

## 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 start the technical appendix with the construction of the sample paths for all related processes. Specifically, our 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, one can write the system dynamics through the following equations:

$$Q^\lambda(t) + Z_1^\lambda(t) = Q^\lambda(0) + Z_1^\lambda(0) + \tilde{\mathcal{N}}_A(t) - \tilde{\mathcal{N}}_{D_1}(t), \quad (\text{A1})$$

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

and

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

where one should recall that  $Y^\lambda(t) = Z_1^\lambda(t) + Z_2^\lambda(t) + Q^\lambda(t)$ , and we define:

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

and finally,

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

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

The construction follows by noting that all input and output processes in the system can be modelled by 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

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

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  $P\{C_i \geq \Gamma + x\} \leq c_1 e^{-c_2 x}, \forall x > 0$ , for all  $i = \dots, 1, \dots, 11$ . Note that all the time changes of the unit rate Poisson processes in equations (A1)-(A3) are bounded by  $c\lambda$  for some positive constant  $c$ . Indeed, this is a result of the fact that we examine only cases in which  $N^\lambda \leq R + \frac{\lambda p}{\mu_{cs}}$ . We can hence write:

$$\begin{aligned}
Q^\lambda(t) + Z_1^\lambda(t) &= Q^\lambda(0) + Z_1^\lambda(0) + \lambda t - \mu_s \int_0^t Z_1^\lambda(u) du - Z_1^\lambda(t) + M_{Z,Q}^\lambda(t) \\
&+ O(\log(2 \vee c\lambda t)), \tag{A5}
\end{aligned}$$

$$\begin{aligned}
Z_2^\lambda(t) &= Z_2^\lambda(0) - \mu_{cs} \int_0^t Z_2^\lambda(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du \\
&\quad - p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Y^\lambda(u) - N^\lambda > K^\lambda\}} du + M_{Z_2}^\lambda(t) + O(\log(2 \vee c\lambda t)), \tag{A6}
\end{aligned}$$

and

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

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  $K^\lambda > 0$ ,

$$\begin{aligned}
M_{Z_1}^\lambda(t) &= B_1 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}} du \right) \\
&\quad - B_5 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) + B_6 \left( (1-p)\mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \\
&\quad + B_7 \left( p\mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Y^\lambda(u) - N^\lambda \leq K^\lambda\}} du \right) \\
&\quad + B_3 \left( \mu_{cs} \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du \right), \tag{A7}
\end{aligned}$$

where  $B_i(\cdot)$ ,  $i = 1, \dots, 7$ , are standard Brownian motions.

Using the Brownian Motion strong law of large numbers (see problem 2.9.3. in [6]) and the fact that the time arguments are all bounded by  $c\lambda t$  for some constant  $\lambda$ , we have that

$$\left( \frac{M_{Z,Q}^\lambda(t)}{\lambda}, \frac{M_{Z_2}^\lambda(t)}{\lambda}, \frac{M_{Z_1}^\lambda(t)}{\lambda} \right) \rightarrow (0, 0, 0), \text{ as } \lambda \rightarrow \infty, \tag{A8}$$

where the convergence is uniform on compact sets. We also define  $C = C_1 + C_2 + \dots + C_{11}$  with  $C_i$ ,  $i = 1, \dots, 11$  as defined in equation (A4). Defining the processes

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

one can re-write equations (A5)-(A7) as follows:

$$\begin{aligned} Q^\lambda(t) &= Q^\lambda(0) + Z_1^\lambda(0) + \lambda t - \mu_s \int_0^t Z_1^\lambda(u) du - Z_1^\lambda(t) + M_{Z,Q}^\lambda(t) \\ &+ O(\log(2 \vee c\lambda t)), \end{aligned} \tag{A9}$$

$$\begin{aligned} Z_2^\lambda(t) &= Z_2^\lambda(0) - \mu_{cs} \int_0^t Z_2^\lambda(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du - T_1^\lambda(t) + M_{Z_2}^\lambda(t) \\ &+ O(\log(2 \vee c\lambda t)), \end{aligned} \tag{A10}$$

and

$$Z_1^\lambda(t) = Z_1^\lambda(0) + T_2^\lambda(t) - \mu_s \int_0^t Z_1^\lambda(u) du + T_3^\lambda(t) + T_4^\lambda(t) - T_5^\lambda(t) + M_{Z_1}^\lambda(t) + O(2 \vee \log(c\lambda t)). \tag{A11}$$

**Notational conventions and organization of the appendix.** We let  $\Xi^\lambda(t) := (Q^\lambda(t), Z_2^\lambda(t), Z_1^\lambda(t))$ , and let  $\mathcal{X}$  be the state-space in which  $\Xi^\lambda(t)$  obtains values. We let  $\nu^\lambda$  be the unique stationary distribution of  $\Xi^\lambda(t)$  (which exists by Lemma 2.2). We use the notation  $\xi$  for a general element in  $\mathcal{X}$  and for a given  $\xi$  we let  $q(\xi)$ ,  $z_2(\xi)$  and  $z_1(\xi)$  be its corresponding coordinates. Finally, we use  $E_\xi[\cdot]$  for the expectation with respect to the initial condition  $\xi$ . Accordingly, we let  $E_{\nu^\lambda}[\cdot]$  be the expectation with respect to an initial condition that is distributed according to the stationary distribution  $\nu^\lambda$ .

The rest of this appendix is organized as follows. Each of the sections A, B, C is dedicated to prove Theorem 4.1 under one of the conditions 1, 2 and 3, respectively. §D is dedicated to proving the result in §5 of the main paper. Finally, §E provides proofs for some Lemmas that were given in the paper and several auxiliary results that are used in this appendix.

## A. Condition 1

The main result of this section is the following Theorem:

**Theorem A.1** *Consider a sequence of systems with  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{\lambda})$  for some  $0 < \beta \leq \frac{p\mu_s}{\mu_{cs}}$  and fix a sequence  $K^\lambda \geq 0$  with*

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow \delta, \text{ as } \lambda \rightarrow \infty,$$

with  $\delta \geq 0$ . Then, under  $TP[K^\lambda]$ ,

$$\frac{E[I^\lambda]}{N^\lambda - R} \rightarrow 0, \text{ as } \lambda \rightarrow \infty, \quad (\text{A12})$$

and

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

The first part of the theorem will be proved in Corollary A.3 and the second part will be proved in A.1. The intuition behind the latter is based on the large extra capacity of the system. Specifically, since  $TP[K^\lambda]$  dictates that whenever the queue length is greater than  $K^\lambda$  - every service or cross-selling completion is followed by an admission of a customer from the queue into service - we have that whenever the queue is longer than  $K^\lambda$ , the depletion rate of the queue is roughly  $\mu_s R + \mu_{cs}(N - R) \gg \lambda$ . In particular, the queue depletion rate is much greater than the input rate to the queue leading to extremely small excess queue above the level of  $K^\lambda$ .

Most of this section is dedicated to the proof of Theorem A.1. The main complication arises from the fact that we are interested in convergence of the steady-state variables  $Q^\lambda$  and  $I^\lambda$  rather than mere convergence on finite time intervals. Before proceeding with the proof we state and prove the asymptotic optimality result for this section:

**Corollary A.1** *Assume that  $N_2^\lambda - R^\lambda \gg N_1^\lambda - R$ , in addition to Assumptions 4.1 and 4.2. Then, the following is asymptotically optimal:*

- **Staffing:** Staff with  $N_2^\lambda$  agents.
- **Control:** Use  $TP[K^\lambda]$  with  $K^\lambda = \lceil \delta \sqrt{R} \rceil$  such that  $\delta \geq 0$  and  $K^\lambda \leq \lambda \bar{W}^\lambda / 2$ .

