > 2\Theta} \frac{E_\xi[Q^\lambda(2t^*/\lambda)^2] - q(\xi)^2}{q(\xi)} \leq -\frac{\Theta}{2}. \quad (\text{A28})$$

The crude inequality (A26) also guarantees that

$$\sup_{\xi \in \mathcal{A}_\epsilon^\lambda: q(\xi) \leq 2\Theta} E[Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2] \leq c_{10}, \quad (\text{A29})$$

where  $c_{10} := 4Kc_8 + c_9$  for some constants  $c_8, c_9$  that are independent of  $\lambda$  and  $\Theta$  so that some simple manipulations lead to

$$q(\xi)^2 - E_\xi[Q^\lambda(2t^*/\lambda)^2] \geq \frac{\Theta}{2}q(\xi) - c_{11} + \left( c_{11} - \frac{\Theta}{2}q(\xi) - E_\xi[Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2] \right) 1_{\{\xi \notin \mathcal{A}_\epsilon^\lambda\}}, \quad (\text{A30})$$

where  $c_{11} = \Theta^2 + c_{10}$ . By definition of stationarity we have that  $E_{\nu^\lambda}[Q^\lambda(0)^2] = E_{\nu^\lambda}[Q^\lambda(2t^*/\lambda)^2]$

where  $\nu^\lambda$  is the steady-state distribution of the process  $(\Xi^\lambda(t), t \geq 0)$ , and in particular,

$$0 = \int_{\xi \in \Xi^\lambda} (q(\xi)^2 - E_\xi[Q^\lambda(2t^*/\lambda)^2]) \nu^\lambda(d\xi).$$

When applied to (A30), this yields

$$E_{\nu^\lambda}[Q^\lambda(0)] \leq \frac{2c_{11}}{\Theta} + \frac{2}{\Theta} \left( E_{\nu^\lambda} \left[ \left( \frac{\Theta}{2} Q^\lambda(0) - c_{11} + E_{\Xi^\lambda(0)}[Q^\lambda(2t^*/\lambda)^2 - Q^\lambda(0)^2] \right) 1_{\{\Xi^\lambda(0) \notin \mathcal{A}_\epsilon^\lambda\}} \right] \right) \quad (\text{A31})$$

To establish a bound on  $E_{\nu^\lambda}[Q^\lambda(0)]$  we need to provide bounds for  $E[Q^\lambda(0)1_{\{\Xi^\lambda(0) \notin \mathcal{A}_\epsilon^\lambda\}}]$  which appears on the right hand side of (A31). The following two lemmas provide us with the necessary tools.

**Lemma B.2** *Under the conditions of Proposition B.1, there exists  $T > 0$  such that*

$$P_{\nu^\lambda} \{ \Xi^\lambda(0) \notin \mathcal{A}_\epsilon^\lambda \} \leq c_3 e^{-c_4 \lambda / \log(2\sqrt{c}\lambda T)}, \quad (\text{A32})$$

for all  $\lambda$  large enough.

**Lemma B.3** *Under the conditions of Theorem B.1*

$$\limsup_{\lambda \rightarrow \infty} E_{\nu^\lambda} \left[ \left( \frac{Q^\lambda(0)}{\lambda} \right)^m \right] < \infty,$$

for any integer  $m \geq 1$ .

The proof of these lemmas are postponed to §F and we now use them to complete the proof of the proposition. To that end, Using Lemmas B.2 and B.3 together with the Cauchy-Schwartz inequality, yields

$$\limsup_{\lambda \rightarrow \infty} E_{\nu^\lambda} \left[ (Q^\lambda(0))^m 1_{\{\xi \notin \mathcal{A}_\epsilon^\lambda\}} \right] = 0. \quad (\text{A33})$$

Applying (A33) and (A26) to (A31) we then have that  $E_{\nu^\lambda}[Q^\lambda(0)] \leq c_{12}$ , for some constant  $c_{12}$  and all  $\lambda$  large enough and, in particular, that

$$\limsup_{\lambda \rightarrow \infty} \frac{E_{\nu^\lambda}[Q^\lambda(0)]}{\sqrt{\lambda}} = 0.$$

This concludes the proof of Proposition B.1. ■

With the proof of Proposition B.1, we have established the first part of Theorem B.1—equation (A13). We turn now to prove (A14). First, we show that the number of idle servers does not exceed the negative part of the threshold. It applies to both cases  $\beta = 0$  and  $\beta > 0$  and will be used also in the proofs in §C and §D.

**Theorem B.2** *Consider a sequence of systems such that: (a) the  $\lambda^{\text{th}}$  system uses  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{\lambda})$  agents for some  $0 \leq \beta \leq \frac{\rho\mu_s}{\mu_{cs}}$ ,  $\max(\beta, \gamma) > 0$  and, (b) the  $\lambda^{\text{th}}$  system uses  $TP[K^\lambda]$  for a sequence  $\{K^\lambda\}_{\lambda \geq 0}$  that satisfies  $K^\lambda/\sqrt{R} \rightarrow \varrho \in (-\infty, \infty)$  as  $\lambda \rightarrow \infty$ . Then,*

$$\frac{E[((N^\lambda - Z^\lambda) - [K^\lambda]^-)^+]}{N^\lambda - R} \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \quad (\text{A34})$$

Corollary B.2 below is a special case of Theorem B.2. Indeed, under the conditions of Theorem B.1 we have  $\beta > 0$  and  $K^\lambda \geq 0$  for all  $\lambda$  so that  $[K^\lambda]^- = 0$ . Corollary B.2 proves (A14) and hence completes the proof of Theorem B.1.

**Corollary B.2** *Under the assumptions of Theorem B.1,*

$$\frac{E[I^\lambda]}{N^\lambda - R} \rightarrow 0 \text{ as } \lambda \rightarrow \infty. \quad (\text{A35})$$

**Proof of Theorem B.2:** Here we prove the theorem only for the case  $\beta > 0$ . The case  $\beta = 0$  is more involved and is proved in §F.

We initialize the  $\lambda^{\text{th}}$  system with  $\Xi^\lambda(0)$  distributed according to its stationary distribution  $\nu^\lambda$ . The process  $(\Xi^\lambda(t), t \geq 0)$  is then stationary. We will show that there exists  $\tilde{t} > 0$  such that  $E_{\nu^\lambda}[I^\lambda(\tilde{t})]/\lambda \rightarrow 0$  as  $\lambda \rightarrow \infty$ . Since, by stationarity,  $E_{\nu^\lambda}[I^\lambda(t)] = E_{\nu^\lambda}[I^\lambda(0)]$  for all  $t \geq 0$ , this will imply that  $E_{\nu^\lambda}[I^\lambda(0)]/\lambda \rightarrow 0$  as  $\lambda \rightarrow \infty$ . Finally, since we assumed that  $\beta > 0$  we have that  $N^\lambda - R > c\lambda$  for some  $c > 0$  and all  $\lambda$  large enough. Consequently, we will conclude that  $E_{\nu^\lambda}[I^\lambda(0)]/(N^\lambda - R) \rightarrow 0$  as  $\lambda \rightarrow \infty$ .

We now gradually fill-in the gaps in the above argument. First, we need some characterization of the fluid-level behavior of the system. Below, the processes  $T_i^\lambda(t)$ ,  $i = 1, \dots, 5$  are as defined in §A.

**Lemma B.4 (fluid Limits)** Consider a finite interval  $[0, T]$  and suppose that

$$\left( \frac{Q^\lambda(0)}{\lambda}, \frac{Z_1^\lambda(0)}{\lambda}, \frac{Z_2^\lambda(0)}{\lambda} \right) \Rightarrow (\bar{Q}(0), \bar{Z}_1(0), \bar{Z}_2(0)).$$

Then, under the assumptions of Theorem B.1, the sequence  $\left( \frac{Q^\lambda(t)}{\lambda}; \frac{Z_1^\lambda(t)}{\lambda}; \frac{Z_2^\lambda(t)}{\lambda}; \frac{T_i^\lambda(t)}{\lambda}, i = 1, \dots, 5 \right)$  is tight in  $D[0, T]$  and every subsequence  $\{\lambda^k\}_{k \geq 1}$  contains a further subsequence that converges in probability to some limit uniformly on compact sets. Moreover, any such limit process

$$(\bar{Q}(t); \bar{Z}_1(t); \bar{Z}_2(t); \bar{T}_i(t), i = 1, \dots, 5),$$

satisfies the following equations:

$$\bar{Z}_1(t) + \bar{Q}(t) = \bar{Q}(0) + \bar{Z}_1(0) + t - \int_0^t \mu_s \bar{Z}_1(u) du, \quad (\text{A36})$$

$$\bar{Z}_2(t) = \bar{Z}_2(0) - \mu_{cs} \int_0^t \bar{Z}_2(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du - \bar{T}_1(t), \quad (\text{A37})$$

$$\bar{Z}_1(t) = \bar{Z}_1(0) + \bar{T}_2(t) - \mu_s \int_0^t \bar{Z}_1(u) du + \bar{T}_3(t) + \bar{T}_4(t) - \bar{T}_5(t), \quad (\text{A38})$$

$$\dot{\bar{T}}_1(t) 1_{\{\bar{Q}(t) > 0\}} = p\mu_s \bar{Z}_1(t), \quad (\text{A39})$$

$$\dot{\bar{T}}_2(t) 1_{\{\bar{Z}_1 + \bar{Z}_2 < \frac{1+\beta}{\mu_s}\}} = 1, \quad (\text{A40})$$

$$\dot{\bar{T}}_3(t) 1_{\{\bar{Q}(t) > 0\}} = \mu_s \bar{Z}_1(t), \quad (\text{A41})$$

$$\dot{\bar{T}}_4(t) 1_{\{\bar{Q}(t) > 0\}} = \mu_{cs} \bar{Z}_2(t), \quad (\text{A42})$$

$$\dot{\bar{T}}_5(t) 1_{\{\bar{Q}(t) > 0\}} = 0. \quad (\text{A43})$$

**Lemma B.5** Fix  $\epsilon > 0$  and assume  $0 < \beta \leq \frac{p\mu_s}{\mu_{cs}}$ . Let

$$(\bar{Q}(t); \bar{Z}_1(t); \bar{Z}_2(t); \bar{T}_i(t), i = 1, \dots, 5),$$

be a non-negative process that satisfies equations (A36)-(A43). Then, there exists  $t^0(\epsilon)$  (independent

of  $\bar{Z}_1(0)$  and  $\bar{Z}_2(0)$ ), such that, for all  $t \geq t^0(\epsilon)$ ,

$$\left| \bar{Z}_1(t) - \frac{1}{\mu_s} \right| \leq \epsilon. \quad (\text{A44})$$

Moreover, there exists  $t^* \geq t^0(\epsilon)$ , such that

$$\bar{I}(t) := \frac{1 + \beta}{\mu_s} - \bar{Z}_1(t) - \bar{Z}_2(t) \leq \epsilon, \quad (\text{A45})$$

for all  $t \geq t^*$ .

Lemmas B.4 and B.5 are proved in §F. We now use them to complete the proof of Theorem B.2 under the assumption that  $\beta > 0$ . To this end, initialize the  $\lambda^{th}$  system according to its stationary distribution. Then, using Proposition B.1 and the fact that  $Z_1^\lambda + Z_2^\lambda \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_{cs}$ , we have that the sequence of steady-state random variables  $(Q^\lambda/\lambda, Z_1^\lambda/\lambda, Z_2^\lambda/\lambda)$  is tight and every limit point is of the form  $(0, \bar{Z}_1(0), \bar{Z}_2(0))$ . By Lemma B.4, the sequence of processes  $(Q^\lambda(t)/\lambda, Z_1^\lambda(t)/\lambda, Z_2^\lambda(t)/\lambda)$  is tight and every limit point  $(\bar{Q}(t), \bar{Z}_1(t), \bar{Z}_2(t))$  satisfies equations (A36)-(A43). We can thus apply Lemma B.5 to conclude the existence of  $t^*$  such that  $\bar{I}(t) \leq \epsilon$ , for all  $t \geq t^*$ . Since this holds for every limit point, we have that

$$\limsup_{\lambda \rightarrow \infty} P_{\nu^\lambda} \left\{ \frac{I^\lambda(t)}{\lambda} > 2\epsilon \right\} = 0. \quad (\text{A46})$$

Since  $I^\lambda(t) \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_{cs}$  we have that

$$\limsup_{\lambda \rightarrow \infty} \frac{E_{\nu^\lambda}[I^\lambda(t)]}{\lambda} \leq 3\epsilon, \quad (\text{A47})$$

for all  $t \geq t^*$ . Finally, since  $E_{\nu^\lambda}[I^\lambda(t)] = E_{\nu^\lambda}[I^\lambda(0)]$ , for all  $t \geq 0$ , we have that

$$\limsup_{\lambda \rightarrow \infty} \frac{E_{\nu^\lambda}[I^\lambda(0)]}{\lambda} \leq 3\epsilon. \quad (\text{A48})$$

Since  $\epsilon$  was arbitrary, this concludes the proof of the theorem for the case  $\beta > 0$ . The case  $\beta = 0$  is proved in §F. ■

## C. Asymptotic optimality under Condition 2

The asymptotic optimality result for this section is stated in the following theorem.

**Theorem C.1** *Suppose that Assumptions 3.1 and 3.2 hold. Also, assume that  $\mu_{cs} \geq \mu_s$ , and*

$$\liminf_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{\bar{N}_1^\lambda - R} \geq 1.$$

*Then, the following is asymptotically optimal in the sense of Definition 3.2:*

- **Staffing:** Staff with  $N_2^\lambda$  agents.
- **Control:** Use  $TP[0]$ .

