e-companion for:

When Promotions Meet Operations: Cross-Selling and Its Effect on Call-Center Performance

Mor Armony\(^1\)  
Itay Gurvich\(^2\)  

In this e-companion we provide proofs for some of the results stated in the manuscript titled: “When promotions meet operations: Cross-selling and its effect on call-center performance”. Specifically, we provide the sample path construction under the TP policy (in §A), and the key steps in the proof of Theorem 3.1 under Condition 1 (in §B).

Due to space limitations, the proofs of Theorem 3.1 under Conditions 2 and 3 (which are more standard) are relegated to a technical appendix that can be downloaded from the authors’ websites (http://www.stern.nyu.edu/∼marmony and http://www.kellogg.northwestern.edu/faculty/gurvich/personal). In addition, that technical appendix also contains proofs of some auxiliary lemmas that are used in this e-companion as well proofs of all the side results in the main manuscript. To allow the reader to easily relate this e-companion to the online technical appendix, the section numbering in that appendix begins from C. Accordingly, references within this e-companion to results in sections C-G are references to the technical appendix.

We begin the e-companion with a formal construction of the sample paths under the TP\([K]\) control.

Appendix A: Sample-path construction

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

\[
Q_1^\lambda(t) + Z_1^\lambda(t) = Q_1^\lambda(0) + Z_1^\lambda(0) + \tilde{N}_A(t) - \tilde{N}_{D_1}(t),
\]

\[
Z_2^\lambda(t) = Z_2^\lambda(0) - \tilde{N}_{D_2}(t) + 1_{\{K^\lambda \geq 0\}} \left[ N_7 \left( \mu_\lambda \int_0^t Z_1^\lambda(u) 1_{\{0 < Q_1^\lambda(u) \leq K^\lambda\}} du \right) 
+ N_5 \left( \mu_\lambda \int_0^t Z_1^\lambda(u) 1_{\{Q_1^\lambda(u) = 0\}} du \right) \right] 
+ 1_{\{K^\lambda < 0\}} N_8 \left( \mu_\lambda \int_0^t Z_1^\lambda(u) 1_{\{Y_1^\lambda(u) - N_1^\lambda \leq K^\lambda\}} du \right).
\]

Here, we use a time change of the processes \(N_7(\cdot)\) and \(N_5(\cdot)\) to construct service completions with cross-selling candidates that are followed by actual cross-selling. These two processes are used if the threshold is non-negative. In this case, a service completion is followed by a cross-selling offer whenever the queue is equal to 0 or when it is greater than 0 but smaller than the threshold. We use \(N_6(\cdot)\) similarly for the cases in which the threshold is negative. In this case, a service completion is followed by a cross-selling offer if the number of idle servers is less than \(K^\lambda\). The processes \(\tilde{N}_A(t), \tilde{N}_{D_1}(t)\) and \(\tilde{N}_{D_2}(t)\) will be defined shortly.

---

\(^1\) Stern School of Business, New York University, marmony@stern.nyu.edu  
\(^2\) Kellogg School of Management, Northwestern University, i-gurvich@kellogg.northwestern.edu  
\(^3\) This work was first submitted when IG was a graduate student at Columbia University
Also, we have that
\[
Z_1^\lambda(t) = Z_1^\lambda(0) + N_1 \left( \lambda \int_0^t \left[ 1_{\{Z_1^\lambda(u) \leq N^\lambda\}} + 1_{\{Z_1^\lambda(u) > N^\lambda\}} \right] du \right) \\
+ N_3 \left( [Y - Z_2(u)]_{(u) > 0} du \right) \\
- 1_{\{K^\lambda \geq 0\}} \left[ N_5 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) + N_6 \left( (1 - p) \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \right] \\
- 1_{\{K^\lambda > 0\}} N_7 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Q^\lambda(u) \leq K^\lambda\}} du \right) \\
+ 1_{\{K^\lambda < 0\}} N_8 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{N^\lambda \leq Y^\lambda(u) \leq K^\lambda\}} du \right) \\
- 1_{\{K^\lambda < 0\}} N_9 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{K^\lambda < Y^\lambda(u) \leq N^\lambda \}} du \right) + N_10 \left( (1 - p) \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \right].
\]

Here, in addition to the previous processes, we use a time change of \( N_1(\cdot) \) to construct the arrivals that occur when some servers are idle. Both processes \( N_6(\cdot) \) and \( N_{10}(\cdot) \) are used to construct service completions with customers that are not cross-selling candidates when the queue is empty. \( N_6(\cdot) \) is used for non-negative thresholds and \( N_{10}(\cdot) \) is used for negative ones. The process \( N_9(\cdot) \) is used to construct service completions with cross-selling candidates when the queue is empty (for negative thresholds), and, finally, the process \( N_3(\cdot) \) is used to construct cross-selling completions when the queue is positive. Note that the event epochs, in which there is a service completion that is followed by admission of a customer to service, are not modeled above as these transitions keep \( Z_1^\lambda \) unchanged.