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

$$\frac{E[W^\lambda]}{\bar{W}^\lambda} = \frac{E[Q^\lambda]}{\lambda \bar{W}^\lambda} \leq \frac{K^\lambda + E[(Q^\lambda - K^\lambda)^+]}{\lambda \bar{W}^\lambda}. \quad (\text{A14})$$

Equation (A13) now implies that

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq \frac{1}{2}, \quad (\text{A15})$$

and in particular that  $TP[K^\lambda]$  is asymptotically feasible. Also, since the system is stable, we have by Little's law that  $E[Z_1^\lambda] = \lambda/\mu_s$ . But recall that  $E[Z_2^\lambda] = N - E[Z_1^\lambda] - E[I^\lambda]$ . Hence, (A12) implies that

$$\frac{\mu_{cs}E[Z_2^\lambda]}{\mu_{cs}(N_2^\lambda - R)} \rightarrow 1, \text{ as } \lambda \rightarrow \infty. \quad (\text{A16})$$

Recall that for each  $\lambda$ ,  $r\mu_{cs}(N_2^\lambda - R) - (C(N_2^\lambda) - C^\lambda(R))$  constitutes an upper bound for the optimal value of the cross-selling problem (10). Equation (A16) implies that the upper bound is asymptotically achieved leading to asymptotic optimality of the sequence  $(N_2^\lambda, TP[K^\lambda])$ . ■

We proceed now to prove Theorem A.1. The proof is composed of several components:

- In Lemma A.1 we show that the process  $Z_1^\lambda(t)$  cannot take values that are much smaller than  $R$ . Using a Lyapunov function argument, the bound is then extended to the steady state distribution in Corollary A.2.
- In Proposition A.1 we show that the steady state queue length is negligible with respect to  $\lambda$ .
- In Lemmas A.3 and A.4 we prove the convergence to fluid limits that satisfy certain characteristics, and, finally,
- Corollary A.3 uses all the previous components to complete the proof of Theorem A.1.

**Lemma A.1** *Consider a sequence of systems, where the  $\lambda^{\text{th}}$  system is staffed with  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{R})$  for  $0 \leq \beta \leq \frac{\nu\mu_s}{\mu_{cs}}$ ,  $\max(\beta, \gamma) > 0$ , and operated with  $TP[K^\lambda]$ , where*

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow \delta, \text{ as } \lambda \rightarrow \infty, \quad (\text{A17})$$

for some  $\delta \in (-\infty, \infty)$ . Then, for all  $\epsilon > 0$ , there exist  $t^0(\epsilon)$  and  $\lambda^0(\epsilon)$  (independent of the initial conditions), such that for all  $\lambda \geq \lambda^0(\epsilon)$  and  $T \geq t^0(\epsilon)$ ,

$$E \left[ \sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- \right] \leq \epsilon\lambda, \quad (\text{A18})$$

and

$$P \left\{ \sup_{t^0(\epsilon) \leq t \leq T} \left( Z_1^\lambda(t) - \frac{\lambda}{\mu_s} \right)^- > \epsilon\lambda \right\} \leq c_3 e^{-c_4\lambda/\log(2\nu c\lambda T)}, \quad (\text{A19})$$

for two positive constants  $c_3$  and  $c_4$ , where  $\left(Z_1^\lambda(t) - \frac{\lambda}{\mu_s}\right)^- = \max\left\{\frac{\lambda}{\mu_s} - Z_1^\lambda(t), 0\right\}$ .

Before proceeding to the proof we have the following corollary which follows directly from Lemma A.1 by initializing the system with its stationary distribution.

**Corollary A.2** *Under the condition of Lemma A.1, there exists  $T$  such that*

$$P\left\{\left(Z_1^\lambda - \frac{\lambda}{\mu_s}\right)^- > \epsilon\lambda\right\} \leq c_3 e^{-c_4\lambda/\log(2\vee c\lambda T)}, \quad (\text{A20})$$

for all  $\lambda$  large enough.

**Remark A.1** *Note that Lemma A.1 and Corollary A.2 are not restricted to any form of staffing sequences. Hence, we will make use of these also for the analysis of the service driven regime.*

**Proof of Lemma A.1:** Fix  $\delta > 0$  and define the set

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

Assume that at time 0,  $Z_1^\lambda(0) < \frac{\lambda}{\mu_s} - \epsilon\lambda$  and define

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

Consider equations (A5) and (A7). Then, on every interval  $[s, t)$  with  $t \leq \tau^\lambda$ , with  $Q^\lambda(u) = 0$  for all  $u \in [s, t)$  we have by equation (A5) that

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

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 (A7) that

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

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 (A5) that

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

Our assumptions on the magnitude of the threshold,  $K^\lambda$ , imply that for  $\lambda$  large enough there exists  $k > 0$  such that  $K^\lambda \leq k\sqrt{\lambda}$  for all  $\lambda$  thereafter. Hence, we have that for  $\lambda$  large enough and on  $\Omega^*(\delta, \lambda)$

$$Z_1^\lambda(t \wedge \tau^\lambda) \geq Z_1^\lambda(0) + \mu_s \wedge \mu_{cs} \frac{\epsilon}{2} \lambda(t \wedge \tau) - \delta \lambda. \quad (\text{A21})$$

Choosing  $\delta = \epsilon/8$  and recalling that by definition  $Z_1^\lambda(0) \geq 0$ , we have that on  $\Omega^*(\delta, \lambda)$ ,

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

Define now

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

and

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

But on  $\Omega^*(\delta, \lambda)$  and for every  $\tau'^\lambda \leq s < t \leq \tau''^\lambda$ , we have by our previous argument that

$$Z_1^\lambda(t) \geq Z_1^\lambda(s) + \eta(\epsilon)\lambda(t-s) - \epsilon\lambda/4$$

implying that  $Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \epsilon\lambda$ , 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 that on  $\Omega^*(\epsilon)$ ,

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

Now, note that

$$P((\Omega^*(\delta, \lambda))^c) \leq P(C \log(2 \vee c\lambda t) \geq \delta\lambda) + \sum_{i=1}^{11} P\left(\sup_{0 \leq t \leq T} \sqrt{c\lambda} |B_i(t)| \geq \delta\lambda\right), \quad (\text{A23})$$

implying that (A19) holds by equation (A4) and recalling that for a Brownian Motion  $B(t)$  and any  $b > 0$

$$P\left(\sup_{0 \leq t \leq T} |B(t)| \geq b\right) \leq \sqrt{\frac{T}{2\pi}} \frac{4}{b} e^{-b^2/2T},$$

whenever  $B(0) = 0$  (see Problem 2.8.2 in [6]). Moreover, since  $Z_1^\lambda \leq c\lambda$  we also have that

$$E\left[\sup_{t^{0(\epsilon)} \leq t \leq T} \left(Z_1^\lambda(t) - \frac{\lambda}{\mu_s}\right)^-\right] \leq \epsilon\lambda + c\lambda c_3 e^{-c_4 \epsilon \lambda / \log(2 \vee c\lambda T)}, \quad (\text{A24})$$

so that there exists  $\lambda$  large enough so that the above is smaller than  $\frac{3\epsilon}{2}\lambda$ . Repeating the argument with  $\frac{2}{3}\epsilon$  instead of  $\epsilon$  we have the result of the lemma.  $\blacksquare$

**Proposition A.1** *Under the assumption of Theorem A.1,*

$$\frac{E[(Q^\lambda - K^\lambda)^+]}{\sqrt{\lambda}} \rightarrow 0. \quad (\text{A25})$$

**Proof:** We prove the result for  $K^\lambda \equiv 0$ . The proof is similar for arbitrary  $K^\lambda > 0$  with  $Q^\lambda(\cdot)$  replaced everywhere with  $(Q^\lambda(\cdot) - K^\lambda)^+$ . We start by re-defining

$$\Omega^*(\delta, \lambda) = \left\{ \omega \in \Omega : 11 \cdot \max_{i=1, \dots, 11} \sup_{0 \leq t \leq T/\lambda} |B_i(c\lambda t)| + C \log(2 \vee c\lambda T) \leq \delta \right\},$$

where we assume that  $T$  is larger than  $2t^*$  (where  $t^*$  is defined in (A29)). Assume first that at time 0,  $[Z_1^\lambda(0) - \lambda/\mu_s]^- \leq \epsilon\lambda$ . Fix a constant  $K > 0$ , assume that  $Q^\lambda(0) > 2K$  and let

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

Then, plugging equation (A7) into equation (A5), and using the fact that  $Z_2^\lambda(t) = N^\lambda - Z_1^\lambda(t)$  whenever  $Q^\lambda(t) > 0$ , we have that on  $\Omega^*(\delta, \lambda)$

$$Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \lambda(t \wedge \tau^\lambda) - \mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du - \mu_{cs} \int_0^{t \wedge \tau^\lambda} N^\lambda - Z_1^\lambda(u) du + \delta. \quad (\text{A26})$$