**Proof:** Proposition C.1 below guarantees the asymptotic feasibility of pairs  $(N_2^\lambda, TP[0])$ . The asymptotic optimality argument is exactly the same as in the proof of Corollary B.1 using Theorem B.2. ■

**Proposition C.1** *Assume  $\mu_{cs} \geq \mu_s$ . Consider a sequence of systems such that: (a) the  $\lambda^{th}$  system uses  $TP[0]$  for control, and (b) the  $\lambda^{th}$  system uses  $N^\lambda$  agents so that*

$$\liminf_{\lambda \rightarrow \infty} \frac{N^\lambda - R}{\bar{N}_1^\lambda - R} \geq 1.$$

*Then,*

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1.$$

**Proof:** Recall that  $W_{\lambda, \mu_s}^{FCFS}(N)$  is the steady-state waiting time in an  $M/M/N$  system with service rate  $\lambda$  and service rate  $\mu_s$ . Also, let  $W^\lambda$  be the steady-state waiting time under  $TP[0]$  for a cross-selling system with  $N$  agents, arrival rate  $\lambda$ , service rate  $\mu_s$  and cross-selling rate  $\mu_{cs}$ . The following is a straightforward result.

**Lemma C.1** *Fix  $\lambda$ . Assume  $N^\lambda \geq R$ . If  $\mu_{cs} \geq \mu_s$  and  $TP[0]$  is used then,*

$$E[W^\lambda | W^\lambda > 0] \leq E[W_{\lambda, \mu_s}^{FCFS}(N) | W_{\lambda, \mu_s}^{FCFS}(N) > 0] = \frac{1}{N\mu_s - \lambda} \quad (\text{A49})$$

We continue with the proof of the proposition. Fix a sequence of staffing levels  $\{N^\lambda, \lambda \geq 0\}$  with

$$\liminf_{\lambda \rightarrow \infty} \frac{N^\lambda - R}{\bar{N}_1^\lambda - R} \geq 1.$$

Then, by Lemma C.1,

$$\frac{E[W^\lambda | W^\lambda > 0]}{\bar{W}^\lambda} \leq \frac{1}{\bar{W}^\lambda (N^\lambda \mu_s - \lambda)}. \quad (\text{A50})$$

By the definition of  $\bar{N}_1^\lambda$ ,  $\limsup_{\lambda \rightarrow \infty} 1/(\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)) \leq 1$ . Hence,

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda | W^\lambda > 0]}{\bar{W}^\lambda} \leq \limsup_{\lambda \rightarrow \infty} \frac{1}{\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)} \frac{\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)}{\bar{W}^\lambda (N^\lambda \mu_s - \lambda)} \leq 1, \quad (\text{A51})$$

Noting that  $E[W^\lambda] \leq E[W^\lambda | W^\lambda > 0]$ , the proof is complete ■

## D. Asymptotic optimality under Condition 3

Our optimality results under Condition 3 are closely related to the results of Gans and Zhou [4]. While [4] considers a system that is essentially different from the cross-selling system, we prove that in this asymptotic regime the two systems are, in some sense, equivalent. Specifically, we prove that the optimal solution in [4] constitutes an upper bound on the expected profit for the cross-selling model and that this upper bound is asymptotically achieved under the appropriate staffing and control.

To simplify the presentation of the results in which we use this asymptotic equivalence, we give first a brief description of the model considered in [4]: Consider a call center with two types of jobs: Type-H and Type-L. Type-H jobs arrive at rate  $\lambda_H$ , are processed at rate  $\mu_H$  and served FCFS within their class. A constraint of the form  $E[W] \leq \bar{W}$  limits the expected delay that these jobs may face. An infinite backlog of type-L jobs awaits processing at rate  $\mu_L$ . A pool of homogeneous servers process all jobs, and a system controller tries to maximize the rate at which type-L jobs are processed, subject to the service-level constraint for the type-H jobs. Given a fixed number of agents, the problem of finding the optimal control is formulated as a constrained average-cost Markov Decision Process (MDP) and the structure of effective routing policies is determined. When  $\mu_H = \mu_L$ , the suggested policies are globally optimal and have a very simple threshold structure. We refer to this model as the G&Z model.

To create a basis for comparison of the two models (Cross-Selling vs. G&Z) one may consider cross-selling transactions against processing of type-L jobs and service transactions against processing of type-H jobs. Clearly, the dynamics of the two models are different. In the cross-selling system, rather than having an infinite backlog of cross-selling “jobs”, these become available only upon a completion of a service “job”, and if they are not processed right away they disappear. Hence, the processing rate of type-L jobs in the G&Z model constitutes an *upper bound* on the cross-selling rate in the cross-selling model. We prove this formally in Lemma D.1.

These differences also illustrate the relative technical complexity of the cross-selling model. While in the G&Z model there is an infinite backlog of type-L jobs, the availability of cross-selling “jobs” is tightly related to the number of customers in the service phase in our model. The technical implication of this difference, is that any description of the system dynamics of the cross-selling system must be at least two-dimensional, regardless of whether  $\mu_s = \mu_{cs}$  or not. Our asymptotic analysis, however, allows us to reduce the dimensionality of the problem whenever  $\mu_s = \mu_{cs}$  and prove that, under  $TP$ , the upper bound, as given by the G&Z model, is asymptotically achieved.

The following is an adaptation of Definition 7 from [4].

**Definition D.1** Fix  $\lambda$ . A randomized threshold reservation policy with threshold  $K^\lambda$  and probability  $p^*$  acts as follows at each event epoch in which there are no type-H calls waiting to be served:

1. A type-H customer will enter service immediately upon arrival if there are any idle agents.
2. Upon service completion (of either a type-L or a type-H job):
  - If there are  $|K^\lambda|$  or fewer idle agents, the policy does nothing.
  - If there are  $|K^\lambda| + 1$  or more idle agents, then with probability  $1 - p^*$  the policy puts enough type-L jobs into service so that exactly  $|K^\lambda|$  agents are idle, and with probability  $p^*$  the policy puts enough type-L jobs into service so that exactly  $|K^\lambda| - 1$  agents are idle.

Note that, if one removes the randomization components from the above definition, the threshold reservation policy in the G&Z model can be thought of as the TP control adapted to the G&Z model. Denote by  $\overline{TP}^\lambda(N^\lambda, p^*)$  the randomized threshold policy of G&Z with threshold  $K^\lambda$  determined through (A52) and with a randomization probability  $p^*$ . The following is a version of the optimality result of [4] for the case  $\mu_s = \mu_{cs}$ . We only cite the parts of the Theorem that are relevant for our results.

**Theorem D.1 (Theorem 1 - Gans and Zhou:)** Consider a G&Z model with arrival rate  $\lambda$ , service rates  $\mu_H = \mu_L = \mu_s = \mu_{cs}$ ,  $N^\lambda$  agents and average delay bound  $\bar{W}^\lambda$ . Then, either

1. the problem is infeasible, or
2. A randomized threshold reservation policy with a threshold  $K^\lambda \leq 0$  and probability  $p^*$  is optimal, for some  $p^* \in [0, 1]$ .

Moreover, the optimal threshold  $K^\lambda$  is chosen so that

$$K^\lambda(N^\lambda) = \max \left\{ k \in [-N^\lambda, 0] \mid \frac{\xi_k(N^\lambda)}{N^\lambda \mu_s - \lambda} \leq \bar{W}^\lambda \right\}. \quad (\text{A52})$$

Here  $\xi_k(N^\lambda) = P\{Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) = N^\lambda \mid Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \geq N^\lambda + k\}$  and  $Z_{\lambda, \mu_s}^{FCFS}(N^\lambda)$  is the steady-state number of busy servers in an  $M/M/N^\lambda$  system with arrival rate  $\lambda$  and service rate  $\mu_s$ .

Given two random variables  $X$  and  $Y$ , we use the notation  $X \geq_{st} Y$  to denote that a random variable  $X$  is stochastically greater than  $Y$ . Let  $CS^\pi(t)$  be the cumulative cross-selling completions up to time  $t$  when the control  $\pi$  is used. Also, let  $TH^{\pi'}(t)$  be the cumulative completion of type-L jobs up to time  $t$  in the G&Z model when the control  $\pi'$  is used. Letting  $\bar{Z}^{\lambda, \pi'}$  be the steady-state number of busy agents in the G&Z model under the control  $\pi'$ , we have that the steady-state throughput rate of type-L jobs equals  $\mu_{cs}(E[\bar{Z}^{\lambda, \pi'}] - R)$ . This is a consequence of a straightforward application of Little's law as in §2 of the main paper.

**Lemma D.1** *Fix  $\lambda, \mu_s, \mu_{cs}, N$  and  $\bar{W}^\lambda$ . Let  $\pi_{g\&z}^*$  be the optimal control in the G&Z system with  $\mu_H = \mu_s$  and  $\mu_L = \mu_{cs}$ . Then, for any policy  $\pi \in \Pi(N)$  we have that*

$$TH^{\pi_{g\&z}^*}(t) \geq_{st} CS^\pi(t), \quad \forall t \geq 0. \quad (\text{A53})$$

*In particular, if  $\pi \in \Pi(N)$  admits a steady-state distribution for the cross-selling system, then*

$$\mu_{cs}(E[Z^{\lambda, \pi}] - R) \leq \mu_{cs}(E[\bar{Z}^{\lambda, \pi_{g\&z}^*}] - R). \quad (\text{A54})$$

**Proof:** Note that this lemma does not require that  $\mu_s = \mu_{cs}$ . The result follows a sample path coupling argument that is independent of the specific values of  $\mu_s, \mu_{cs}$  and  $\lambda$ . We will show that under our sample path construction the inequality (A53) holds a.s. This, in turn, implies the stochastic ordering in (A53).

We construct the coupled sample paths as follows: fix a common sample path of arrivals, service times and cross-selling times for both systems. Specifically, let  $\{t_n\}_{n=1}^\infty, \{s_n\}_{n=1}^\infty$ , and  $\{c_n\}_{n=1}^\infty$  be, respectively, the sequence of arrival times, service times and potential cross-selling times (that is, if cross-selling is exercised on customer  $n$ , his cross-selling time will be  $c_n$ ). Then, our sample path construction uses the same sequences,  $\{t_n\}, \{s_n\}$  and  $\{c_n\}$  for both systems.

For simplicity of notation label the cross-selling system by 1 and the G&Z system by 2. Fix a scheduling policy  $\pi_1$  for system 1 and use the same scheduling policy for system 2. This is clearly possible because whenever system 1 can schedule a customer to cross-sell system 2 can schedule a type-L job to service. It is now straightforward to show by induction on the event epochs (arrival, or service completion of any type) that both systems will have exactly the same sample paths, and

we would have that pathwise

$$TH_{\pi_1}(t) = CS_{\pi_1}(t), \forall t \geq 0, \quad (\text{A55})$$

and

$$\mu_{cs}(E[\bar{Z}^{\lambda, \pi_1}] - R) = \mu_{cs}(E[Z^{\lambda, \pi_1}] - R). \quad (\text{A56})$$

Since we fixed the scheduling policy for system 1 we have, in particular, that

$$\mu_{cs}(E[\bar{Z}^{\lambda, \pi_{g\&z}^*}] - R) \geq \sup_{\pi_1 \in \Pi(N)} \mu_{cs}(E[Z^{\lambda, \pi_1}] - R),$$

for any policy  $\pi_1$  that admits a steady-state distribution for the cross-selling system.  $\blacksquare$

For future reference let  $\bar{V}(N^\lambda) = r\mu_{cs}(E[\bar{Z}^{\lambda, \pi_{g\&z}^*}] - R) - (C^\lambda(N^\lambda) - C^\lambda(R))$ , so that  $\bar{V}(N^\lambda)$  is the optimal throughput rate in the G&Z model with  $N^\lambda$  agents. Now, let  $\{N^\lambda, \lambda \geq 0\}$  be a sequence with  $N^\lambda \geq N_1^\lambda$  for all  $\lambda$  and such that

$$\frac{N^\lambda - R}{\sqrt{R}} \rightarrow \hat{\gamma} > 0 \text{ as } \lambda \rightarrow \infty. \quad (\text{A57})$$

The existence of such a sequence is guaranteed since, by §9 of [2], we have that

$$\frac{N_1^\lambda - R}{\sqrt{R}} \rightarrow \gamma, \quad (\text{A58})$$

for some  $\gamma > 0$ . Let  $\bar{Y}^{\lambda, p^*}$  be the steady-state overall number of customers in a G&Z system with  $N^\lambda$  agents and using the control  $\bar{TP}^\lambda(N^\lambda, p^*)$ . Also, let  $Y^\lambda$  be the steady-state overall number of customers in a cross-selling system with  $N^\lambda$  agents and using  $TP[K^\lambda]$  with  $K^\lambda$  determined through (A52). Accordingly, we let  $\bar{Z}^\lambda$  and  $Z^\lambda$  be the number of busy agents in the above two systems. Define the scaled variables

$$\bar{X}^{\lambda, p^*} = \frac{\bar{Y}^{\lambda, p^*} - N^\lambda}{N^\lambda - R}, \text{ and } X^\lambda = \frac{Y^\lambda - N^\lambda}{N^\lambda - R}. \quad (\text{A59})$$

For the following result, let  $D := D[0, \infty)$  be the space right continuous processes with left limits endowed with the  $J_1$  Skorohod topology. We say that a sequence of processes  $x^\lambda(\cdot) \Rightarrow x(\cdot)$  in  $D_-$  if the convergence holds in  $D[s, T]$  for each  $0 < s < T < \infty$ . We let  $\bar{Y}^{\lambda, p^*}(t), t \geq 0$  be the process representing the overall number of customers in a G&Z system with  $N^\lambda$  agents and using

the control  $\overline{TP}^\lambda(N^\lambda, p^*)$ . Also, let  $(Y^\lambda(t), t \geq 0)$  be the process representing the overall number of customers in a cross-selling system with  $N^\lambda$  agents and using  $TP[K^\lambda]$  with  $K^\lambda$  determined through (A52). Define the scaled processes

$$\bar{X}^{\lambda, p^*}(t) = \frac{\bar{Y}^{\lambda, p^*}(t) - N^\lambda}{N^\lambda - R}, \text{ and } X^\lambda(t) = \frac{Y^\lambda(t) - N^\lambda}{N^\lambda - R}. \quad (\text{A60})$$

Proposition D.1 below is the main component in the proof of asymptotic optimality of our proposed solution under Condition 2. The proposition shows that the G&Z system operated under their optimal threshold reservation policy is asymptotically equivalent to the cross-selling system operated with the TP rule. Since we established that the G&Z throughput serves as an upper bound for the cross-selling rate, the asymptotic equivalence will allow us to prove that the upper bound is achieved.