We construct the aggregate-arrival process, \( \tilde{N}_A(t) \), and the process of cross-selling completions, \( \tilde{N}_{D_2}(t) \), as follows:
\[
\tilde{N}_A(t) := \tilde{N}_1 \left( \lambda \int_0^t \left[ 1_{\{Z_1^\lambda(u) + Z_2(u) < N^\lambda\}} + 1_{\{Z_1^\lambda(u) + Z_2(u) \geq N^\lambda\}} \right] du \right) + \tilde{N}_2 \left( \lambda \int_0^t \left[ 1_{\{Z_1^\lambda(u) \geq N^\lambda\}} \right] du \right).
\]
\[
\tilde{N}_{D_2}(t) := \tilde{N}_3 \left( \mu_s \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) > 0\}} du \right) + \tilde{N}_4 \left( \mu_s \int_0^t Z_2^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right).
\]

Here, we use \( \tilde{N}_2(\cdot) \) to construct arrivals when all servers are busy and \( \tilde{N}_4(\cdot) \) to construct cross-selling completions when the queue is empty. Finally, we have
\[
\tilde{N}_{D_1}(t) = \tilde{N}_1 \left( \lambda \int_0^t \left[ 1_{\{Z_1^\lambda(u) \leq N^\lambda\}} \right] du \right) + \tilde{N}_5 \left( (1 - p) \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \\
+ \tilde{N}_6 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \\
+ \tilde{N}_7 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{0 < Q^\lambda(u) \leq K^\lambda\}} du \right) \\
+ \tilde{N}_8 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{N^\lambda \leq Y^\lambda(u) \leq K^\lambda\}} du \right) \\
+ \tilde{N}_9 \left( \mu_s \int_0^t Z_1^\lambda(u) 1_{\{K^\lambda < Y^\lambda(u) \leq N^\lambda\}} du \right) \\
+ \tilde{N}_{10} \left( (1 - p) \mu_s \int_0^t Z_1^\lambda(u) 1_{\{Q^\lambda(u) = 0\}} du \right) \\
+ \tilde{N}_{11} \left( \int_0^t \lambda(u) du \right),
\]

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

This construction follows by noting that all input and output processes in the system can be generated through thinning of Poisson processes. The fact that the representation of the queueing dynamics via time changes of Poisson process has a unique solution can be proved as in Theorem 9.2 of [7].
By Lemma 9.4 in [7], there exists a probability space \((\Omega, \mathcal{F}, P)\), a filtration \(\{\mathcal{F}_t\}_{t \geq 0}\) and an 11-dimensional Brownian Motion \((B_1(\cdot), \ldots, B_{11}(\cdot))\) such that the random variable
\[
E_i := \sup_{t \geq 0} \frac{N_i(t) - t - B_i(t)}{\log(2 \vee t)},
\]
has a moment generating function in a neighborhood of the origin and in particular, there exist constants \(c_1, c_2\) and \(\Gamma\), such that, for all \(i = 1, \ldots, 11\), and all \(x \geq 0\),
\[
P\{E_i \geq \Gamma + x\} \leq c_1 e^{-c_2 x}.
\]
We let
\[
E := \sum_{i=1}^{11} E_i.
\]

As we consider only cases in which \(N^\lambda \leq R + \frac{3\lambda}{\mu_a}\), the value of the time change at time \(t\) in each of the equations (A1)-(A3) is bounded by \(c\lambda t\) for some positive constant \(c\), so that we may apply the strong-approximations result.

We now use (A4) to express the dynamics as follows:
\[
Q^\lambda(t) = Q^\lambda(0) + Z_1^\lambda(0) + \lambda t - \mu_s \int_0^t Z_1^\lambda(u)du - Z_1^\lambda(t) + M_{\lambda, Q}^2(t) + O(\log(2 \vee c\lambda t)),
\]
\[
Z_2^\lambda(t) = Z_2^\lambda(0) - \mu_{cs} \int_0^t Z_2^\lambda(u)du + \mu_s \int_0^t Z_1^\lambda(u)du
- \mu_{cs} \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0) + M_{\lambda, z_2}^2(t) + O(\log(2 \vee c\lambda t)),
\]
and
\[
Z_1^\lambda(t) = Z_1^\lambda(0) + \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}}du - \mu_s \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0) + M_{\lambda, z_1}^2(t) + O(\log(2 \vee c\lambda t)).
\]

Here, \(M_{\lambda, Q}^2(\cdot), M_{\lambda, z_1}^2(\cdot)\) and \(M_{\lambda, z_2}^2(\cdot)\) are sums of time changed Brownian motions. For example, if \(K^\lambda > 0\),
\[
M_{\lambda, z_1}(t) = B_1 \left( \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}}du \right)
+ B_3 \left( \mu_{cs} \int_0^t Z_2^\lambda(u)du Z_1^\lambda(0) \right)
- B_5 \left( \mu_s \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0) \right)
- B_6 \left( (1-p) \mu_s \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0) \right)
+ B_7 \left( \mu_{cs} \int_0^t Z_2^\lambda(u)du Z_1^\lambda(0) \right),
\]
where \(B_i(\cdot), i \in \{1, 3, 5, 6, 7\}\), are standard Brownian motions.