Recall from the proof of Lemma A.1, that there exists  $\delta$  small enough so that on  $\Omega^*(\delta, \lambda)$ ,  $Z_1^\lambda(t) \geq \frac{\lambda}{\mu_s} - \epsilon\lambda$  for all  $t \geq 0$ . In particular, on  $\Omega^*(\delta, \lambda)$  and for all  $t \leq \tau^\lambda \wedge T$ ,

$$\begin{aligned} \lambda - \mu_s Z_1^\lambda(t) - \mu_{cs} Z_2^\lambda(t) &= \mu_s \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^- - \mu_s \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^+ - \mu_{cs} \left( \frac{1+\beta}{\mu_s} - Z_1^\lambda(t) \right) \\ &\leq \mu_s \epsilon - \underline{\mu} \left( \left( Z_1^\lambda(t) - \frac{1}{\mu_s} \right)^+ + \frac{1+\beta}{\mu_s} - Z_1^\lambda(t) \right) \\ &\leq \mu_s \epsilon \lambda - \underline{\mu} \frac{\beta}{\mu_s} \lambda, \end{aligned} \quad (\text{A27})$$

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

$$Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \left( \mu_s \epsilon \lambda - \underline{\mu} \frac{\beta}{\mu_s} \lambda \right) (t \wedge \tau^\lambda) + \delta. \quad (\text{A28})$$

Choosing  $\epsilon$  and  $\delta$  small enough and letting  $\eta := -(\mu_s \epsilon - \underline{\mu} \frac{\beta}{\mu_s})$  and

$$t^* = \frac{3K/2 + \delta}{\eta}, \quad (\text{A29})$$

we must have that  $\tau^\lambda \leq t^*/\lambda$  on  $\Omega^*(\delta, \lambda)$ . By similar considerations as in the proof of Lemma A.1, we now have that for all  $t^*/\lambda \leq t \leq T/\lambda$ ,  $Q^\lambda(t) \leq Q^\lambda(0) - K$ . In particular, on  $\Omega^*(\delta, \lambda)$ ,

$$\sup_{\xi \in \mathcal{A}: q(\xi) > 2K} \frac{Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2}{q(\xi)} \leq -K, \quad (\text{A30})$$

where  $\mathcal{A} := \{\xi \in \mathcal{X} : (z_1(\xi) - \frac{\lambda}{\mu_s})^- \leq \epsilon\lambda\}$ .

As in the proof of Lemma A.1, one can prove that  $P((\Omega^*(\delta, \lambda))^c) \leq c_5 e^{-c_6 \lambda \delta}$ , for some positive constants  $c_5$  and  $c_6$ . Since  $Q^\lambda(t) \leq Q^\lambda(0) + A^\lambda(t)$  we then have that

$$\sup_{\xi \in \mathcal{A}: q(\xi) > 2K} \frac{E_\xi[Q^\lambda(2t^*/\lambda)^2] - q(\xi)^2}{q(\xi)} \leq -K + E[A^\lambda(2t^*/\lambda) 1_{(\Omega^*(\delta, \lambda))^c}]. \quad (\text{A31})$$

Using the Cauchy-Schwartz inequality and noting that  $E[2A^\lambda(2t^*\lambda)] \leq c_7$  for some constant  $c_7$  and for all  $\lambda$ , we can choose  $\lambda$  large enough so that

$$\sup_{\xi \in \mathcal{A}: q(\xi) > 2K} \frac{E_\xi[Q^\lambda(2t^*/\lambda)^2] - q(\xi)^2}{q(\xi)} \leq -\frac{K}{2}. \quad (\text{A32})$$

Also, using again the fact that  $Q^\lambda(t) \leq Q^\lambda(0) + A^\lambda(t)$ , we readily have that,

$$\sup_{\xi \in \mathcal{A}: q(\xi) \leq 2K} E[Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2] \leq c_8, \quad (\text{A33})$$

for some constant  $c_8$  (that depends on  $K$  but is independent of  $\lambda$ ). After some simple manipulations we have that

$$q(\xi)^2 - E_\xi[Q^\lambda(2t^*/\lambda)^2] \geq \frac{K}{2}q(\xi) - c_8 + \left( \frac{K}{2}q(\xi) - c_8 - E_\xi[Q^\lambda(2t^*/\lambda)^2 - q(\xi)^2] \right) 1\{\xi \notin \mathcal{A}\} \quad (\text{A34})$$

We now follow an argument very similar to the proof of Theorem 5 in [3]. Specifically, let  $\nu^\lambda(\cdot)$  be the stationary distribution of the process  $\Xi^\lambda(t)$ . Then,

$$E_{\nu^\lambda}[Q^\lambda(0)^2] = E_{\nu^\lambda}[Q^\lambda(2t^*/\lambda)^2], \quad (\text{A35})$$

and in particular,

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

where  $q(\xi)$  is the queue component of the state  $\xi$ . By equation (A34) we now have that

$$E_{\nu^\lambda}[Q^\lambda(0)] \leq \frac{2c_8}{K} - \frac{2}{K} \left( E_{\nu^\lambda} \left[ \left( \frac{K}{2}Q^\lambda(0) - c_8 - E[Q^\lambda(2t^*/\lambda)^2 - Q^\lambda(0)^2 | Q^\lambda(0)] \right) 1\{\Xi^\lambda(0) \notin \mathcal{A}\} \right] \right) \quad (\text{A36})$$

We now have the following Lemma whose proof we postpone to §E.

**Lemma A.2** *Under the assumption of Theorem A.1, and for any integer  $m \geq 1$ ,*

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

Using this Lemma together with Corollary A.2, we have by the Cauchy-Schwartz inequality that

$$\limsup_{\lambda \rightarrow \infty} E_{\nu^\lambda} [Q^\lambda(0)^m 1\{\xi \notin \mathcal{A}\}] = 0.$$

Applying this with some minor manipulation to (A36) we get that

$$E_{\nu^\lambda} [Q^\lambda(0)] \leq c_9,$$

for some constant  $c_9$  and all  $\lambda$  large enough. Consequently,

$$\limsup_{\lambda \rightarrow \infty} \frac{E[Q^\lambda]}{\sqrt{\lambda}} = 0. \quad (\text{A37})$$

■

**Lemma A.3 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 A.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 to some limit almost surely 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:

$$\bar{Z}_1(t) + \bar{Q}(t) = \bar{Q}(0) + \bar{Z}_1(0) + t - \int_0^t \mu_s \bar{Z}_1(u) du, \quad (\text{A38})$$

$$\bar{Z}_2(t) = \bar{Z}_2(0) - \mu_{cs} \int_0^t \bar{Z}_2(u) du + p\mu_s \int_0^t Z_1^\lambda(u) du - \bar{T}_1(t), \quad (\text{A39})$$

$$\bar{Z}_1(t) = \bar{Z}_1(0) + \bar{T}_2(t) - \mu_s \int_0^t \bar{Z}_1(u) du + \bar{T}_3(t) + \bar{T}_4(t) - \bar{T}_5(t), \quad (\text{A40})$$

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

$$\dot{\bar{T}}_2(t) 1_{\{\bar{Z}_1 + \bar{Z}_2 < \frac{1+\beta}{\mu_s}\}} = 1, \quad (\text{A42})$$

$$\dot{\bar{T}}_3(t)1_{\{\bar{Q}(t)>0\}} = \mu_s \bar{Z}_1(t), \quad (\text{A43})$$

$$\dot{\bar{T}}_4(t)1_{\{\bar{Q}(t)>0\}} = \mu_{cs} \bar{Z}_2(t), \quad (\text{A44})$$

$$\dot{\bar{T}}_5(t)1_{\{\bar{Q}(t)>0\}} = 0. \quad (\text{A45})$$

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

$$\sum_{i=1}^5 \frac{|T_i^\lambda(t) - T_i^\lambda(s)|}{\lambda} \leq c(t - s). \quad (\text{A46})$$

This follows directly from the fact that  $Z_1^\lambda(t) + Z_2^\lambda(t) \leq N^\lambda \leq c\lambda$  for some  $c > 0$ . Invoking the Arzelà-Ascoli Theorem (see for example [1]) together with (A8) we have that the sequence

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

is relatively compact. In particular, from equations (A9)-(A11) it follows that the sequence