**Proposition D.1 (Diffusion Limits:)** *Suppose that  $\{N^\lambda, \lambda \geq 0\}$  is a sequence that satisfies (A57). If, in addition,*

$$\bar{X}^{\lambda, p^*}(0) \Rightarrow \bar{X}(0), \text{ and } X^\lambda(0) \Rightarrow \bar{X}(0) \text{ as } \lambda \rightarrow \infty, \quad (\text{A61})$$

then,

$$\bar{X}^{\lambda, p^*}(\cdot) \Rightarrow \bar{X}(\cdot) \text{ in } D_- \text{ as } \lambda \rightarrow \infty, \quad (\text{A62})$$

and

$$X^\lambda(\cdot) \Rightarrow \bar{X}(\cdot) \text{ in } D_- \text{ as } \lambda \rightarrow \infty. \quad (\text{A63})$$

Here,  $\bar{X}(\cdot)$  is a diffusion process with infinitesimal drift function

$$m(x) = \begin{cases} -\beta\mu_s, & x \geq 0 \\ -(\beta + x)\mu_s, & -\delta \leq x \leq 0 \end{cases} \quad (\text{A64})$$

and infinitesimal variance term  $\sigma^2 = 2\mu_s$  and  $\delta := \delta(\hat{\gamma}, \hat{W})$  is the limit from Lemma D.3.

We prove proposition D.1 at the end of this section and proceed towards the proof of asymptotic optimality. Since our profits are measured in terms of steady-state performance we first have the following corollary by which the convergence on compact intervals can be extended to convergence of the corresponding steady-state variables.

**Corollary D.1** Assume that  $\mu_s = \mu_{cs}$  and that  $\{N^\lambda, \lambda \geq 0\}$  is a sequence that satisfies (A57). Then, for any  $p^* \in [0, 1]$ ,

$$\bar{X}^{\lambda, p^*} \Rightarrow \bar{X}, \text{ as } \lambda \rightarrow \infty, \quad (\text{A65})$$

and

$$X^\lambda \Rightarrow \bar{X}, \text{ as } \lambda \rightarrow \infty, \quad (\text{A66})$$

where  $\bar{X}$  has the steady-state distribution of the diffusion process  $\bar{X}(\cdot)$  in Proposition D.1. Furthermore, the convergence in (A66) also holds in expectation.

The proof of Corollary D.1 is given in §F. Note that under  $\overline{TP}(N^\lambda, 0)$  (i.e. when setting  $p^* = 0$ ), the steady-state number of busy agents in the G&Z system, denoted by  $E[\bar{Z}^\lambda]$ , satisfies

$$E[\bar{Z}^\lambda] = E[Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \mid Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \geq N^\lambda + K^\lambda]. \quad (\text{A67})$$

Since the limit in Corollary D.1 is independent of the precise value  $p^*$ , the fact that  $\bar{Z}^\lambda = N^\lambda - (\bar{Y}^{\lambda, p^*} - N^\lambda)^-$  implies that

$$\frac{E[Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) - R \mid Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) \geq N^{*\lambda} + K^\lambda] - (E[\bar{Z}^\lambda] - R)}{N^\lambda - R} \rightarrow 0 \text{ as } \lambda \rightarrow \infty, \quad (\text{A68})$$

provided that the sequence of staffing levels  $\{N^\lambda, \lambda \geq 0\}$  satisfies (A57). This observation is formalized in the following Corollary.

**Corollary D.2** Assume  $\mu_s = \mu_{cs}$  and consider a sequence of cross-selling systems such that (a) the  $\lambda^{\text{th}}$  system uses  $N^\lambda$  agents with  $\{N^\lambda, \lambda \geq 0\}$  satisfying equation (A57), and (b) the  $\lambda^{\text{th}}$  system uses  $TP[K^{*\lambda}]$  for control with  $K^{*\lambda}$  determined through equation (A52). Then,

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1, \quad (\text{A69})$$

and

$$\frac{V^\lambda(N^\lambda, TP[K^\lambda])}{\bar{V}^\lambda(N^\lambda)} \rightarrow 1, \text{ as } \lambda \rightarrow \infty. \quad (\text{A70})$$

The following lemma will help us to translate the result of Corollary D.2 to the more general asymptotic optimality result that we need.

**Lemma D.2** Assume  $\mu_s = \mu_{cs}$  in addition to Assumptions 3.1 and 3.2. Let  $N^{*\lambda}$  and  $K^{*\lambda}$  be determined through (13) and (14) and assume that  $\limsup_{\lambda \rightarrow \infty} (N_2^\lambda - R)/(\bar{N}_1^\lambda - R) < 1$ . Then,

$$\liminf_{\lambda \rightarrow \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} > 0, \quad (\text{A71})$$

and

$$\limsup_{\lambda \rightarrow \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} < \infty. \quad (\text{A72})$$

Lemma D.2 then shows that the sequence of staffing recommendations  $N^{*\lambda}$  given by (13) satisfies that the sequence  $(N^{*\lambda} - R)/\sqrt{R}$  is bounded. Consequently, it has convergent subsequences. We can then apply Corollary D.2 to any convergent subsequence to show that  $TP[K^{*\lambda}]$  will outperform asymptotically any other routing rule over this subsequence. This reasoning can be applied to any convergent subsequence to conclude the asymptotic optimality of  $(N^{*\lambda}, TP[K^{*\lambda}])$ , as stated in the following corollary which concludes the proof of asymptotic optimality for this section.

**Corollary D.3** Assume that  $\mu_s = \mu_{cs}$  in addition to Assumptions 3.1 and 3.2. Also, assume that

$$\limsup_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{\bar{N}_1^\lambda - R} < 1.$$

Then, the following is asymptotically optimal for the cross-selling system in the sense of Definition 3.2:

- **Staffing:** Staff with  $N^{*\lambda}$  agents where  $N^{*\lambda}$  is given by equations (13) and (14).
- **Control:** Use  $TP[K^\lambda(N^{*\lambda})]$  where  $K^\lambda(N^{*\lambda})$  is given by equation (14).

We end this section with the proof of the diffusion-limits result.

**Proof of Proposition D.1:** We begin with the sequence of processes  $\{X^\lambda(\cdot)\}$ . We will first prove the convergence of a certain Birth and Death (B&D) process. We will then show, via coupling, that this B&D process is asymptotically equivalent to  $X^\lambda(\cdot)$ . This will allow us to apply the convergence together theorem to conclude the convergence of  $\{X^\lambda(\cdot)\}$ .

To this end, consider the B&D process with birth rates rates:  $\hat{\lambda}_i = \lambda$  for all  $i$ , where  $i$  is the number of customers in the system, and

$$\hat{\mu}_i^\lambda = \begin{cases} (N^\lambda + K^\lambda + i)\mu_s & 1 \leq i \leq -K^\lambda - 1 \\ N^\lambda\mu_s & i \geq K^\lambda \\ 0 & \text{otherwise} \end{cases}, \quad (\text{A73})$$

where one should recall that  $K^\lambda$  is negative in this setting. We denote this process by  $\hat{Y}^\lambda(\cdot)$  and define its scaled version by  $\hat{X}^\lambda(\cdot) = \frac{\hat{Y}^\lambda(\cdot) + K^\lambda}{N^\lambda - R}$ . We now have the following lemma:

**Lemma D.3** *Assume that  $\mu_s = \mu_{cs}$ ,  $N^\lambda$  satisfies (A57) and  $K^\lambda$  is defined through (A52). Then, there exists a function  $\delta(\cdot, \cdot)$  such that for all  $(\gamma, \hat{W})$ ,*

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow \delta(\hat{\gamma}, \hat{W}) < \infty \text{ as } \lambda \rightarrow \infty. \quad (\text{A74})$$

Lemma D.3 is proved in §F. We apply it now in the proof of the proposition. To this end, initializing the  $\lambda^{\text{th}}$  B&D process at state  $\hat{X}^\lambda(0) \vee (-K^\lambda)$  and using the convergence of  $K^\lambda/\sqrt{R}$  in Lemma D.3, the sequence  $\hat{X}^\lambda(\cdot)$ , converges weakly to  $\bar{X}(\cdot)$  with the diffusion parameters given in equation (A64); see for example [8]. To complete the proof we will show that for all  $T > 0$

$$d^T(\hat{X}^\lambda, X^\lambda) \xrightarrow{P} 0, \text{ as } \lambda \rightarrow \infty, \quad (\text{A75})$$

where  $d^T(\cdot, \cdot) = \sup_{0 < t \leq T} \|\hat{X}^\lambda(t) - X^\lambda(t)\|$ . The convergence of  $\{X^\lambda(\cdot)\}$  will then follow from the convergence together theorem (see e.g. Theorem 11.4.7 of [10]).

In order to evaluate  $d^T(\hat{X}^\lambda, X^\lambda)$ , we use a coupling argument and deduce that

$$d^T(\hat{X}^\lambda, X^\lambda) \leq \sup_{0 < t \leq T} \frac{[Z^\lambda(t) - (N^\lambda + K^\lambda)]^-}{\sqrt{N^\lambda - R}}. \quad (\text{A76})$$

In Remark F.1 we show that the right-hand side above converges to 0 in probability. Hence, to complete the proof, it only remains to provide the coupling argument between the cross-selling system and the B&D process that we constructed.

To this end, fix  $\lambda$  (and omit it from the notation). Initialize the cross-selling system with all agents busy and no customer in queue and we initialize the B&D process with  $-K$  customers in system. We generate arrivals from the same Poisson process. We generate the departures from the same Poisson process with thinning. Let  $\hat{Y}^\lambda(t)$  be the value of the state dependent  $M/M/1$  process at time  $t$ .  $Y^\lambda(t)$ , as before, is the number of customers in the cross-selling system at time  $t$ . We prove by induction that

- $\hat{Y}(t) \geq Y(t) - (N + K)$ , for all  $t \geq 0$ .
- $\hat{Y}(t) - (Y(t) - (N + K)) \leq \sup_{0 \leq s \leq t} [Z(s) - (N + K)]^-$ , for all  $t \geq 0$

By our initial conditions the assumption holds at the first departure from the system. Assume that it holds for the first  $n - 1$  departures and consider the  $n^{\text{th}}$ , let the time of this departure be  $t_n$ . By our inductive assumption the  $n^{\text{th}}$  departure will be a departure in both systems if  $\hat{Y}(t_n -) = Y(t_n -) - (N + K)$  while preserving the ordering. It will be a departure in the  $M/M/1$  system, and not in the cross-selling system, only if  $\hat{Y}(t_n -) > 0$  and  $\hat{Y}(t_n -) > Y(t_n -) - (N + K)$  thus preserving the ordering. It will be a departure in the cross-selling system and not in the  $M/M/1$  queue only if  $0 = \hat{Y}(t_n -) > Y(t_n -) - (N + K)$  again preserving the ordering. Also, whenever  $\hat{Y}(t_n -) > Y(t_n -) - (N + K) > 0$  the difference between the two processes cannot increase, since every departure will necessarily be a departure in the  $M/M/1$  system. The difference can only increase when  $0 = \hat{Y}(t_n -) > Y(t_n -) - (N + K)$ , in which case  $\hat{Y}(t_n) - (Y(t_n) - (N + K)) = [Y(t_n) - N + K]^- = [Z(t_n) - N + K]^-$ , where the last equality follows from the fact  $Y(t) = Z(t)$  whenever  $Y(t) \leq N$ . Thus, the second part of the inductive assumption is preserved. Note that the result would still hold as long as  $\hat{Y}(0) = (Y(0) - (N - K))^+$ .