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

is relatively compact so that every subsequence contains a further subsequence that converges to some limit. It is trivial that every limit must satisfy the equations (A38)-(A40). To see why equations (A41)-(A45) must hold consider for example equation (A41): Choose  $t \geq 0$  with  $\bar{Q}(t) > 0$ . It is then possible to choose  $\lambda_0$  large enough along the subsequence 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  $\dot{\bar{T}}_1(t) = p\mu_s \bar{Z}_1(t)$ . ■

**Lemma A.4** Fix  $\epsilon > 0$  and assume  $0 < \beta \leq \frac{p\mu_s}{\mu_{cs}}$ . For any process

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

satisfying equations (A38)-(A45), there exists  $t^0(\epsilon)$  (independent of  $\bar{Z}_1(0)$  and  $\bar{Z}_2(0)$ ), such that for all  $t \geq t^0(\epsilon)$

$$\left| \bar{Z}_1(t) - \frac{1}{\mu_s} \right| \leq \epsilon, \quad (\text{A47})$$

for every fluid limit  $(\bar{Q}(t), \bar{Z}_1(t), \bar{Z}_2(t))$ . Moreover, there exists  $t^* \geq t^0(\epsilon)$ , such that for all  $t \geq t^*$ :

$$\bar{I}(t) \leq \epsilon, \quad (\text{A48})$$

where

$$\bar{I}(t) = \frac{1 + \beta}{\mu_s} - \bar{Z}_1(t) - \bar{Z}_2(t).$$

**Proof:** 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 (A38), 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

$$\frac{d(\bar{Q}(t) + \bar{Z}_1(t))}{dt} \geq 1 - \mu_s(1/\mu_s - \epsilon), \quad (\text{A49})$$

or equivalently

$$\frac{d(\bar{Q}(t) + \bar{Z}_1(t))}{dt} \geq \mu_s \epsilon. \quad (\text{A50})$$

Also, by equation (A40), 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

$$\frac{d\bar{Z}_1(u)}{du} \geq \mu_{cs} \bar{Z}_2(u), \quad (\text{A51})$$

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

$$\frac{d\bar{Z}_1(u)}{du} \geq \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right). \quad (\text{A52})$$

Combining equations (A50) and (A52) we have that for all  $t \geq 0$

$$\frac{d\bar{Z}_1(t)}{dt} \geq \left[ \mu_s \epsilon \wedge \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right) \right], \quad (\text{A53})$$

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

$$\bar{Z}_1(t) \geq 1/\mu_s - \epsilon, \forall t \geq \tilde{t}^0(\epsilon). \quad (\text{A54})$$

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

$$\begin{aligned} \frac{d\bar{Q}(t)}{dt} &= 1 - \mu_s \bar{Z}_1(t) - \mu_{cs} \bar{Z}_2(t) = \mu_s \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^- - \mu_s \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^+ \\ &\quad - \mu_{cs} \left( \frac{1+\beta}{\mu_s} - \bar{Z}_1(t) \right) \leq \mu_s \epsilon - \underline{\mu} \left( \left( \bar{Z}_1(t) - \frac{1}{\mu_s} \right)^+ + \frac{1+\beta}{\mu_s} - \bar{Z}_1(t) \right) \\ &\leq \mu_s \epsilon - \underline{\mu} \frac{\beta}{\mu_s}, \end{aligned} \quad (\text{A55})$$

where  $\bar{\mu} = \mu_s \wedge \mu_{cs}$ . Choosing  $\epsilon$  small enough, we have that  $d\bar{Q}(t) \leq -\eta \leq 0$  for some  $\eta > 0$ . In particular,  $\underline{t} \leq \bar{Q}(\tilde{t}^0(\epsilon))/\eta$ . Moreover, since  $d\bar{Q}(t) \leq 0$  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 (A38) that  $\frac{d\bar{Z}_1(t)}{dt} = 1 - \mu_s \bar{Z}_1(t)$  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  $[t_0, \bar{t}]$ ,

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

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

$$d\bar{I}(t) \geq -\mu_s \epsilon - \mu_{cs} \left( \frac{\beta}{\mu_s} + \epsilon \right) + p\mu_s \left( \frac{1}{\mu_s} - \epsilon \right). \quad (\text{A56})$$

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

$t \geq t^*$ . ■

The following Theorem shows that the number of idle server doesn't 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 §B and §C.

**Theorem A.2** *Consider a sequence of systems, where the  $\lambda^{\text{th}}$  system is staffed with  $N^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{R})$  for  $0 \leq \beta \leq \frac{\rho\mu_s}{\mu_{cs}}$ ,  $\max(\beta, \gamma) > 0$ , and operated with  $TP[K^\lambda]$ , where*

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow \delta, \text{ as } \lambda \rightarrow \infty, \quad (\text{A57})$$

for some  $\delta \in (-\infty, \infty)$ . Then,

$$\frac{E[((N^\lambda - Z^\lambda) - [K^\lambda]^-)^+]}{N^\lambda - R} \rightarrow 0, \text{ as } \lambda \rightarrow \infty, \quad (\text{A58})$$

where for a real number  $x$ ,  $[x]^- = \max(-x, 0)$ , and  $[x]^+ = \max(x, 0)$ .

**Proof:** We prove here the theorem only for the case  $\beta > 0$ . The case  $\beta = 0$  which is more involved is proved in §E. To that end, initialize the  $\lambda^{\text{th}}$  system with  $(Q^\lambda(0), Z_1^\lambda(0), Z_2^\lambda(0))$  distributed according to the stationary distribution. Then, the process  $(Q^\lambda(t), Z_1^\lambda(t), Z_2^\lambda(t))$  is a stationary process implying that  $E[I^\lambda(t)] = E[I^\lambda]$  for all  $t \geq 0$ . Now, by Proposition A.1 and since  $Z_1^\lambda + Z_2^\lambda \leq N^\lambda$ , we have that the sequence

$$\left( \frac{Q^\lambda}{\lambda}, \frac{Z_1^\lambda}{\lambda}, \frac{Z_2^\lambda}{\lambda} \right)$$

is tight and every limit point is of the form  $(0, \bar{Z}_1(0), \bar{Z}_2(0))$ . In particular, the sequence of processes  $\left( \frac{Q^\lambda(t)}{\lambda}, \frac{Z_1^\lambda(t)}{\lambda}, \frac{Z_2^\lambda(t)}{\lambda} \right)$  is tight and every limit point  $(\bar{Q}(t), \bar{Z}_1(t), \bar{Z}_2(t))$  satisfies the fluid limit equations (A38)-(A45). We can thus apply Lemma A.4 to conclude the existence of  $t^*$  such that for all  $t \geq t^*$ ,  $\bar{I}(t) \leq \epsilon$  and this holds for any limit point. In particular, we have that for all  $t \geq t^*$ ,

$$\limsup_{\lambda \rightarrow \infty} \frac{I^\lambda(t)}{\lambda} \leq \epsilon, \text{ a.s.}, \quad (\text{A59})$$

where the bound also holds when applying expectation since  $I^\lambda(t) \leq N^\lambda \leq c\lambda$ , for some constant  $c > 0$ . That is,

$$\limsup_{\lambda \rightarrow \infty} \frac{E[I^\lambda(t)]}{\lambda} \leq \epsilon. \quad (\text{A60})$$

Recall now that  $I^\lambda(t)$  has the distribution of  $I^\lambda$ . Thus, we can conclude that

$$\limsup_{\lambda \rightarrow \infty} \frac{E[I^\lambda]}{\lambda} \leq \epsilon. \quad (\text{A61})$$

Since this is true for any  $\epsilon$  we have the assertion of the Lemma. ■

The following corollary follows directly from Theorem A.2 by noting that under the assumptions of Theorem A.2 we have that  $[K^\lambda]^- = 0$ .

**Corollary A.3** *Under the assumptions of Theorem A.1,*

$$\frac{E[I^\lambda]}{N^\lambda - R} \rightarrow 0. \quad (\text{A62})$$

## B. Condition 2

The following proposition establishes the asymptotic feasibility of the proposed solution under Condition 2.

**Proposition B.1** *Assume  $\mu_{cs} \geq \mu_s$  and  $N^\lambda$  agents are used for the  $\lambda^{\text{th}}$  system with*

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

*Then, using  $TP[0]$  for all  $\lambda$ , we have that*

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1. \quad (\text{A63})$$

This is a rather straightforward result. Recall that by the definition of  $\bar{N}_1^\lambda$ ,

$$E[W_{\lambda, \mu_s}^{FCFS}(N) | W_{\lambda, \mu_s}^{FCFS}(N) > 0] \leq \bar{W}^\lambda.$$

Now,  $TP[0]$  dictates that in the presence of a positive queue length, every service completion will be followed by an admission of a customer from the queue. Thus, whenever all agents are busy the system will deplete customers at a faster rate than the associated  $M/M/N$  (using here the fact that  $\mu_{cs} \geq \mu_s$ ) and the result will hold recalling that, by Condition 2,  $N^\lambda - R \geq \bar{N}_1^\lambda - R$  for  $\lambda$  large enough.

The asymptotic optimality result is then the following corollary:

**Corollary B.1** *Assume that  $\mu_{cs} \geq \mu_s$ , and*

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

*as well as that Assumptions 4.1 and 4.2 hold. Then, the following is asymptotically optimal:*

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

**Proof:** Proposition B.1 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 A.1 using Theorem A.2. ■

**Proof of Proposition B.1:** 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  $\mu_{cs}$ . Then, we have the following intuitive result.

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

$$E[W^\lambda | W^\lambda > 0] \leq E[W_{\lambda, \mu_s}^{FCFS}(N) | W_{\lambda, \mu_s}^{FCFS}(N) > 0] = \frac{1}{N\mu_s - \lambda} \quad (\text{A64})$$

Having Lemma B.1 the proposition readily follows. Specifically, fix a sequence of staffing levels  $N^\lambda$  with

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

Then, by Lemma B.1,

$$\frac{E[W^\lambda | W^\lambda > 0]}{\bar{W}^\lambda} \leq \frac{1}{\bar{W}^\lambda (N^\lambda \mu_s - \lambda)}. \quad (\text{A65})$$

But, from the definition of  $\bar{N}_1^\lambda$  we have that

$$\limsup_{\lambda \rightarrow \infty} \frac{1}{\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)} \leq 1, \quad (\text{A66})$$

so that, recalling that  $R = \frac{\lambda}{\mu_s}$ , we have

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda | W^\lambda > 0]}{\bar{W}^\lambda} \leq \limsup_{\lambda \rightarrow \infty} \frac{1}{\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)} \frac{\bar{W}^\lambda (\bar{N}_1^\lambda \mu_s - \lambda)}{\bar{W}^\lambda (N^\lambda \mu_s - \lambda)} \leq 1, \quad (\text{A67})$$

and the proof is completed by noting that

$$E[W^\lambda] = E[W^\lambda | W^\lambda > 0] P\{W^\lambda > 0\} \leq E[W^\lambda | W^\lambda > 0].$$

■

### C. Condition 3

Our optimality results under Condition 3 are closely related to the results of Gans and Zhou [4]. We start then with a description of the system analyzed in [4] as well as with some results comparing this system with the cross-selling system. While [4] considers a system that is essentially different from the cross-selling system, we prove that in this asymptotic regime the two problems are, in some sense, equivalent. Specifically, we prove that the model 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 here 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 must maximize the rate at which type-L jobs are processed, subject to the service-level constraint placed on the type-H work. 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. Intuitively then, 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 C.1.

The above 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 strongly dependent on 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 using *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 C.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 without randomization the threshold reservation policy defined in Definition C.1 can be thought of as the TP control adapted to the G&Z model. Denote by  $\bar{TP}^\lambda(N^\lambda, p^*)$  the randomized threshold policy of G&Z with threshold  $K^\lambda$  determined through (A68) 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 C.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 one of two cases holds: 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 threshold  $K^\lambda$  is chosen so that

$$K^\lambda(N^\lambda) = \max \left\{ k \in [-N^\lambda, 0] \mid \frac{\xi_k(N^\lambda)}{N^\lambda \mu_s - \lambda} \leq \bar{W}^\lambda \right\}. \quad (\text{A68})$$

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  $\mu_s$ .

**Remark C.1** Note that under  $\bar{TP}(N^\lambda, 0)$  (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  $E[\bar{Z}^\lambda] = E[Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \mid Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \geq N^\lambda + K^\lambda]$ .

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. Note that, by the same argument as in Lemma 2.1, 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  $\mu_{cs}(E[\bar{Z}^{\lambda, \pi'}] - R)$ . The following lemma does not assume  $\mu_s = \mu_{cs}$ .

**Lemma C.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

$$TH^{\pi_{g\&z}^*}(t) \geq_{st} CS^\pi(t), \quad \forall t \geq 0. \quad (\text{A69})$$

In particular,

$$\mu_{cs}(E[Z^{\lambda, \pi}] - R) \leq \mu_{cs}(E[\bar{Z}^{\lambda, \pi_{g\&z}^*}] - R). \quad (\text{A70})$$

**Proof:** We use a sample path construction and a coupling argument. We will show that under our sample path construction the inequality (A69) holds a.s. This, in turn, implies the stochastic ordering in (A69).

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

$$TH_{\pi_1}(t) = CS_{\pi_1}(t), \forall t \geq 0, \quad (\text{A71})$$

and

$$\mu_{cs}(E[\bar{Z}^{\lambda, \pi_1}] - R) = \mu_{cs}(E[Z^{\lambda, \pi_1}] - R). \quad (\text{A72})$$

Since we fixed the scheduling policy for system 1 we can continue by saying that

$$\max_{\pi} \mu_{cs}(E[\bar{Z}^{\lambda, \pi}] - R) \geq \mu_{cs}(E[Z^{\lambda, \pi_1}] - R), \quad (\text{A73})$$

and in particular,

$$\mu_{cs}(E[\bar{Z}^{\lambda, \pi_{g\&z}^*}] - R) \geq \sup_{\pi_1} \mu_{cs}(E[Z^{\lambda, \pi_1}] - R). \quad (\text{A74})$$

■

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  $N^\lambda$  be a sequence with  $N^\lambda \geq N_1^\lambda$  for all  $\lambda$  and

$$\frac{N^\lambda - R}{\sqrt{R}} \rightarrow \hat{\gamma}, \text{ as } \lambda \rightarrow \infty, \quad (\text{A75})$$

for some  $\hat{\gamma} > 0$ . The existence of such a sequence is guaranteed since by §9 of [2] we have that

$$\frac{N_1^\lambda - R}{\sqrt{R}} \rightarrow \gamma, \quad (\text{A76})$$

for some  $\gamma > 0$ . Also, 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  $T\bar{P}^\lambda(N^\lambda, p^*)$ . 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 (A68). Accordingly, we let  $\bar{Z}^\lambda$  and  $Z^\lambda$  be the number of busy agents in the above two systems. Define the scaled variables

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

**Lemma C.2** *Assume that  $\mu_s = \mu_{cs}$ ,  $N^\lambda$  satisfies (A75) and  $K^\lambda$  is defined through (A68). Then, there exists a function  $\delta(\cdot, \cdot)$  such that*

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow \delta(\hat{\gamma}, \hat{W}), \text{ as } \lambda \rightarrow \infty, \quad (\text{A77})$$

where  $\delta(\cdot, \cdot)$  is finite for all positive and finite arguments.