The proof for the sequence  $\bar{X}^{\lambda, p^*}(\cdot)$  is much simpler. It is trivial to show through a coupling argument that, for any  $p^* \in [0, 1]$ , one can construct the sample path of the processes  $\bar{X}^{\lambda, 0}(\cdot)$ ,  $\bar{X}^{\lambda, p^*}(\cdot)$ ,  $\bar{X}^{\lambda, 1}(\cdot)$ , so that

$$\bar{X}^{\lambda, 0}(t) \leq \bar{X}^{\lambda, p}(t) \leq \bar{X}^{\lambda, 1}(t), \forall t \geq 0.$$

Note that for  $p^* = 0$  the overall number of customers in the G&Z system has exactly the same law as the state dependent  $M/M/1$  defined through equation (A73) above. For  $p^* = 1$  the same holds with  $K^\lambda$  replaced with  $K^\lambda + 1$ . But, by [8] the scaled versions of these two  $M/M/1$  systems will have the same limit  $\bar{X}(\cdot)$ . The proof is completed by applying the convergence together theorem. ■

## E. Proofs for §4

### E.1 Proof of Lemma 4.1

The argument is straightforward and we only provide a sketch of the proof. For any policy  $\pi \in \Pi(N)$  we construct the corresponding sample paths as follows: We generate arrivals from a Poisson stream. In addition, we generate an infinite sequence of service times  $\{s_i\}_{i \geq 1}$  and cross-selling times  $\{c_i\}_{i \geq 1}$ . When constructing the sample paths, the service  $s_i$  will be assigned to the  $i^{\text{th}}$  customer to begin service and the cross-selling time  $c_i$  will be assigned to the  $i^{\text{th}}$  customer to begin cross-selling. Under this construction the process  $(Z_2(t), Y(t))$  is invariant to the order in which customers are admitted from the queue. In particular,  $\pi'$  which is obtained from  $\pi$  by admitting customers to service in a FCFS manner induces the same sample path under this construction. This invariance guarantees (through Little's Law) that if  $\pi$  is feasible so will be  $\pi'$ . Moreover, both controls will admit the same cross-selling rate since the steady-state number of customers in cross-selling,  $Z_2$ , has the same distribution probability law under both  $\pi$  and  $\pi'$ . ■

### E.2 Proof of Lemma 4.2

We use a coupling argument to prove this assertion. Consider two cross-selling systems with the same number of agents,  $N$ , in both systems. Let system 1 be the system that uses the policy  $\pi$  and system 2 be the system that uses  $\pi'$ . The latter is the work conserving system. We assume that both systems are initialized empty and we let  $\{t_i\}_{i \geq 1}$  and  $\{s_i\}_{i \geq 1}$  and  $\{c_i\}_{i \geq 1}$  be, respectively, the sequences of arrival times, service times and cross-selling times in system 1. Specifically, customer  $i$  arrives at time  $t_i$  and requires a service time of  $s_i$ . If cross-selling is exercised at customer  $i$  the cross-selling will require  $c_i$  units of time. If cross-selling is not exercised on customer  $i$  we set  $c_i = 0$ .

We construct the sample path of system 2 from system 1 as follows: we use the same stream of arrivals, services and cross-selling times. We cross-sell to customer  $i$  in system 2 if and only if we cross-sell to him in system 1. To differ from system 1, upon service completion of customer  $i$ , if cross-selling is not exercised, a customer from the queue will be admitted to service (unless the queue is empty). Let  $b_i^j, j = 1, 2$ , be the time at which customer  $i$  begins service in system  $j$  (In particular, the waiting time of customer  $i$  in system  $j$  is given by  $t_i - b_i^j$ ).

Let  $Q^j(t), j = 1, 2$ , be the queue length at time  $t$  in system  $j$ . Also, let  $CS^j(t)$  be the number

of customers that left system  $j$  up to time  $t$  after cross-selling was exercised on them. In order to prove that the assertion of the lemma holds it suffices to show the following:

1.  $Q^1(t) \geq Q^2(t)$ , for all  $t \geq 0$ .
2.  $CS^1(t) \leq CS^2(t)$ .

Indeed, if  $Q^1(t) \geq Q^2(t)$ , the assumed feasibility of system 1 will imply the feasibility of system 2. Moreover, if  $CS^1(t) \leq CS^2(t)$ , then system 2 performs at least as well as system 1 in terms of cross-selling rate. Since we exercise cross-selling on customer  $i$  in system 2 only if we exercise cross-selling on this customer in system 1, it suffices to prove that for all  $i \geq 1$ ,  $b_i^2 \leq b_i^1$ . That is, in system 2 all the customers begin service earlier.

We will now proceed by induction on the customer number to prove that indeed  $\forall i \geq 1$ ,  $b_i^2 \leq b_i^1$ . The conditions clearly holds for the first customer since both systems are initialized empty. Assume the condition holds up to customer  $n - 1$  and consider customer  $n$ . Specifically consider the following cases:

- If at time  $t_n$  there are idle agents in system 2 the customer will be admitted to service immediately upon arrival (in system 2) and the inductive assumption will be kept.
- Otherwise, let us consider the time  $b_{n-1}^2$  at which customer  $n - 1$  will begin service in system 2 (while he might still be waiting for service in system 1). Let  $r_i^j$  be the remaining handling time of customer  $i \leq n - 1$  in system  $j$  at time  $b_{n-1}^2$ . That is,

$$r_i^j = [s_i + c_i - [b_{n-1}^2 - b_i^j]^+]^+. \quad (\text{A77})$$

In particular, by our inductive assumption  $r_i^2 \leq r_i^1$ ,  $\forall i \leq n - 1$ , and by work conservation and the fact that customer  $n$  had to wait in queue, we have that  $t_n < b_{n-1}^2 + \min_{i \leq n-1} \{r_i^2 | r_i^2 > 0\}$  and  $b_n^2 = b_{n-1}^2 + \min_{i \leq n-1} \{r_i^2 | r_i^2 > 0\}$ . If we can show that for system 1  $b_n^1 \geq b_{n-1}^2 + \min_{i \leq n-1} \{r_i^1 | r_i^1 > 0\}$ , then we are done.

To see that this indeed the case note that since  $b_i^2 \leq b_i^1$  for all  $i \leq n - 1$  and since the handling times are common for both system, we have that at time  $b_{n-1}^2$  the overall number of customers in system 1 is at least as large as the overall number of customers in system 2.

Now, recall that we assumed that all agents are busy in system 2 at time  $t_n -$ , this implies that on the interval  $[b_{n-1}^2, t_n)$  all agents are busy (otherwise customer  $n$  would not have to wait by

work conservation). Hence, at time  $b_{n-1}^2$  the number of customers in system 2 (and then in both systems) will be at least  $N$ . For system 1 this implies that the number of idle agents at time  $b_{n-1}^2$  is smaller than the queue length. Formally, if  $Z^1(t)$  is the number of busy agents in system 1 at time  $t$ , then we just argued that  $Z^1(b_{n-1}^2) + Q^1(b_{n-1}^2) \geq N$ , and in particular  $I^1(b_{n-1}^2) \leq Q^1(b_{n-1}^2)$ , where  $I^1(t)$  is the number of idle agents in system 1 at time  $t$ . Hence, even if at time  $b_{n-1}^2$  system 1 admits all waiting customers to service, by the assumption that  $t_n < b_{n-1}^2 + \min_{i \leq n-1} \{r_i^2 | r_i^2 > 0\} \leq b_{n-1}^2 + \min_{i \leq n-1} \{r_i^1 | r_i^1 > 0\}$ , customer  $n$  must find either a non-empty queue or an empty queue but with all agents busy. If he finds an empty queue with all agents busy he will enter at time  $b_{n-1}^2 + \min_{i \leq n-1} \{r_i^1 | r_i^1 > 0\}$ , otherwise he will have to wait more. In any case we have that  $b_n^1 \geq b_{n-1}^2 + \min_{i \leq n-1} \{r_i^1 | r_i^1 > 0\}$ . ■

### E.3 Proof of Proposition 4.1

Since we fix  $\lambda$  we omit the superscript from the all the notation. Let  $\pi^\infty$  be any feasible policy for the original cross-selling problem which exists by our assumption that  $N \geq N_1$ . Define  $\pi^L$  to be an adaptation of  $\pi^\infty$  to a system where there is a limited number of trunk lines,  $L$ . The adaptation of  $\pi^\infty$  to the system with finite buffer is straightforward. Note that  $\pi^\infty(i, j)$  defines what action to take in an event epoch when the system is in state  $i, j$ . Then, we take  $\pi^L(i, j) = \pi^\infty(i, j), \forall i, j : j \leq L, z_2 \leq N$ . Also, from any feasible policy,  $\pi^L$ , in the finite buffer the system we can construct a corresponding policy,  $\pi_L^\infty$  for the infinite buffer system by setting  $\pi_L^\infty(i, j) = 0, \forall i, j : j > L$ .

To establish the result of the proposition, it suffices to show that: (a) starting from a work conserving policy  $\pi^\infty$ , the sequence that we construct  $\pi^L$  (which is also work conserving and hence within the set of possible solutions for the LP), achieves asymptotically the same value, as  $L \rightarrow \infty$ , and (b) that starting from a sequence of policies  $\{\pi^L\}$ , the sequence of adapted policies for the infinite buffer system,  $\{\pi_L^\infty\}$ , achieves asymptotically the same value for the infinite buffer system. Formally, we want to show that, given  $\epsilon > 0$ ,

$$|\tilde{V}(N, \pi^\infty) - \hat{V}(N, L, \pi^L)| \leq \epsilon, \quad (\text{A78})$$

and

$$|\tilde{V}(N, \pi_L^\infty) - \hat{V}(N, L, \pi^L)| \leq \epsilon., \quad (\text{A79})$$

for all  $L$  large enough. Here,  $\tilde{V}(N, \pi^\infty)$  and  $\hat{V}(N, L, \pi^L)$  are, respectively, the cross-selling rates in the infinite and finite buffer systems, equipped with  $\pi^\infty$  and  $\pi^L$ ,  $N$  agents and  $L$  trunk lines (in the finite buffer system). Then, by definition  $V^*_{LP}(N, L) = \sup_{\pi^L} \hat{V}(N, L, \pi^L)$  and  $V^*(N) = \sup_{\pi} \tilde{V}(N, \pi)$ , where the supremum is taken over feasible policies for each system. Recalling that the cross-selling rate under any policy  $\pi'$  equals  $\mu_{cs}E[Z_2^{\pi'}]$ , in order to prove (A78) it suffices to show that  $E[Z_2^{\pi^\infty}] = \lim_{L \rightarrow \infty} E[Z_{2B}^{L, \pi^L}]$ , where  $Z_{2B}^{L, \pi^L}$  is the steady-state number of agents busy cross-selling in the finite buffer system with  $L$  trunk lines and using a control  $\pi^L$ .

We first fix a feasible policy  $\pi^\infty$  for the infinite buffer system and prove (A78). Consider the truncation of the resulting Markov chain to the subspace of the domain in which  $\{j \leq L\}$ . Then, the restricted Markov chain has the same law as the finite buffer system with  $\pi^L$ . Hence,

$$E[Z_2^{\pi^\infty}] = E[Z_{2B}^{L, \pi^L}] P\{Y^\pi \leq L\} + E[Z_2^{\pi^\infty} 1_{\{Y^{\pi^\infty} > L\}}]. \quad (\text{A80})$$

The feasibility of  $\pi^\infty$  implies that  $E[Q^\pi] \leq \lambda \bar{W}$ . Using Markov's inequality we have

$$P\{Y^\pi > L\} = P\{Q^\pi > L - N\} \leq \frac{\lambda \bar{W}}{L - N} \quad (\text{A81})$$

Using the Cauchy-Schwartz inequality we have that

$$E[Z_2^{\pi^\infty} 1_{\{Y^{\pi^\infty} > L\}}] \leq \sqrt{E[(Z_2^{\pi^\infty})^2]} P\{Y^{\pi^\infty} > L\}. \quad (\text{A82})$$

Since, by definition,  $Z_2^{\pi^\infty} \leq N$ , we then have

$$E[Z_2^{\pi^\infty} 1_{\{Y^{\pi^\infty} > L\}}] \rightarrow 0, \text{ as } L \rightarrow \infty. \quad (\text{A83})$$

Plugging (A81) and (A83) back into equation (A80) we have that

$$E[Z_2^{\pi^\infty}] = \lim_{L \rightarrow \infty} E[Z_{2B}^{L, \pi^L}]. \quad (\text{A84})$$

This completes the proof (A78). The proof of (A79) is very similar and is omitted. ■

## F. Proofs of auxiliary results

For the following recall that  $\Xi^\lambda(t) := (Q^\lambda(t), Z_2^\lambda(t), Z_1^\lambda(t))$ .

**Lemma F.1** *Fix  $\lambda$ . Assume the system is staffed with  $N > R$  agents and that  $TP[K]$  is used for control with some  $K \geq -N$ . Then, the Markov process  $(\Xi^\lambda(t), t \geq 0)$  admits a steady-state distribution. Consequently,  $TP$  is admissible in the sense that*

$$\lim_{t \rightarrow \infty} \frac{E[Q^\lambda(t)]}{t} = 0. \quad (\text{A85})$$

**Proof:** It is immediate to see that the chain is irreducible. Because the rates are bounded we can use uniformization and define a related Discrete Time Markov Chain (DTMC). Define the set

$$C = \{(i, \max\{N \vee N + K\}) : 0 \leq i \leq N\}.$$

Let  $\tau_C$  be the first hitting time in the set  $C$ . Accordingly,  $E_x[\tau_C]$  is the expected hitting time given that  $S(0) = x$ .  $C$  is a compact set and it is easy to prove that  $\sup_{x \in C} E_x[\tau_C] < M_C < \infty$  (an elaborate derivation of the bound,  $M_C$ , would be similar to the proof of Lemma 8 in [4] and we omit the detailed argument). Stability is now established by applying theorem 10.4.10 from [9]. Equation (A85) follows directly from stability. ■

**Lemma F.2** *Let*

$$\Omega^*(\delta, \lambda, T, \Theta) := \left\{ \omega \in \Omega : 11 \cdot \max_{i=1, \dots, 11} \sup_{0 \leq t \leq T} B_i(c\lambda t) + \mathcal{E}_i \log(2 \vee c\lambda t) - \delta\lambda t \leq \Theta \right\}.$$

*Then,*

$$P \{(\Omega^*(\delta, \lambda, T, \Theta))^c\} \leq 11 \left( 2c_1 e^{-c_2(\Theta - \Gamma)} + e^{-\frac{\delta\Theta\sqrt{\lambda}}{2c}} \right),$$

*where  $\Gamma, c_1, c_2$  are as in (A5).*

**Proof:** Fix  $i = 1, \dots, 11$ . Note that

$$\begin{aligned} P \{(\Omega^*(\delta, \lambda, T, \Theta))^c\} &\leq P \left\{ \mathcal{E}_i \log(2 \vee c\lambda T) > \frac{\Theta}{2} \right\} \\ &\quad + P \left\{ \sup_{0 \leq t \leq T} B_i(c\lambda t) - \delta\lambda t > \frac{\Theta}{2} \right\} \end{aligned} \quad (\text{A86})$$

Now we treat each element on the right-hand side separately. From (A5) it now follows that

$$P \left\{ \mathcal{E}_i \log(2 \vee c\lambda T) > \frac{\Theta}{2} \right\} \leq c_1 e^{-c_2(\frac{\Theta}{2} - \Gamma) / (\log(2 \vee c\lambda T))}.$$

The second element on the right hand side of (A86) is bounded by  $e^{-\frac{\delta\sqrt{\lambda}\Theta}{2c}}$  by a well known result for negative-drift Brownian motions; see e.g. Exercise 4.3.13 in [6].  $\blacksquare$

The following auxiliary lemma is used later in the proofs of Lemmas B.1 and B.2.

**Lemma F.3** *Consider a sequence of systems such that: (a) the  $\lambda^{\text{th}}$  system uses  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{\lambda})$  agents for some  $0 \leq \beta \leq \frac{\rho\mu_s}{\mu_{cs}}$ ,  $\max(\beta, \gamma) > 0$  and, (b) the  $\lambda^{\text{th}}$  system uses  $TP[K^\lambda]$  for a sequence  $\{K^\lambda\}_{\lambda \geq 0}$  that satisfies  $K^\lambda/\sqrt{R} \rightarrow \varrho \in (-\infty, \infty)$  as  $\lambda \rightarrow \infty$ . Then, for all  $\epsilon > 0$ , there exist  $t^0(\epsilon)$ ,  $\lambda^0(\epsilon)$  (independent of the initial state  $\Xi^\lambda(0)$ ), such that,*

$$\sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \leq \epsilon\lambda, \quad (\text{A87})$$

for all  $\omega \in \Omega^*(0, \lambda, T, \epsilon\lambda/8)$  and  $\lambda \geq \lambda^0(\epsilon)$ . Here,  $t^0(\epsilon) = 0$  whenever  $\Xi^\lambda(0) \in \mathcal{A}_{\epsilon/2}^\lambda$ . Consequently,

$$P \left\{ \sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- > \epsilon\lambda \right\} \leq c_3 e^{-c_4\lambda/\log(2 \vee c\lambda T)}, \quad (\text{A88})$$

for two positive constants  $c_3$  and  $c_4$  and, finally,

$$E \left[ \sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \right] \leq 2\epsilon\lambda. \quad (\text{A89})$$

**Proof:** We fix  $\omega \in \Omega^*(0, \lambda, T, \zeta\lambda)$ . Assume that  $Z_1^\lambda(0) < \frac{\lambda}{\mu_s} - \epsilon\lambda$ . The other case is treated at the end of this proof. Define

$$\tau^\lambda = \inf \left\{ t \geq 0 : Z_1^\lambda(t) \geq \lambda/\mu_s - \frac{\epsilon}{2}\lambda \right\}.$$

Fixing an interval  $[s, t)$  with  $t \leq \tau^\lambda$  and such that  $Q^\lambda(u) = 0$  for all  $u \in [s, t)$  we have, by equation (A6), that

$$Z_1^\lambda(t) - Z_1^\lambda(s) \geq \lambda(t-s) - \mu_s \left( \frac{\lambda}{\mu_s} - \frac{\epsilon}{2}\lambda \right) (t-s) - \zeta\lambda = \mu_s \frac{\epsilon}{2}\lambda(t-s) - \zeta\lambda.$$

On the other hand for intervals  $[s, t)$  with  $t \leq \tau^\lambda$  and  $Q^\lambda(u) > K^\lambda \vee 0$ , for all  $u \in [s, t)$ , we have by equation (A8) that

$$Z_1^\lambda(t) - Z_1^\lambda(s) \geq \mu_{cs} \int_s^t (N^\lambda - Z_1^\lambda(u)) du - \zeta\lambda \geq \mu_{cs} \frac{\epsilon}{2}\lambda(t-s) - \zeta\lambda,$$

and finally, on intervals  $[s, t)$  with  $t \leq \tau^\lambda$  and such that  $0 < Q^\lambda(u) \leq K^\lambda$  for all  $u \in [s, t)$ , we have by equation (A6) that

$$Z_1^\lambda(t) - Z_1^\lambda(s) \geq -(K^\lambda \vee 0) + \lambda(t-s) - \mu_s \left( \frac{\lambda}{\mu_s} - \frac{\epsilon}{2}\lambda \right) - \zeta\lambda \geq -(K^\lambda \vee 0) + \mu_s \frac{\epsilon}{2}\lambda(t-s) - \zeta\lambda.$$

By assumption,  $K^\lambda/\sqrt{R} \rightarrow \varrho \in (-\infty, \infty)$  as  $\lambda \rightarrow \infty$ . In particular, there exists  $k > 0$  such that  $K^\lambda \leq k\sqrt{\lambda}$  for all  $\lambda$  large enough. For such values of  $\lambda$  we then have that

$$Z_1^\lambda(t \wedge \tau^\lambda) \geq Z_1^\lambda(0) + \mu_s \wedge \mu_{cs} \frac{\epsilon}{2}\lambda(t \wedge \tau^\lambda) - \zeta\lambda. \quad (\text{A90})$$

Choosing  $\zeta = \epsilon/8$  and since  $Z_1^\lambda(0) \geq 0$ , we have that

$$Z_1^\lambda(t \wedge \tau^\lambda) \geq \mu_s \wedge \mu_{cs} \frac{\epsilon}{2}\lambda(t \wedge \tau^\lambda) - \frac{\epsilon}{8}\lambda,$$

so that, by the definition of  $\tau^\lambda$ ,

$$\tau^\lambda \leq \frac{\frac{1}{\mu_s} - \frac{\epsilon}{4}}{(\mu_s \wedge \mu_{cs})\epsilon/2}, \quad (\text{A91})$$

and this holds for all  $\omega \in \Omega^*(0, \lambda, T, \epsilon\lambda/8)$ . Define now

$$\tau'^\lambda = \sup \left\{ t \geq \tau^\lambda : Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \frac{\epsilon}{2}\lambda \right\} \wedge T,$$

and

$$\tau''^\lambda = \inf \left\{ t \geq \tau'^\lambda : Z_1^\lambda(t) < \frac{\lambda}{\mu_s} - \epsilon\lambda \right\} \wedge T.$$

We note that, on  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ ,  $|Z_1^\lambda(t) - Z_1^\lambda(s)| \leq \tilde{c}\lambda t + \epsilon\lambda/8$  for some  $\tilde{c} > 0$ . This follows from (A8) and the definition of  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ . Hence,  $\tau''^\lambda > \tau'^\lambda$  on  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ . Repeating closely the arguments that lead to (A90) we have that

$$Z_1^\lambda(t) \geq Z_1^\lambda(s) + \mu_s \wedge \mu_{cs} \frac{\epsilon}{2} \lambda (t - s) - \frac{\epsilon}{8} \lambda, \quad (\text{A92})$$

for all  $\tau'^\lambda \leq s < t \leq \tau''^\lambda$ . There are now two cases to consider: if  $\tau'^\lambda = T$ , then  $Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \epsilon\lambda/2$ , for all  $t \geq \tau^\lambda$ . If, on the other hand,  $\tau'^\lambda < T$ , then by (A92), we must have that  $\tau''^\lambda(t) = T$  so that  $Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \epsilon\lambda$  for all  $t \geq \tau^\lambda$ . Consequently, we conclude that, on  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ ,  $Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \epsilon\lambda$  for all  $t \geq \tau^\lambda$ . In particular, choosing

$$t^0(\epsilon) = \frac{\frac{1}{\mu_s} - \frac{\epsilon}{4}}{(\mu_s \wedge \mu_{cs})\epsilon/2},$$

we have by (A91) that,

$$\sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \leq \epsilon\lambda, \quad (\text{A93})$$

on  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ . Using Lemma F.2 together with  $Z_1^\lambda \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_{cs}$  we also have that

$$E \left[ \sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \right] \leq \epsilon\lambda + c\lambda c_3 e^{-c_4 \epsilon \lambda / \log(2\vee c\lambda T)}, \quad (\text{A94})$$

so that there exists  $\lambda$  large enough to guarantee that the above is smaller than  $2\epsilon\lambda$ .

Finally, the last statement of the lemma follows from the above by noting that, if  $\Xi^\lambda(0) \in \mathcal{A}_\epsilon^\lambda$  then  $\tau'^\lambda = \tau^\lambda = 0$  so that  $Z_1^\lambda(t) \geq \lambda/\mu_s - \epsilon^\lambda$  for all  $t \geq 0$ .  $\blacksquare$

**Proof of Lemma B.1:** Lemma B.1 is obtained as a special case of Lemma F.3, and specifically the first part of this lemma, by assuming that  $\beta, \delta > 0$ , replacing  $\epsilon/2$  and assuming  $\Xi^\lambda(0) \in \mathcal{A}_\epsilon^\lambda$  so

that  $t_0(\epsilon) = 0$ . We note that in Lemma B.1 we focus on  $\Omega^*(\delta, \lambda, T/\lambda, \Theta)$  while Lemma F.3 uses  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$ . The proof, however, remains practically unchanged with the replacement of the time horizon from  $T$  to  $T/\lambda$  and the replacement of the set  $\Omega^*(0, \lambda, T, \epsilon\lambda/8)$  with  $\Omega^*(\delta, \lambda, T/\lambda, \Theta)$ . ■

**Proof of Lemma B.2:** Lemma B.2 is a direct consequence of Lemma F.3. Indeed, initialize the  $\lambda^{\text{th}}$  system with its steady-state distribution  $\nu^\lambda$ . Using Lemma F.3 we have that there exists  $t > 0$  such that

$$P \left\{ \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- > \epsilon\lambda \right\} \leq c_3 e^{-c_4\lambda/\log(2\nu c\lambda T)}.$$

By stationarity of  $\nu^\lambda$  we have that  $Z^\lambda(0) \stackrel{D}{=} Z^\lambda(t) \sim \nu^\lambda$  for all  $t$  hence the result of the lemma. ■

**Proof of Lemma B.3:** The proof relies on two building blocks—the analysis of the queue dynamics and the use of a Lyapunov function along the lines of [3]. Its relative simplicity is a consequence of the fact that we are only interested here in fluid-scale behavior. We provide only a sketch of the proof.

Towards that end, using the equations for the evolution of the queue length (as in the proof of Theorem B.1) as well as Lemma F.3, it is straightforward to establish the existence of  $t^1 \geq 0$  and strictly positive constants  $\delta$  and  $\gamma$  such that

$$\sup_{\xi: q(\xi) > \delta\lambda} \frac{E_\xi \left[ e^{\frac{Q^\lambda(t^1)}{\lambda}} \right]}{e^{\frac{Q^\lambda(0)}{\lambda}}} \leq e^{-\gamma},$$

for some constant  $\gamma > 0$ . Fix  $\Phi^\lambda(\xi) = e^{q(\xi)/\lambda}$  for all  $\xi \in \mathcal{X}^\lambda$ . Let

$$\phi^\lambda(t) := \sup_{\xi \in \mathcal{X}^\lambda} \frac{E_\xi(\Phi^\lambda(\Xi^\lambda(t^1)))}{(\Phi^\lambda)(\xi)}.$$

Then, using the fact that  $Q^\lambda(t) \leq Q^\lambda(0) + A^\lambda(t)$  for all  $t \geq 0$ , it is straightforward that  $\phi^\lambda(t^1) < \infty$  for all  $\lambda$  and, moreover, that

$$\limsup_{\lambda \rightarrow \infty} \phi^\lambda(t^1) < \infty.$$

It is in establishing this last bound that the restriction to fluid-level bounds simplifies things signif-

icantly.