Lemma C.2 is proved in §E. It is a key component in the proof of the following convergence result where we 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  $T\bar{P}^\lambda(N^\lambda, p^*)$ . 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 (A68). Define the scaled processes

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

**Proposition C.1 Diffusion Limits:** *With  $N^\lambda$  satisfying (A75), and assuming that*

$$\bar{X}^{\lambda,p^*}(0) \Rightarrow \bar{X}(0), \text{ and } X^\lambda(0) \Rightarrow \bar{X}(0), \text{ as } \lambda \rightarrow \infty, \quad (\text{A78})$$

*we have that*

$$\bar{X}^{\lambda,p^*}(\cdot) \Rightarrow \bar{X}(\cdot) \text{ in } D \text{ as } \lambda \rightarrow \infty, \quad (\text{A79})$$

*and*

$$X^\lambda(\cdot) \Rightarrow \bar{X}(\cdot) \text{ in } D \text{ as } \lambda \rightarrow \infty \quad (\text{A80})$$

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

$$m(x) = \begin{cases} -\beta\mu_s, & x \geq 0 \\ -(\beta + x)\mu_s, & -\delta \leq x \leq 0 \end{cases} \quad (\text{A81})$$

*and infinitesimal variance term  $\sigma^2 = 2\mu_s$ , where  $\delta := \delta(\hat{\gamma}, \hat{W})$  is given in (A115).*

**Proof:** We start by proving the result for the sequence  $X^\lambda(\cdot)$ . We use coupling with a Birth and Death (B&D) process corresponding to a state dependent  $M/M/1$  system for which the limits are known. In particular, consider the B&D process with rates:  $\hat{\lambda}_i = \lambda$  for all  $i$ , where  $i$  is the number of customers in the system, and

$$\hat{\mu}_i^\lambda = \begin{cases} (N^\lambda + K^\lambda + i)\mu_s & 1 \leq i \leq -K^\lambda - 1 \\ N^\lambda\mu_s & i \geq K^\lambda \\ 0 & \text{otherwise} \end{cases}, \quad (\text{A82})$$

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  $\hat{X}^\lambda(\cdot) = \frac{\hat{Y}^\lambda(\cdot) + K^\lambda}{N^\lambda - R}$ . Initializing the  $\lambda^{\text{th}}$  B&D process with  $\hat{X}^\lambda(0) \vee -K^\lambda$  and recalling the convergence of  $K^\lambda/\sqrt{R}$ , the sequence  $\hat{X}^\lambda(\cdot)$ , converges weakly to  $\bar{X}(\cdot)$  with the diffusion parameters given in equation (A81) (see for example [8]). In order to complete our proof

we have to show that for any  $T > 0$

$$d^T(\hat{X}^\lambda, X^\lambda) \xrightarrow{P} 0, \text{ as } \lambda \rightarrow \infty \quad (\text{A83})$$

where  $d^T(\cdot, \cdot) = \sup_{0 < t \leq T} \|\hat{X}^\lambda(t) - X^\lambda(t)\|$ . By the convergence together theorem (see for example Theorem 11.4.7 of [10]) we will have the desired result. In order to evaluate  $d^T(\hat{X}^\lambda, X^\lambda)$ , we can use a coupling argument and deduce that

$$d^T(\hat{X}^\lambda, X^\lambda) \leq \sup_{0 < t \leq T} \frac{[Z^\lambda(t) - (N^\lambda + K^\lambda)]^-}{\sqrt{N^\lambda - R}}, \quad (\text{A84})$$

which converges to zero by Remark E.1. To complete the proof, then, we present the coupling argument, in which we omit the superscript  $\lambda$  for simplicity of notation. We 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 (A82) 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. ■

We now show that the process-wise convergence also implies in this case convergence of the steady-state variables.

**Corollary C.1** *With  $\mu_s = \mu_{cs}$  and  $N^\lambda$  satisfying (A75), there exists a random variable  $\bar{X}$  such that for any  $p^* \in [0, 1]$*

$$\bar{X}^{\lambda,p^*} \Rightarrow \bar{X}, \text{ as } \lambda \rightarrow \infty, \quad (\text{A85})$$

and

$$X^\lambda \Rightarrow \bar{X}, \text{ as } \lambda \rightarrow \infty, \quad (\text{A86})$$

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

The proof of Corollary C.1 is given in §E. Note that since the result is independent of  $p^*$  we have by Remark C.1 that whenever  $N^\lambda$  satisfies (A75)

$$\frac{E[Z_{\lambda,\mu_s}^{FCFS}(N^{*\lambda}) - R \mid Z_{\lambda,\mu_s}^{FCFS}(N^{*\lambda}) \geq N^{*\lambda} + K^\lambda] - (E[\bar{Z}^\lambda] - R)}{N^\lambda - R} \rightarrow 0, \text{ as } \lambda \rightarrow \infty. \quad (\text{A87})$$

This readily follows by noting that one can write  $\bar{Z}^\lambda = N^\lambda - (\bar{Y}^{\lambda,p^*} - N^\lambda)^-$ .

Corollary C.1 leads immediately to the following result, which establishes asymptotic optimality for a fixed sequence of staffing levels,  $N^\lambda$ .

**Corollary C.2** *Assume  $\mu_s = \mu_{cs}$  and consider a sequence of cross-selling systems such that  $N^\lambda$  satisfies equation (A75), and for each  $\lambda$ ,  $TP[K^\lambda]$  is used with  $K^\lambda$  determined through equation (A68). Then,*

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1, \quad (\text{A88})$$

and

$$\frac{V^\lambda(N^\lambda, TP[K^\lambda])}{\bar{V}^\lambda(N^\lambda)} \rightarrow 1, \text{ as } \lambda \rightarrow \infty. \quad (\text{A89})$$

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

**Lemma C.3** *Assume  $\mu_s = \mu_{cs}$  in addition to Assumptions 4.1 and 4.2. Let  $N^{*\lambda}$  and  $K^\lambda$  be determined through (18) and (19) and assume that  $\limsup_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} < 1$ . Then,*

$$\liminf_{\lambda \rightarrow \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} > 0, \quad (\text{A90})$$

and

$$\limsup_{\lambda \rightarrow \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} < \infty. \quad (\text{A91})$$

**Corollary C.3** *Assume that  $\mu_s = \mu_{cs}$  in addition to Assumptions 4.1 and 4.2. Also, assume that  $\limsup_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} < 1$ . Then, the following is asymptotically optimal for the cross-selling system:*

- **Staffing:** Staff with  $N^{*\lambda}$  agents where  $N^{*\lambda}$  is given by equations (18) and (19).
- **Control:** Use  $TP[K^\lambda(N^{*\lambda})]$  where  $K^\lambda(N^{*\lambda})$  is given by equation (19).

## D. Proofs for §5

### D.1 Proof of Lemma 5.1

The argument is straightforward and we only briefly outline it. For any policy  $\pi \in \Pi(N)$  we construct the corresponding sample paths as follows: We generate the 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 actual 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. Clearly, 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, we will have exactly the same sample paths under  $\pi'$  which is obtained from  $\pi$  by admitting customers to service in a FCFS manner. 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  $Z_2$  has the same probability law under both  $\pi$  and  $\pi'$ . ■

## D.2 Proof of Lemma 5.2

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 work conserving system). We assume that both systems are initiated 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 will have  $c_i = 0$ .

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 initiated 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,

$$r_i^j = [s_i + c_i - [b_{n-1}^2 - b_i^j]^+]^+ . \quad (\text{A92})$$

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  $b_n^1 \geq b_{n-1}^2 + \min_{i \leq n-1} \{r_i^1 | r_i^1 > 0\}$ .

■

### D.3 Proof of Proposition 5.1

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

Our aim is now to show that the problems have an asymptotically equal optimal value. For this purpose it suffices to show that starting from a work conserving policy  $\pi$ , 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 vice versa, i.e. that starting from a sequence of policies  $\{\pi^L\}$ , the sequence of adapted policies for the infinite buffer system,  $\{\pi_L\}$ , achieves asymptotically the same value for the infinite buffer system. Hence, it suffices to show that for any  $\epsilon > 0$ , for all  $L$  large enough

$$|\tilde{V}(N, \pi) - \hat{V}(N, L, \pi^L)| \leq \epsilon, \quad (\text{A93})$$

and

$$|\tilde{V}(N, \pi_L) - \hat{V}(N, L, \pi^L)| \leq \epsilon. \quad (\text{A94})$$

Here,  $\tilde{V}(N, \pi)$  and  $\hat{V}(N, L, \pi^L)$  are, respectively, the cross-selling rates in the infinite and finite buffer systems, equipped with  $\pi$  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 (A93) it suffices to show that

$$E[Z_2^{\pi}] = \lim_{L \rightarrow \infty} E \left[ Z_{2B}^{L, \pi^L} \right],$$

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 start, by fixing  $\pi$ , a feasible policy for the infinite buffer system and proving (A93). We will consider only this direction since the proof of (A94) is analogous. 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,

$$E[Z_2^\pi] = E \left[ Z_{2B}^{L, \pi^L} \right] P\{Y^\pi \leq L\} + E[Z_2^\pi 1_{\{Y^\pi > L\}}]. \quad (\text{A95})$$