Applying now Theorem 5 of [3] we have for all  $\lambda$  that

$$E_{\nu^\lambda}[\Phi^\lambda(\Xi^\lambda(0))] \leq \frac{e^\delta \phi^\lambda(t^1)}{1 - e^{-\gamma}},$$

and by our definition of  $\Phi^\lambda(\cdot)$  we have that

$$\limsup_{\lambda \rightarrow \infty} E \left[ e^{\frac{Q^\lambda(0)}{\lambda}} \right] < \infty.$$

The result of the lemma now follows. ■

**Proof of Lemma B.4:** We first establish the existence of fluid limits. To this end, note that  $T_i^\lambda(\cdot)$  are increasing continuous functions with  $T_i^\lambda(0) = 0$  and for  $t > s$

$$\sum_{i=1}^5 \frac{|T_i^\lambda(t) - T_i^\lambda(s)|}{\lambda} \leq \tilde{c}(t - s), \quad (\text{A95})$$

for some constant  $\tilde{c} > 0$ . This follows directly from the fact that  $Z_1^\lambda(t) + Z_2^\lambda(t) \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_{cs}$ . Invoking the Arzelà-Ascoli Theorem (see for example [1]) together with (A9) we have that the sequence

$$\left\{ \left( \frac{T_1^\lambda}{\lambda}, \dots, \frac{T_5^\lambda(t)}{\lambda}, \frac{M_{Z,Q}^\lambda(t)}{\lambda}, \frac{M_{Z_1}^\lambda(t)}{\lambda}, \frac{M_{Z_2}^\lambda(t)}{\lambda} \right), \lambda \geq 0 \right\}$$

is C-Tight in the sense of Theorem 15.5 in [1]. From equations (A10)-(A12) it then follows that the sequence

$$\left\{ \left( \frac{T_i^\lambda(t)}{\lambda}, i = 1, \dots, 5; \frac{Z_1^\lambda(t)}{\lambda}; \frac{Z_2^\lambda(t)}{\lambda}; \frac{Q^\lambda(t)}{\lambda}, \frac{M_{Z,Q}^\lambda(t)}{\lambda}, \frac{M_{Z_1}^\lambda(t)}{\lambda}, \frac{M_{Z_2}^\lambda(t)}{\lambda} \right), \lambda \geq 0 \right\}$$

is C-Tight so that every subsequence contains a further subsequence that converges to some limit. It is now straightforward to verify that every limit must satisfy equations (A36)-(A38). Consider, for example, equation (A39): Choose  $t \geq 0$  with  $\bar{Q}(t) > 0$ . It is then possible to choose  $\lambda_0$  large enough such that for all  $\lambda > \lambda_0$ , on the subsequence,  $Q^\lambda(t)/\lambda > \epsilon$  for some  $\epsilon > 0$  (and this can be

shown to also hold in some small neighborhood of  $t$ ). In particular for  $\lambda$  large enough and for any  $s$  in some neighborhood of  $t$ ,  $Q^\lambda(s) > K^\lambda$  (by the assumption on  $K^\lambda$ ), so that  $\dot{T}_1(t) = p\mu_s\bar{Z}_1(t)$ . ■

**Proof of Lemma B.5:** The argument is very simple. Assume that the statement  $[\bar{Z}_1(t) - \frac{1}{\mu_s}]^- \leq \epsilon$  is violated at time 0, that is  $\bar{Z}_1(0) < 1/\mu_s - \epsilon$ . By equation (A36), for every interval  $[s, t]$ , on which  $\bar{Q}(u) = 0$  and  $\bar{Z}_1(u) < 1/\mu_s - \epsilon$  for  $u \in [s, t]$ , we have that

$$\frac{d(\bar{Q}(t) + \bar{Z}_1(t))}{dt} \geq 1 - \mu_s(1/\mu_s - \epsilon), \quad (\text{A96})$$

or equivalently

$$\frac{d(\bar{Q}(t) + \bar{Z}_1(t))}{dt} \geq \mu_s\epsilon. \quad (\text{A97})$$

Also, by equation (A38), on intervals  $[s, t]$  such that  $\bar{Q}(u) > 0$  and  $\bar{Z}_1(u) < 1/\mu_s - \epsilon$  for  $u \in [s, t]$ , we have that

$$\frac{d\bar{Z}_1(u)}{du} \geq \mu_{cs}\bar{Z}_2(u), \quad (\text{A98})$$

and since we assumed that  $\bar{Z}_1(u) < 1/\mu_s - \epsilon, \forall u \in [s, t]$ , we have that  $\bar{Z}_2(u) \geq \frac{\beta}{\mu_s} + \epsilon$  on this interval and

$$\frac{d\bar{Z}_1(u)}{du} \geq \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right). \quad (\text{A99})$$

Combining equations (A97) and (A99) we have that for all  $t \geq 0$

$$\frac{d\bar{Z}_1(t)}{dt} \geq \left[ \mu_s\epsilon \wedge \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right) \right], \quad (\text{A100})$$

for each  $u$  with  $\bar{Z}_1(u) < 1/\mu_s - \epsilon$ . In particular, if  $\bar{Z}_1(0) < 1/\mu_s - \epsilon$ , we have that  $\exists \tilde{t}^0(\epsilon) \leq \bar{Z}_1(0)/(\mu_s(\frac{\beta}{\mu_s} + \epsilon) \wedge \mu_{cs}\epsilon)$ , with  $\bar{Z}_1(\tilde{t}^0(\epsilon)) \geq 1/\mu_s - \epsilon$ . Note that by this argument  $\bar{Z}_1$  is increasing as long as it is below  $1/\mu_s - \epsilon$ , implying that

$$\bar{Z}_1(t) \geq 1/\mu_s - \epsilon, \forall t \geq \tilde{t}^0(\epsilon). \quad (\text{A101})$$

Now, we claim that there exists a time  $\tilde{t} \geq \tilde{t}^0(\epsilon)$ , such that  $\forall t \geq \tilde{t}$ ,  $\bar{Q}(t) = 0$ . Indeed, assume that at time  $\tilde{t}^0(\epsilon)$ ,  $\bar{Q}(t) > 0$  and let

$$\underline{t} = \inf \{t \geq \tilde{t}^0(\epsilon) : \bar{Q}(t) = 0\}.$$

Then, for all  $\tilde{t}^0(\epsilon) \leq t \leq \underline{t}$ ,

$$\begin{aligned} \frac{d\bar{Q}(t)}{dt} &= 1 - \mu_s \bar{Z}_1(t) - \mu_{cs} \bar{Z}_2(t) = \mu_s \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^- - \mu_s \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^+ \\ &\quad - \mu_{cs} \left( \frac{1+\beta}{\mu_s} - \bar{Z}_1(t) \right) \leq \mu_s \epsilon - \underline{\mu} \left( \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^+ + \frac{1+\beta}{\mu_s} - \bar{Z}_1(t) \right) \\ &\leq \mu_s \epsilon - \underline{\mu} \frac{\beta}{\mu_s}, \end{aligned} \quad (\text{A102})$$

where  $\bar{\mu} = \mu_s \wedge \mu_{cs}$ . Choosing  $\epsilon$  small enough, we have that  $\dot{\bar{Q}}(t) \leq -\eta \leq 0$  for some  $\eta > 0$ . In particular,  $\underline{t} \leq \bar{Q}(\tilde{t}^0(\epsilon))/\eta$ . Moreover, since  $\dot{\bar{Q}}(t) \leq 0$  for all  $t \geq \tilde{t}^0(\epsilon)$ , we also have that  $\bar{Q}(t) = 0$  for all  $t \geq \underline{t}$ . We can now set  $\tilde{t} = \underline{t}$ . Now, since for all  $t \geq \tilde{t}$ ,  $\bar{Q}(t) = 0$ , we have by equation (A36) that  $\dot{\bar{Z}}_1(t) = 1 - \mu_s \bar{Z}_1(t)$  for all  $t \geq \tilde{t}$  and it is straightforward to show the existence of a time  $t^0(\epsilon) \geq \tilde{t}$ , such that for all  $t \geq t^0(\epsilon)$ ,  $|\bar{Z}_1(t) - \frac{1}{\mu_s}| \leq \epsilon$ .

To prove the second part of the lemma, assume that at some time  $t_0 \geq t^0(\epsilon)$ ,  $\bar{I}(t) > 0$ . Then, letting  $\bar{t} = \inf \{t \geq t_0 : \bar{I}(t) = 0\}$ , we have that on  $[t_0, \bar{t}]$ ,

$$\dot{\bar{I}}(t) = \lambda - \mu_s \bar{Z}_1(t) - \mu_{cs} \bar{Z}_2(t) + p\mu_s \bar{Z}_1(t).$$

But since  $|\bar{Z}_1(t) - \frac{1}{\mu_s}| \leq \epsilon$  for all  $t \geq t^0(\epsilon)$ , we also have that

$$\dot{\bar{I}}(t) \geq -\mu_s \epsilon - \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right) + p\mu_s \left( \frac{1}{\mu_s} - \epsilon \right). \quad (\text{A103})$$

Hence, choosing  $\epsilon$  small enough, we have the existence of  $\eta > 0$ , such that  $d\bar{I}(t) \geq \eta > 0$ , for all  $t \geq t_0$ . In particular, there exists a time  $t^*$  at which  $\bar{I}(t) = 0$ . Moreover, by repeating a similar argument starting at the first time after  $t^*$  in which  $\bar{I}(t) \geq \epsilon/2$ , we have that  $\bar{I}(t) \leq \epsilon$  for all  $t \geq t^*$ . ■

**Completing the proof of Theorem B.2.** The theorem was proved for the case  $\beta > 0$  in §B. It remains, to provide the proof for the case  $\beta = 0$ . Note that, in this case,  $N^\lambda = R + \gamma\sqrt{R} + o(\sqrt{\lambda})$ , so that we necessarily have that  $(Z_1^\lambda - \frac{\lambda}{\mu_s})^+ \leq \gamma\sqrt{R} + o(\sqrt{R})$ . Consequently, the statement of Lemma F.3 holds with  $(Z_1^\lambda - \frac{\lambda}{\mu_s})^-$  replaced with  $|Z_1^\lambda - \frac{\lambda}{\mu_s}|$ .

We will assume for the rest of the proof that  $K^\lambda \equiv 0$  but the more general case follows similarly by replacing  $I^\lambda(\cdot)$  with  $(I^\lambda(\cdot) - [K^\lambda]^-)^+$  and  $I^\lambda$  with  $(I^\lambda - [K^\lambda]^-)^+$  throughout. The remainder of the proof is similar to, but simpler than, the proof of Proposition B.1.

Fix  $\Theta > 0$ ,  $\epsilon > 0$  and assume that  $t^0(\epsilon)$  in Lemma F.3 is 0. We will later remove this last assumption. Assume further that  $I^\lambda(0) > 2\Theta$  and let

$$\tau^\lambda = \inf\{t \geq 0 : I^\lambda(t) \leq I^\lambda(0) - \Theta\}.$$

Using the identity  $I^\lambda(t) = N^\lambda - Z_1^\lambda(t) - Z_2^\lambda(t)$  as well as equations (A11) and (A12), we have that on  $\Omega^*(0, \lambda, T/\lambda, \Theta/2)$ ,

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) - \lambda(t \wedge \tau^\lambda) + \mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du + \mu_{cs} \int_0^{t \wedge \tau^\lambda} Z_2^\lambda(u) du - p\mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du + \Theta/2, \quad (\text{A104})$$

with  $\Omega^*(\cdot, \cdot, \cdot, \cdot)$  is as in (A20). From Lemma F.3 we have that, on  $\Omega^*(0, \lambda, T/\lambda, \Theta/2)$ ,  $|Z_1^\lambda(t) - \frac{\lambda}{\mu_s}| \leq \epsilon\lambda$  for all  $t \geq 0$ . By definition  $Z_2^\lambda(t) \leq N^\lambda - Z_1^\lambda(t)$  for all  $t \geq 0$ . In particular, on  $\Omega^*(0, \lambda, T/\lambda, \Theta/2)$ , we have from equation (A104) that

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) + ((1+p)\mu_s\epsilon\lambda + \mu_{cs}\epsilon\lambda)(t \wedge \tau^\lambda) - p\lambda(t \wedge \tau^\lambda) + \Theta/2. \quad (\text{A105})$$

Let  $\eta := (1+p)\mu_s\epsilon + \mu_{cs}\epsilon - p$  and choose  $\epsilon$  small enough so that  $\eta > 0$ . Let now  $t^* = 3\Theta/(2\eta)$ . Then, we must have that  $\tau^\lambda \leq t^*/\lambda$  on  $\Omega^*(0, \lambda, T/\lambda, \Theta/2)$ . By similar considerations as in the proof of Lemma F.3, we now have that for all  $t^*/\lambda \leq t \leq T/\lambda$ ,  $I^\lambda(t) \leq I^\lambda(0) - \Theta$ . From here the proof follows the proof of Proposition B.1 almost verbatim with the appropriate replacements of  $Q^\lambda$  with  $I^\lambda$ . We point out, however, that an analogue of Lemma B.3 is not required here. Indeed, the fact that

$$\limsup_{\lambda \rightarrow \infty} E \left[ \left( \frac{I^\lambda}{\lambda} \right)^m \right] < \infty,$$

for any integer  $m$  follows trivially from the fact that  $I^\lambda \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_{cs}$ . ■

The following Remark is used in the proof of Proposition D.1.

**Remark F.1 (Convergence of idleness on compact intervals)** The argument that we used in the proof of Theorem B.2 can to prove convergence of the steady-state variables, can be modified (and simplified) to establish convergence over compact intervals provided a proper convergence is assumed at time 0. Specifically, we make the following claim: Assume that, in addition to the conditions of Theorem B.2, we have that

$$\frac{I^\lambda(0)}{\sqrt{\lambda}} \Rightarrow \hat{I}(0) \text{ as } \lambda \rightarrow \infty.$$

Then, we have that

$$\frac{(I^\lambda(t) - [K^\lambda]^-)^+}{\sqrt{\lambda}} \Rightarrow 0 \text{ as } \lambda \rightarrow \infty,$$

where the convergence is uniform on compact subsets of  $(0, \infty)$ .

The proof of this claim is similar to that of Theorem B.2 and we provide only a sketch. As in that proof, we focus initially on a subset of the sample space. Specifically, we focus on  $\Omega^*(\delta, \lambda, T, \epsilon\sqrt{\lambda}/2)$ , where  $\Omega^*(\cdot, \cdot, \cdot, \cdot)$  is as in (A20)). Using Lemma F.2, we have that

$$P \left\{ (\Omega^*(\delta, \lambda, T, \epsilon\sqrt{\lambda}/2))^c \right\} \leq \epsilon/2$$

for all  $\lambda$  large enough. To establishing the convergence over compact intervals of  $(0, T]$ , it then suffices to show that for every  $0 < s < T$ , there exists for all  $\lambda$  large enough and  $\omega \in \Omega^*(\delta, \lambda, T, \epsilon\sqrt{\lambda}/2)$ ,

$$\sup_{s < t \leq T} \left| \frac{(I^\lambda(t) - [K^\lambda]^-)^+}{\sqrt{\lambda}} \right| \leq \epsilon.$$

This is what we prove next. To that end, define the random time  $\tau^\lambda = \inf\{t \geq 0 : I^\lambda(t) \leq \epsilon/2\sqrt{\lambda}\}$ . Following the arguments in the beginning of the proof of Theorem B.2, paralleling (A105), one shows that on  $\Omega^*(\delta, \lambda, T, \epsilon\sqrt{\lambda}/2)$ ,

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) + ((1+p)\mu_s\epsilon\lambda + \mu_{cs}\epsilon\lambda)(t \wedge \tau^\lambda) - p\lambda(t \wedge \tau^\lambda) + \delta\lambda t + \frac{\epsilon}{2}\sqrt{\lambda}. \quad (\text{A106})$$

The convergence of  $I^\lambda(0)/\sqrt{\lambda}$  allows us to choose  $\eta(\epsilon)$  and possibly so that  $P\{I^\lambda(0) > \eta(\epsilon)\sqrt{\lambda}\} \leq \epsilon/2$  for all  $\lambda$  large enough. Let

$$\tilde{\Omega}(\epsilon, \delta, \lambda, T) := \{\omega \in \Omega : I^\lambda(0) > \eta(\epsilon)\sqrt{\lambda}\} \cap \Omega^*(\delta, \lambda, T, \epsilon\sqrt{\lambda}/2).$$

It is now straightforward to modify the argument in the proof of Theorem B.2 to show that there exists  $t^*(\epsilon)$  such that  $I^\lambda(t) \leq \epsilon\sqrt{\lambda}$ , for all  $\omega \in \tilde{\Omega}(\epsilon, \delta, \lambda, T)$  and all  $t^*(\epsilon)/\sqrt{\lambda} \leq t \leq T$ . Consequently,

$$P \left\{ \sup_{t^*(\epsilon)/\sqrt{\lambda} \leq t \leq T} (I^\lambda(\cdot) - [K^\lambda]^-)^+ \geq \epsilon\sqrt{\lambda} \right\} \leq \epsilon,$$

for all  $\lambda$  large enough which, in turn, implies the desired convergence. ■

**Proof of Lemma C.1:** Since we fix  $\lambda$  we omit the superscript  $\lambda$  throughout the proof of the Lemma. Recall the state descriptor  $S(t) = \{Z_2(t), Y(t)\}$ . Consider the set  $A$  where all agents are busy, that is  $A = \{(i, j) : j \geq N, i \geq 0\}$ . Let  $S^A(t)$  be the process one gets when restricting the Markov chain to the set  $A$  (and in particular  $Q^A(t) = (Y^A(t) - N)^+$ ). Since the new state space is clearly irreducible  $S^A(t)$  is a Markov chain. In particular, we will have that  $E[Q|Y \geq N] = E[Q^A]$ . For the restricted Markov chain we can couple the queue length with an  $M/M/1$  queue with service rate  $N\mu_s$  as follows. Let  $Q^B(t)$  be the queue length in this  $M/M/1$  queue. Initiate  $Q^A(0) = Q^B(0) = 1$ . Generate arrivals from the same Poisson process and departures from the same Poisson process with rate  $N\mu_s + N\mu_{cs}$  with thinning. Since we assumed that  $\mu_{cs} \geq \mu_s$ , it is straightforward to show by induction on the event epochs (arrivals and departures), that for all  $t \geq 0$ ,  $Q^B(t) \geq Q^A(t)$ . But we know that  $E[Q^B] = \frac{\lambda}{N\mu_s - \lambda}$ , which implies that

$$E[Q^A] \leq \frac{\lambda}{N\mu_s - \lambda},$$

and the assertion of the lemma is now obtained by applying Little's law. ■

**Proof of Lemma D.3.** Recall the definition of  $\hat{\gamma}$  and  $\gamma$  from (A57) and (A58). Then, our assumption that  $N^\lambda \geq N_1^\lambda$  for all  $\lambda$  implies that  $\hat{\gamma} \geq \gamma$ . Let  $Y_{\lambda, \mu}^{FCFS}(N^\lambda)$  be the steady-state number of customers in system in the  $\lambda^{th}$   $M/M/N$  system and let

$$X_{\lambda, \mu}^{FCFS}(N^\lambda) := \frac{Y_{\lambda, \mu}^{FCFS}(N^\lambda) - N^\lambda}{\sqrt{R}}.$$

Then, by [5] we have that

$$X_{\lambda, \mu}^{FCFS}(N^\lambda) \Rightarrow X^{FCFS}, \tag{A107}$$

where the convergence holds also in expectation and  $X^{FCFS}$  has a density function

$$f_{\hat{\gamma}}(x) = \begin{cases} (1 - \alpha(\hat{\gamma})) \frac{\phi(\hat{\gamma}+x)}{\Phi(\hat{\gamma})}, & x \leq 0, \\ \alpha(\hat{\gamma}) e^{-\hat{\gamma}x}, & x > 0, \end{cases} \quad (\text{A108})$$

where for  $x \geq 0$ ,  $\alpha(x) := \left[1 + \frac{x\Phi(x)}{\phi(x)}\right]^{-1}$ , and  $\phi(\cdot)$  and  $\Phi(\cdot)$  are the standard normal distribution and cumulative distribution functions. In particular, for all  $x \leq 0$  the cdf is given

$$F_{\hat{\gamma}}(x) = (1 - \alpha(\hat{\gamma})) \frac{\Phi(\hat{\gamma} + x)}{\Phi(\hat{\gamma})}, \quad (\text{A109})$$

and  $E[(X^{FCFS})^+] = \frac{\alpha(\hat{\gamma})}{\hat{\gamma}}$ . By [2],  $\gamma$  is such that  $\frac{\alpha(\gamma)}{\gamma} = \sqrt{\mu_s \hat{W}}$ . Also, by our assumption that  $N^\lambda \geq N_1^\lambda$  for all  $\lambda$  we then have that  $E[(X^{FCFS})^+] \leq \sqrt{\mu_s \hat{W}}$  and the inequality is strict whenever  $\hat{\gamma} > \gamma$ .

We now turn to the actual proof of the lemma. For  $x \geq 0$ , define the functions

$$h^\lambda(x) \triangleq \xi_{\lceil -x\sqrt{N^\lambda} \rceil}(N^\lambda), \quad (\text{A110})$$

where  $h^\lambda(x) \equiv h^\lambda(\sqrt{N^\lambda})$  when  $x \geq \sqrt{N^\lambda}$ . For any fixed  $\lambda$ ,  $h^\lambda(\cdot)$  is a decreasing function. Also by (A107) and (A109), we have that  $h^\lambda(x) \rightarrow h(x) \triangleq \frac{\alpha(\hat{\gamma})}{1 - F_{\hat{\gamma}}(-x)}$ , as  $\lambda \rightarrow \infty$ . Since these are non-increasing functions the convergence is locally uniform in  $x$ . Re-writing (A52), we have

$$K^\lambda = -\sqrt{N^\lambda} \cdot \min \left\{ x \geq 0 \mid \frac{h^\lambda(x)}{N^\lambda \mu_s - \lambda} \leq \hat{W} / \sqrt{\lambda} \right\}, \quad (\text{A111})$$

or

$$\frac{K^\lambda}{\sqrt{N^\lambda}} = - \min \left\{ x \geq 0 \mid h^\lambda(x) \leq \hat{W} \times \frac{N^\lambda \mu_s - \lambda}{\sqrt{\lambda}} \right\}. \quad (\text{A112})$$

Assume first that  $\hat{\gamma} > \gamma$ . Then, we can further bound  $x$  in the following way. Since  $h(x)$  is continuous in  $x$  with  $h(x) \rightarrow \alpha(\hat{\gamma})$ , as  $x \rightarrow \infty$  and  $h(x) \rightarrow 1$ , as  $x \rightarrow 0$ , we know that as long as  $\alpha(\hat{\gamma})/(\hat{\gamma}) \leq \sqrt{\mu_s \hat{W}} - \sqrt{\mu_s} \epsilon$  for some  $\epsilon > 0$ , there exists  $\bar{x}$  such that  $h(\bar{x}) < \sqrt{\mu_s} \hat{\gamma} \hat{W} - \sqrt{\mu_s} \hat{\gamma} \epsilon$ . In particular by the pointwise convergence of the sequence  $h^\lambda(\cdot)$  we have that for every  $\lambda$  large enough  $h^\lambda(\bar{x}) \leq \sqrt{\mu_s} \hat{\gamma} \hat{W} \hat{\gamma}$ . This, together with the monotonicity of  $h^\lambda(x)$ , allow us to “localize” equation (A112) so that

$$\frac{K^\lambda}{\sqrt{N^\lambda}} = -\min \left\{ 0 \leq x \leq \bar{x} \mid h^\lambda(x) \leq \hat{W} \times \frac{N^\lambda \mu_s - \lambda}{\sqrt{\lambda}} \right\}. \quad (\text{A113})$$

The locally uniform convergence of  $h^\lambda(\cdot)$  together with the condition (A57) implies

$$\frac{K^\lambda}{\sqrt{N^\lambda}} \rightarrow -\min \left\{ 0 \leq x \leq \bar{x} \mid h(x) \leq \hat{\gamma} \hat{W} \sqrt{\mu_s} \right\}. \quad (\text{A114})$$

$h(x)$  is monotone decreasing in  $x$  and continuous. Hence, using  $N^\lambda = R + O(\sqrt{R})$ , we have that

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow -\delta(\hat{\gamma}, \hat{W}) \triangleq \{x \mid h(x) = \hat{\gamma} \hat{W} \sqrt{\mu_s}\}, \quad (\text{A115})$$

so that,

$$\delta(\hat{\gamma}, \hat{W}) = F_{\hat{\gamma}}^{-1} \left( 1 - \frac{\alpha(\hat{\gamma})}{\hat{\gamma} \hat{W} \sqrt{\mu_s}} \right). \quad (\text{A116})$$

Recall that we assumed that  $\hat{\gamma} > \gamma$ . We claim that (A115) still holds, however, when  $\hat{\gamma} = \gamma$ , and in particular, since by the definition of  $\gamma$ ,  $\alpha(\hat{\gamma}) = \hat{\gamma} \hat{W} \sqrt{\mu_s}$ , this would imply that  $\frac{K^\lambda}{\sqrt{R}} \rightarrow -\infty$ . Indeed, assume that  $\hat{\gamma} = \gamma$  and

$$\liminf_{\lambda \rightarrow \infty} \frac{K^\lambda}{\sqrt{R}} = -\hat{\delta} > -\infty.$$

Then, repeating our previous arguments (and using uniform convergence on  $[0, 2\hat{\delta}]$ ) we have that

$$h^\lambda(K^\lambda/\sqrt{R}) = \xi_{\lceil K^\lambda \rceil} \rightarrow h(\hat{\delta}) > \alpha(\gamma), \quad (\text{A117})$$

where the last inequality follows from the definition of  $h(\cdot)$ . In particular,

$$\sqrt{\lambda} \frac{\xi_{K^\lambda}}{N^\lambda \mu - \lambda} \rightarrow \frac{h(\hat{\delta})}{\sqrt{\mu_s} \gamma} > \frac{\alpha(\gamma)}{\sqrt{\mu_s} \gamma} > \hat{W}, \quad (\text{A118})$$

so that there exists  $\lambda$  large enough for which the average delay constraint is violated, contradicting the definition of  $K^\lambda$ . ■

**Proof of Lemma D.2.** By §9 of [2],  $N_1^\lambda$  is such that

$$\frac{N_1^\lambda - R}{\sqrt{R}} \rightarrow \underline{\gamma},$$

for some  $\underline{\gamma} > 0$ . In particular, (A71) follows from the fact that  $N^{*\lambda} \geq N_1^\lambda$  by definition. Assume, to reach a contradiction, that there exists a subsequence  $\lambda^k$  such that

$$\lim_{k \rightarrow \infty} \frac{N^{*\lambda^k} - R^k}{\sqrt{R^k}} = \infty. \quad (\text{A119})$$

Also, recall that in the statement of the lemma that  $\limsup_{\lambda \rightarrow \infty} (N_2^\lambda - R)/(\bar{N}_1^\lambda - R) < 1$ . Hence, we can choose a further subsequence  $\lambda^{k_l}$  so that  $(N_2^{\lambda^{k_l}} - R)/(\bar{N}_1^{\lambda^{k_l}} - R) < 1$  for all  $l$  large enough.

We fix such a subsequence.

Using Assumption 3.2 and the definition of  $\bar{N}_1^\lambda$  in (12) we also have that

$$\limsup_{\lambda \rightarrow \infty} \frac{\bar{N}_1^{\lambda^{k_l}} - R}{\sqrt{R}} < \infty. \quad (\text{A120})$$

In particular, there exists  $l^*$  so that, for all  $l \geq l^*$ , both  $N^{*\lambda^{k_l}} \geq \bar{N}_1^{\lambda^{k_l}}$  and  $(N_2^{\lambda^{k_l}} - R)/(\bar{N}_1^{\lambda^{k_l}} - R) < 1$ .

1. Using the definition of  $\bar{N}_1^{\lambda^{k_l}}$ , we then have that

$$E[W_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) | Z_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) \geq N] = E[W_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) | W_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) > 0] \leq \bar{W}^{\lambda^{k_l}},$$

for all  $l \geq l^*$  and, consequently, that  $K^{*\lambda} \geq 0$  with  $K^{*\lambda}$  as in (14) is the least threshold that satisfies the service-level constraint. It is trivially the case that  $E[Z_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) | Z_{\lambda^{k_l}, \mu_s}^{FCFS}(N^{*\lambda^{k_l}}) \geq N^{*\lambda^{k_l}}] = N^{*\lambda^{k_l}}$ , so that we can replace (13) with

$$N^{*\lambda^{k_l}} = \arg \max_{N \geq \bar{N}_1^{\lambda^{k_l}}} r\mu_{cs}(N - R) - C^{\lambda^{k_l}}(N) - C^{\lambda^{k_l}}(R), \quad (\text{A121})$$

where we always pick the smallest maximizer. The convexity of  $C^{\lambda^{k_l}}(\cdot)$  and the definition of  $N_2^{\lambda^{k_l}}$  imply that for all  $l$ ,  $\mu_{cs}(N - R) - (C^{\lambda^{k_l}}(N) - C^{\lambda^{k_l}}(R))$  is non-increasing on  $[N_2^{\lambda^{k_l}}, \infty)$ . Recall that  $(N_2^{\lambda^{k_l}} - R)/(\bar{N}_1^{\lambda^{k_l}} - R) < 1$  for all  $l \geq l^*$  and we must have that  $N^{*\lambda^{k_l}} = \bar{N}_1^{\lambda^{k_l}}$  for all  $\lambda$  large enough, implying that for all  $l \geq l^*$ . Equation (A120) now leads to a contradiction to (A119) and

in particular to equation (A72). ■

**Proof of Corollary D.1.** Let  $X^\lambda$  and  $\bar{X}^{\lambda,p^*}$  be the steady-state variables defined in (A59). Having Proposition D.1, the proof of Corollary D.1 requires only establishing that the sequences  $\{X^\lambda, \lambda \geq 0\}$  and  $\{\bar{X}^{\lambda,p^*}, \lambda \geq 0\}$  are tight sequence of random variables. Once this tightness is established, the convergence in the corollary follows from exactly the argument in Corollary 2 of [5] which is by now standard.