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

$$P\{Y^\pi > L\} = P\{Q^\pi > L - N\} \leq \frac{\lambda \bar{W}}{L - N} \quad (\text{A96})$$

Moreover, by the Cauchy-Schwartz inequality

$$E[Z_2^\pi 1_{\{Y^\pi > L\}}] \leq \sqrt{E[(Z_2^\pi)^2]} P\{Y^\pi > L\}. \quad (\text{A97})$$

Since, by definition,  $Z_2^\pi \leq N$ , we then have that

$$E[Z_2^\pi 1_{\{Y^\pi > L\}}] \rightarrow 0, \text{ as } L \rightarrow \infty. \quad (\text{A98})$$

Plugging (A96) and (A98) back into equation (A95) we have that

$$E[Z_2^\pi] = \lim_{L \rightarrow \infty} E \left[ Z_{2B}^{L, \pi^L} \right]. \quad (\text{A99})$$

## E. Proofs of auxiliary results

**Proof of Lemma 2.2.** 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 we start at state  $x$ .  $C$  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]. ■

**Proof of Lemma A.2:** The proof uses a Lyapunov function argument and applies Theorem 5 of [3]. We give only the outline of the proof. Towards that end, using the equations for the evolution of the queue length (as in the proof of Theorem A.1) as well as Lemma A.1, we can easily show that there exists a time  $t^1$  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  $\phi^{\lambda}(t) := \sup_{\xi \in \mathcal{X}^{\lambda}} (\Phi^{\lambda})^{-1}(\xi) E_{\xi}(\Phi^{\lambda}(\Xi^{\lambda}(t)))$ . 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

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

Applying 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 consequently 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.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. ■

**Completing the proof of Theorem E.1.** The theorem was proved for the case  $\beta > 0$  within the proof of Theorem A.1. Indeed, in Theorem A.1 a non-negative value is assumed for  $K^\lambda$ , but the proof remains practically the same for any sequence of thresholds  $K^\lambda$  with  $K^\lambda = O(\sqrt{\lambda})$ . It remains, then, to prove Theorem A.2 for the case  $\beta = 0$ . Recall from Remark A.1 that Lemma A.1 holds irrespective of whether  $\beta > 0$  or not. Moreover, since  $N^\lambda = R + \gamma\sqrt{R} + o(\sqrt{\lambda})$ , we have that for all  $\lambda$ , Lemma A.1 holds with  $(Z_1^\lambda - \frac{\lambda}{\mu_s})^-$  replaced with  $|Z_1^\lambda - \frac{\lambda}{\mu_s}|$ . The proof of Theorem A.2 for the case  $\beta = 0$  is now similar to the proof of Proposition A.1 but we give it for completeness. We assume for 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.

Recall the definition

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

Fix  $\epsilon > 0$  and assume that  $t^0(\epsilon)$  in Lemma A.1 is 0. Fix  $K > 0$ , assume that  $I^\lambda(0) > 2K$  and let  $\tau^\lambda = \inf\{t \geq 0 : I^\lambda(t) - I^\lambda(0) \leq -K\}$ . Using the identity  $I^\lambda(t) = N^\lambda - Z_1^\lambda(t) - Z_2^\lambda(t)$  as well as equations (A10) and (A11), we have that on  $\Omega^*(\delta, \lambda)$ ,

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) - \lambda(t \wedge \tau^\lambda) + \mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du + \mu_{cs} \int_0^{t \wedge \tau^\lambda} Z_2^\lambda(u) du - p\mu_s \int_0^{t \wedge \tau^\lambda} Z_1^\lambda(u) du + \delta. \quad (\text{A100})$$

Recall from the proof of Lemma A.1, that there exists  $\delta$  small enough so that on  $\Omega^*(\delta, \lambda)$ ,  $|Z_1^\lambda(t) - \frac{\lambda}{\mu_s}| \leq \epsilon\lambda$  for all  $t \geq 0$ . From now on we assume that  $\delta$  does indeed satisfy this requirement. By definition  $Z_2^\lambda(t) \leq N^\lambda - Z_1^\lambda(t)$  for all  $t \geq 0$ . In particular, on  $\Omega^*(\delta, \lambda)$ , we have from equation (A100) that

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) + ((1+p)\mu_s\epsilon\lambda + \mu_{cs}\epsilon\lambda)(t \wedge \tau^\lambda) - p\lambda(t \wedge \tau^\lambda) + \delta. \quad (\text{A101})$$

Let  $\eta := (1+p)\mu_s\epsilon + \mu_{cs}\epsilon - p$  and choose  $\epsilon$  and  $\delta$  small enough so that  $\eta > 0$ . Also, let

$$t^* = \frac{K + \delta}{\eta}, \quad (\text{A102})$$

then, we must have that  $\tau^\lambda \leq t^*/\lambda$  on  $\Omega^*(\delta, \lambda)$ . By similar considerations as in the proof of Lemma A.1, we now have that for all  $t^*/\lambda \leq t \leq T/\lambda$ ,  $I^\lambda(t) \leq I^\lambda(0) - K$ . From here the proof follows almost verbally the proof of Proposition A.1 with the appropriate replacements of  $Q^\lambda$  with  $I^\lambda$ . We point out, however, that an analogue of Lemma A.2 is not required here as  $I^\lambda \leq N^\lambda$  and consequently  $I^\lambda/\lambda$  trivially has finite moments of all orders.  $\blacksquare$

**Remark E.1** *The argument in the Proof of Theorem A.2 can be easily modified to show that if*

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

and the assumptions of Theorem A.2 hold, then

$$\frac{(I^\lambda(\cdot) - [K^\lambda]^-)^+}{\sqrt{\lambda}} \Rightarrow 0, \text{ as } \lambda \rightarrow \infty, \quad (\text{A103})$$

where the convergence is uniform on compact subsets of  $(0, \infty)$ . Indeed, fix  $\epsilon > 0$  and consider the set

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

Then, it is straightforward to show that there exists  $\lambda_0$  such that for all  $\lambda \geq \lambda_0$ ,  $P((\Omega^*(\delta, \lambda))^c) \leq \epsilon/2$ . Define the stopping 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 A.2, paralleling (A101), one can show that on  $\Omega^*(\delta, \lambda)$ , we now write

$$I^\lambda(t \wedge \tau^\lambda) \leq I^\lambda(0) + ((1+p)\mu_s\epsilon\lambda + \mu_{cs}\epsilon\lambda)(t \wedge \tau^\lambda) - p\lambda(t \wedge \tau^\lambda) + \delta\lambda t + \frac{\epsilon}{2}\sqrt{\lambda}. \quad (\text{A104})$$

By the convergence of  $I^\lambda(0)/\sqrt{\lambda}$  we may choose  $\eta(\epsilon)$  and possibly re-define  $\lambda_0$  so that for all  $\lambda \geq \lambda_0$ ,  $P\{I^\lambda(0) > \eta(\epsilon)\sqrt{\lambda}\} \leq \epsilon/2$ . By choosing  $\delta$  appropriately, it is now straightforward to modify the argument in the proof of Theorem A.2 to show that whenever  $I^\lambda(0) \leq \eta(\epsilon)\sqrt{\lambda}$ , there

exists  $t^*(\epsilon)$  such that on  $\Omega^*(\delta, \lambda)$  and for all  $t^*(\epsilon)/\sqrt{\lambda} \leq t \leq T$  we have that  $I^\lambda(t) \leq \epsilon\sqrt{\lambda}$ . In particular, for all  $\lambda$  large enough,

$$P \left\{ \sup_{t^*(\epsilon)/\sqrt{\lambda} \leq t \leq T} (I^\lambda(\cdot) - [K^\lambda]^-)^+ \geq \epsilon\sqrt{\lambda} \right\} \leq \epsilon, \quad (\text{A105})$$

which implies the desired convergence. ■

**Proof of Lemma C.2.** Recall the definition of  $\hat{\gamma}$  and  $\gamma$  from (A75) and (A76). 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  $\lambda^{\text{th}}$   $M/M/N$  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

$$X_{\lambda,\mu}^{FCFS}(N^\lambda) \Rightarrow X^{FCFS}, \quad (\text{A106})$$

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

$$f_{\hat{\gamma}}(x) = \begin{cases} (1 - \alpha(\hat{\gamma})) \frac{\phi(\hat{\gamma}+x)}{\Phi(\hat{\gamma})}, & x \leq 0, \\ \alpha(\hat{\gamma}) e^{-(\hat{\gamma}x)}, & x > 0, \end{cases} \quad (\text{A107})$$

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

$$F_{\hat{\gamma}}(x) = (1 - \alpha(\hat{\gamma})) \frac{\Phi(\hat{\gamma} + x)}{\Phi(\hat{\gamma})}, \quad (\text{A108})$$

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

$$h^\lambda(x) \triangleq \xi_{\lceil -x\sqrt{N^\lambda} \rceil}(N^\lambda), \quad (\text{A109})$$

where  $h^\lambda(x) \equiv h^\lambda(\sqrt{N^\lambda})$  when  $x \geq \sqrt{N^\lambda}$ . It is trivial to show that for any fixed  $\lambda$ ,  $h^\lambda(\cdot)$  is a decreasing function. Also by (A106) and (A108), 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 (A68), we have

$$K^\lambda = -\sqrt{N^\lambda} \cdot \min \left\{ x \geq 0 \mid \frac{h^\lambda(x)}{N^\lambda \mu_s - \lambda} \leq \hat{W}/\sqrt{\lambda} \right\}, \quad (\text{A110})$$

or

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

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 (A111) so that

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

The locally uniform convergence of  $h^\lambda(\cdot)$  together with the condition (A75) implies

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

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

$$\frac{K^\lambda}{\sqrt{R}} \rightarrow -\delta(\hat{\gamma}, \hat{W}) \triangleq \{x \mid h(x) = \hat{\gamma} \hat{W} \sqrt{\mu_s}\}, \quad (\text{A114})$$

so that,

$$\delta(\hat{\gamma}, \hat{W}) = F_{\hat{\gamma}}^{-1} \left( 1 - \frac{\alpha(\hat{\gamma})}{\hat{\gamma} \hat{W} \sqrt{\mu_s}} \right). \quad (\text{A115})$$

Recall that we assumed that  $\hat{\gamma} > \gamma$ . We claim that (A114) still holds 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 would have that

$$h^\lambda(K^\lambda/\sqrt{R}) = \xi_{\lceil K^\lambda \rceil} \rightarrow h(\hat{\delta}) > \alpha(\gamma), \quad (\text{A116})$$

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

$$\sqrt{\lambda} \frac{\xi_{K^\lambda}}{N^\lambda \mu - \lambda} \rightarrow \frac{h(\hat{\delta})}{\sqrt{\mu_s} \gamma} > \frac{\alpha(\gamma)}{\sqrt{\mu_s} \gamma} > \hat{W}, \quad (\text{A117})$$

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

**Proof of Lemma C.3.** 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, (A90) follows from the fact that  $N^{*\lambda} \geq N_1^\lambda$  by definition. Assume, now, that there exists a subsequence  $\lambda^k$  such that

$$\lim_{k \rightarrow \infty} \frac{N^{*\lambda^k} - R^k}{\sqrt{R^k}} = \infty. \quad (\text{A118})$$

It is trivial to check that

$$\limsup_{\lambda \rightarrow \infty} \frac{\bar{N}_1^\lambda - R}{\sqrt{R}} < \infty. \quad (\text{A119})$$

In particular, there exists  $k^*$  such that for all  $k \geq k^*$ ,  $N^{*\lambda} \geq \bar{N}_1^\lambda$  so that for all  $k \geq k^*$ ,

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

so that for all  $k \geq k^*$ , we will have that  $K^\lambda \geq 0$ . Clearly

$$E[Z_{\lambda^k, \mu_s}^{FCFS}(N^{*\lambda^k}) | Z_{\lambda^k, \mu_s}^{FCFS}(N^{*\lambda^k}) \geq N^{*\lambda^k}] = N^{*\lambda^k},$$

so that for all  $k \geq k^*$ ,

$$N^{*\lambda^k} = \arg \max_{N \geq \bar{N}_1^{\lambda^k}} r\mu_{cs}(N - R) - C^{\lambda^k}(N) - C^{\lambda^k}(R), \quad (\text{A120})$$

where we always pick the smallest maximizer. The convexity of  $C^\lambda(\cdot)$  and the definition of  $N_2^\lambda$  imply that for all  $\lambda$ ,  $\mu_{cs}(N - R) - C^\lambda(N) - C^\lambda(R)$  is non-increasing on  $[N_2^\lambda, \infty)$ . The smallest maximizer in (A120) must then be  $\bar{N}_1^\lambda$ , implying that for all  $k \geq k^*$ ,  $N^{*\lambda^k} = \bar{N}_1^\lambda$ . Equation (A119) now leads to a contradiction of (A118) and in particular of equation (A91). ■

## E.1 Proof of Corollary C.3.

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

$$V^\lambda(N^\lambda, \pi^\lambda) \leq r\mu_{cs}(E[\bar{Z}^\lambda] - R) - (C^\lambda(N^\lambda) - C^\lambda(R)), \quad (\text{A121})$$

where,  $\bar{Z}^\lambda$  is the steady state number of busy agents in the G&Z model controlled by  $T\bar{P}(N^\lambda, p^*)$ . Moreover, since the limit of  $\bar{X}^{\lambda, p^*}$  is the same regardless of the value of  $p^*$ , we can use Remark C.1 to write:

$$\begin{aligned} V^\lambda(N^\lambda, \pi^\lambda) &\leq r\mu_{cs}(E[Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) - R | Z_{\lambda, \mu_s}^{FCFS}(N^\lambda) \geq N + K^\lambda] - R) \\ &\quad - (C^\lambda(N^\lambda) - C^\lambda(R)) + o(N^\lambda - R), \end{aligned} \quad (\text{A122})$$

where  $K^\lambda$  is determined through equation (A68). In particular, by the definition of  $N^{*\lambda}$ , we have that

$$\begin{aligned} \sup_{N^\lambda, \pi^\lambda} V^\lambda(N^\lambda, \pi^\lambda) &\leq \mu_{cs}(E[Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) - R \mid Z_{\lambda, \mu_s}^{FCFS}(N^{*\lambda}) \geq N^{*\lambda} + K^\lambda] - R) \\ &+ C^\lambda(N^{*\lambda}) - C^\lambda(R) + o(N^\lambda - R). \end{aligned} \quad (\text{A123})$$

By Corollary C.1, equation (A87) and the second part of Assumption 4.1 we have now that

$$\liminf_{\lambda \rightarrow \infty} \frac{V^\lambda(N^{*\lambda}, TP[K^\lambda])}{\bar{V}^\lambda(N^{*\lambda})} \rightarrow 1, \text{ as } \lambda \rightarrow \infty, \quad (\text{A124})$$

so that the upper bound is achieved. Together with equation (A88) 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. ■

**Proof of Corollary C.1.** Let  $X^\lambda$  and  $\bar{X}^{\lambda, p^*}$  be the steady state variables for  $X^\lambda(\cdot)$  and  $\bar{X}^{\lambda, p^*}(\cdot)$ . Then, having Proposition C.1, the proof of Corollary C.1 is based on the standard argument used in [5] and hence we omit most of it. We only specify the lower bound and upper bound systems for each  $\lambda$ . These bounds are the same for  $\bar{X}^{\lambda, p^*}$  and  $X^\lambda$  and they imply the tightness of the sequences  $X^\lambda$ . Specifically, the lower bound system is an  $M/M/N$  system 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 take the state dependent  $M/M/1$  defined in the proof of Proposition C.1 which was already proved to constitute an upper bound. The sequence of steady state variables for the lower bound converges as  $\lambda \rightarrow \infty$  by [5]. As for the upper bound sequence, note that this 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

$$\hat{X}^\lambda \stackrel{d}{=} S^\lambda \mid S^\lambda > N^\lambda - K^\lambda, \quad (\text{A125})$$

where  $S^\lambda$  is the steady state random variable for the corresponding  $M/M/N$  system, so that by [5]

$$\hat{X}^\lambda \Rightarrow S \mid S > -\delta, \quad (\text{A126})$$

where  $S$  is the steady state distribution of the Halfin-Whitt limit. This convergence implies the tightness of  $\hat{X}^\lambda$ . ■

## 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., “Markov Chains and Stochastic Stability”, Springer-Verlag, London, 1993.
- [10] Whitt W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer-Verlag, New York.