To establish tightness we will identify sequences of random variables  $\{L^\lambda, \lambda \geq 0\}$  and  $\{U^\lambda, \lambda \geq 0\}$  such that, for each  $\lambda$ ,  $L^\lambda \leq_{st} X^\lambda \leq U^\lambda$  and the same holds for  $\bar{X}^{\lambda,p^*}$ . We will show that the sequence  $\{L^\lambda, \lambda \geq 0\}$  and  $\{U^\lambda, \lambda \geq 0\}$  are both tight, thus implying the tightness of  $\{X^\lambda, \lambda \geq 0\}$  and  $\{\bar{X}^{\lambda,p^*}, \lambda \geq 0\}$ .

For the lower bound, we will let  $L^\lambda$  be the steady-state number of customers in an  $M/M/N$  system (i.e. without cross-selling). The fact that this is indeed a lower bound can be proved using a straightforward coupling argument that is omitted. For the upper bound, we will let  $U^\lambda$  be the steady-state number of jobs in the state dependent  $M/M/1$  defined in the proof of Proposition D.1. In that proof, this  $M/M/1$  was already proved to constitute an upper bound for both process  $X^\lambda(\cdot)$  and  $\bar{X}^{\lambda,p^*}(\cdot)$ .

The fact that  $\{L^\lambda, \lambda \geq 0\}$  converges follows directly from Theorem 1 in [5]. As for  $\{U^\lambda, \lambda \geq 0\}$ . Note that the state dependent  $M/M/1$  queue is just a state space reduction of the  $M/M/N$  system from [5], and hence can also be shown to converge using their results. Specifically, we have that

$$U^\lambda \stackrel{d}{=} S^\lambda \mid S^\lambda > N^\lambda - K^\lambda, \quad (\text{A122})$$

where  $S^\lambda$  is the steady-state number of customers in system in the corresponding  $M/M/N$  system, so that by Theorem 1 in [5]

$$U^\lambda \Rightarrow S \mid S > -\delta, \quad (\text{A123})$$

where  $S$  is the steady-state distribution of the limit in Theorem 1 of [5]. The convergence of the upper and lower bound sequence implies their tightness and, in turn, that of  $\{X^\lambda, \lambda \geq 0\}$  and  $\{\bar{X}^{\lambda,p^*}, \lambda \geq 0\}$ .

Having the tightness the proof of the corollary is concluded by mimicking the argument in the proof of Corollary 2 in [5]. We omit that argument. ■

**Proof of Corollary D.3:** By Lemma D.2 we can always choose a convergent subsequence of  $N^{*\lambda}$ . Assume first that the whole sequence converges. By Lemma D.1 we have that

$$V^\lambda(N^\lambda, \pi^\lambda) \leq r\mu_{cs}(E[\bar{Z}^\lambda] - R) - (C^\lambda(N^\lambda) - C^\lambda(R)), \quad (\text{A124})$$

where,  $\bar{Z}^\lambda$  is the steady-state number of busy agents in the G&Z model controlled by  $\overline{TP}(N^\lambda, p^*)$ . Moreover, since the limit of  $\bar{X}^{\lambda, p^*}$  is the same regardless of the value of  $p^*$ , we can use the fact that (A67) holds for the G&Z model to write:

$$\begin{aligned} V^\lambda(N^\lambda, \pi^\lambda) &\leq r\mu_{cs}(E[Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) - R \mid Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \geq N + K^\lambda] - R) \\ &\quad - (C^\lambda(N^\lambda) - C^\lambda(R)) + o(N^\lambda - R), \end{aligned} \quad (\text{A125})$$

where  $K^\lambda$  is determined through equation (A52). In particular, by the definition of  $N^{*\lambda}$ , we have that

$$\begin{aligned} \sup_{N^\lambda, \pi^\lambda} V^\lambda(N^\lambda, \pi^\lambda) &\leq \mu_{cs}(E[Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) - R \mid Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) \geq N^{*\lambda} + K^\lambda] - R) \\ &\quad + C^\lambda(N^{*\lambda}) - C^\lambda(R) + o(N^\lambda - R). \end{aligned} \quad (\text{A126})$$

By Corollary D.1, equation (A68) and the second part of Assumption 3.1 we have now that

$$\liminf_{\lambda \rightarrow \infty} \frac{V^\lambda(N^{*\lambda}, TP[K^\lambda])}{\bar{V}^\lambda(N^{*\lambda})} \rightarrow 1, \text{ as } \lambda \rightarrow \infty, \quad (\text{A127})$$

so that the upper bound is achieved. Together with equation (A69) this implies that  $TP[K^\lambda]$  and  $N^{*\lambda}$  are an asymptotically optimal staffing and control pair. Since these arguments can be repeated for every convergent subsequence the assumption that  $\frac{N^{*\lambda} - R}{\sqrt{R}}$  converges can be removed. ■

## G. Numerical examples

Here, we augment the numerical study from §5 of the paper. In the latter we examined the efficacy of our prescription of implementation of Section 5.1, but focused on cases i. and iv. of that prescription (i. is the PD regime under Condition 1 of Theorem 3.1, while iv. is the case where

none of the conditions 1, 2, and 3 holds and hence a search procedure is needed). It remains to examine the efficacy of our prescription under cases ii. and iii., consistent with Conditions 2 and 3, respectively. We start with Condition 3. This condition assumes that  $N_2^\lambda \leq \bar{N}_1^\lambda$ , where  $\bar{N}_1^\lambda$  is as given in (12), and that the service- and cross-selling rates are equal, i.e,  $\mu_s = \mu_{cs}$ . We return to the call centers that we examined in §5. A small call center with  $R = 30$  and a medium-sized one with  $R = 100$ . Here we assume that  $\mu_s = \mu_{cs} = 1$ . For each of these two systems we vary the staffing level between the least feasible staffing level  $N_1$ , as defined in (11), and  $\bar{N}_1$ . For each of these staffing levels we compute  $K^\lambda(N)$  as in (14). We then find the revenue obtained when using  $N$  servers and  $TP[K^\lambda(N)]$ . The resulting series is depicted by the solid line in Figure G. In addition, for each staffing level we compute the revenue using the MDP from §4. These points are used to construct the dashed series in Figure 7. Evidently, the performance of the prescribed threshold approaches that of the MDP as the system size grows, but it is reasonably good also for the small call center.

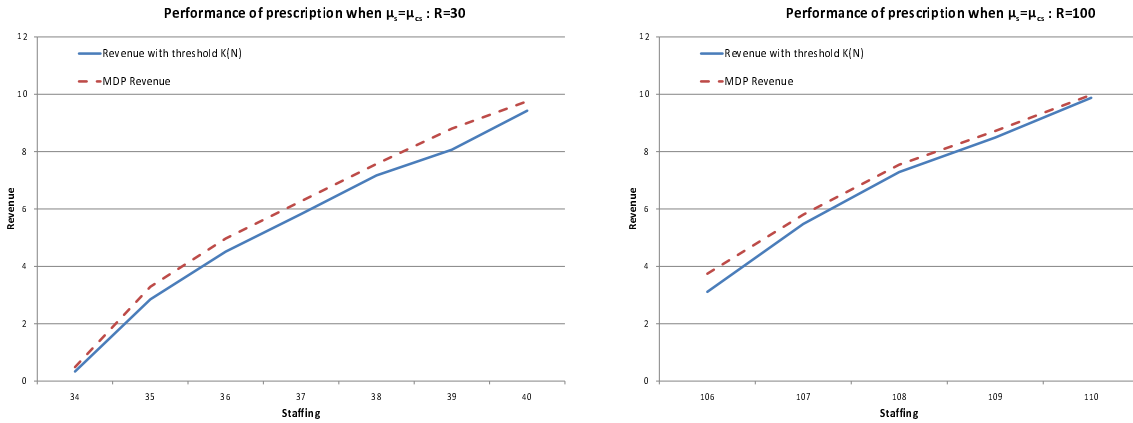


Figure 7: Performance of the threshold in (14) for call centers with equal rates:  $\mu_s = \mu_{cs}$ : (a) for  $R = 30$ , and (b)  $R = 100$ .

We now study the recommendation in Condition 2 of Theorem 3.1. Namely, we consider a case in which  $\mu_{cs} = 2 \geq \mu_s = 1$  and staffing range  $[\bar{N}_1, M]$  such that  $M$  is not much larger than  $\bar{N}_1$ —note that the recommendation in condition 2 assumes that condition 1 does not hold but still  $N_2 \geq \bar{N}_1$ . In turn, this implies that  $N_2 - \bar{N}_1 = O(\sqrt{\lambda})$ . Hence, we use  $M = \lceil \bar{N}_1 + \sqrt{\lambda} \rceil$ . For the small call center (with  $R = 30$ ) this region is almost non-existent for all practical purposes since  $\bar{N}_1 = 40$  is already 33% greater than  $R$ , so that if  $N_2 \geq \bar{N}_1$  it is fair to say that  $N_2 \gg R$ . Hence, we focus on  $R = 100$ . The results are plotted in Figure 8. The solid series depicts the result when

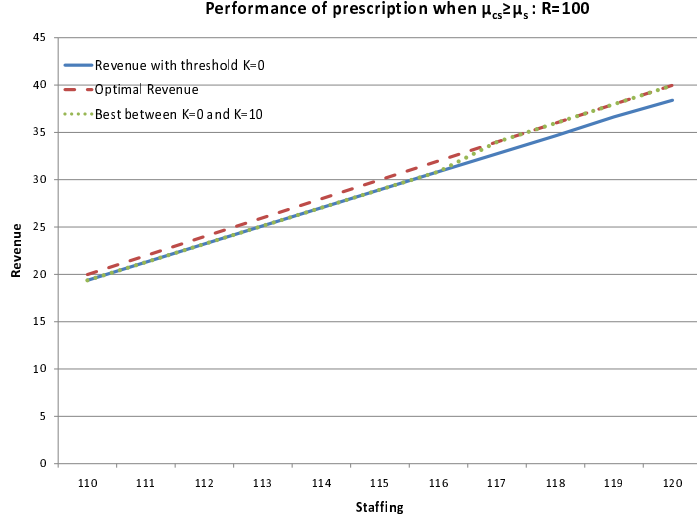


Figure 8: Performance of the threshold  $K=0$ , for  $N \geq \bar{N}_1$  and  $\mu_{cs} \geq \mu_s$

using  $K = 0$ , as recommended in Theorem 3.1 when Condition 2 holds, and the dashed series represents the optimal solution from the MDP.

Finally, an interesting question is what is the point of transition from Condition 2 to Condition 1 of Theorem 3.1. By Theorem 3.1 we expect that, when Condition 1 holds, thresholds that are strictly greater than 0 will be feasible and outperform the threshold  $K = 0$ . Hence, to identify the transition point, we compare the performance of the threshold  $K = 0$  to the performance with the best feasible threshold in the range  $[0, \lambda\bar{W}]$  which is found through a search. We recall that  $\bar{W} = 0.1$  so that  $\lambda\bar{W} = 10$ . As seen in Figure 8, for all staffing levels that are less than 115 agents, the best threshold in this range does not significantly outperform the performance with  $K = 0$ . A transition happens around 115, where we see that the using thresholds that are strictly greater than 0 leads to a performance improvement. Note that with  $N = 115$ ,  $\mu_{cs}(N - R) = 0.3\lambda$  so that the system has the capacity to cross-sell to roughly 30% of its customers.

## References

- [1] Billingsley P., “Convergence of Probability Measures”, J. Wiley & Sons, New York, 1968.
- [2] Borst S., Mandelbaum A. and Reiman M., “Dimensioning Large Call Centers”, *Operations Research*, 52(1), pp. 17-34, 2004.
- [3] Gamarnik D. and Zeevi A., “Validity of heavy traffic steady-state approximations in generalized Jackson networks”, *Annals of Applied Probability*, 16(1), pp. 5690, 2006.
- [4] Gans N. and Zhou Y.-P., “A call-routing problem with service-level constraints”, *Operations Research*, 51(2), pp. 255-271, 2003.
- [5] Halfin S. and Whitt W., “Heavy-traffic limits for queues with many exponential servers”, *Operations Research*, 29, pp. 567-587, 1981.
- [6] Karatzas I. and Shreve S.E., “Brownian Motion and Stochastic Calculus”, 2nd Edition, Springer Verlag, 1991.
- [7] Mandelbaum A., Massey W.A and Reiman M.I., “Strong Approximations for Markovian Service Networks”. *Queueing Systems* 30, pp. 149-201, 1998.
- [8] Mandelbaum A. and Pats G., “State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits”, *Annals of Applied Probability*, 8-2, pp. 569-646, 1998.
- [9] Meyn S.P. and Tweedie R.L. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag, London.
- [10] Whitt W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.