Using the Brownian motion strong law of large numbers (see problem 2.9.3. in [6]) and the bound \(c\lambda t\) on the time change values, we have that, uniformly on compact sets,
\[
\left( \frac{M_{\lambda, Q}^2(t)}{\lambda}, \frac{M_{\lambda, z_1}^2(t)}{\lambda}, \frac{M_{\lambda, z_2}^2(t)}{\lambda} \right) \rightarrow (0, 0, 0), \text{ as } \lambda \rightarrow \infty.
\]

Put
\[
T_1^\lambda(t) := \mu_s \int_0^t Z_1^\lambda(u)1_{\{Y^\lambda(u) - N^\lambda > K^\lambda\}}du, \quad T_2^\lambda(t) := \lambda \int_0^t 1_{\{Z_1^\lambda(u) + Z_2^\lambda(u) < N^\lambda\}}du,
\]
\[
T_3^\lambda(t) := \mu_{cs} \int_0^t Z_2^\lambda(u)du Z_1^\lambda(0), \quad T_4^\lambda(t) := \mu_s \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0),
\]
\[
T_5^\lambda(t) := \mu_s \int_0^t Z_1^\lambda(u)du Z_2^\lambda(0) 1_{\{Y^\lambda(u) - N^\lambda \leq K^\lambda\}}du,
\]
and re-write (A6)-(A8) as follows:

\[
Q_\lambda(t) = Q_\lambda(0) + Z_\lambda^1(0) + \lambda t - \mu_s \int_0^t Z_\lambda^1(u)du - Z_\lambda^1(t) + M_{Z_\lambda,Q}(t) \\
+ O(\log(2 \vee c\lambda t)), \quad (A10)
\]

\[
Z_\lambda^2(t) = Z_\lambda^2(0) - \mu_{cs} \int_0^t Z_\lambda^2(u)du + p\mu_s \int_0^t Z_\lambda^1(u)du - T_\lambda^1(t) + M_{Z_\lambda^2}(t) \\
+ O(\log(2 \vee c\lambda t)), \quad (A11)
\]

\[
Z_\lambda^3(t) = Z_\lambda^3(0) + T_\lambda^3(t) - \mu_s \int_0^t Z_\lambda^1(u)du + T_\lambda^3(t) + T_\lambda^2(t) - T_\lambda^3(t) \\
+ M_{Z_\lambda}(t) + O(2 \vee \log(\lambda c t)). \quad (A12)
\]

**Notational conventions and organization of the e-companion and the technical appendix.** Under a $TP[K]$ policy, the process $(\Xi^\lambda(t), t \geq 0)$ defined by $\Xi^\lambda(t) := (Q^\lambda(t), Z_\lambda^1(t), Z_\lambda^3(t))$, is a Markov process. We denote $X$ as the state-space of $\Xi^\lambda(t)$ and use the notation $\xi$ for a general element in $X$. For a given $\xi \in X$ we let $q(\xi)$, $z_2(\xi)$ and $z_3(\xi)$ be its corresponding coordinates. We use $P_\xi \{ \cdot \} = P_\xi \{ \cdot | \Xi^\lambda(0) = \xi \}$ and $E_\xi[\cdot] := E_\xi[\cdot | \Xi^\lambda(0) = \xi]$. If $\nu^\lambda$ is the steady-state distribution of $\Xi^\lambda(t)^d$, we let $P_{\nu^\lambda} \{ \cdot \}$ be the probability with respect to an initial condition that is distributed according to $\nu^\lambda$. $E_{\nu^\lambda}[\cdot]$ is the corresponding expectation. Finally, for $x \in \mathbb{R}$, we let $x^- := \max\{0, -x\}$ and $x^+ := \max\{0, x\}$.

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

**Appendix B: Asymptotic optimality under Condition 1**

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

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

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

and

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

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

**Corollary B.1** Suppose that Assumptions 3.1 and 3.2 hold. If, in addition, $N^\lambda_2 - R \gg N^\lambda_1 - R$, then the following is asymptotically optimal in the sense of Definition 3.2:

- **Staffing:** Staff with $N^\lambda_2$ agents.
- **Control:** Use $TP[K^\lambda]$ with $K^\lambda \in [0, [\lambda \tilde{W}])$.

\footnote{In Lemma F.1 we prove the existence of a steady-state distribution under $TP[K]$ assuming $N^\lambda > \lambda / \mu_s$.}
Proof: By Little’s law:

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

Equation (A13) now implies that

\[
\limsup_{\lambda \to \infty} \frac{E[W^\lambda]}{W^\lambda} \leq 1,
\] (A16)

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

To establish optimality, recall that \(E[Z^\lambda_t] = N^\lambda_2 - E[Z^\lambda_0] - E[I^\lambda]\). Since, by Little’s law \(E[Z^\lambda_0] = \lambda/\mu_s := R\), (A14) implies that

\[
\frac{\mu_{cs}E[Z^\lambda_0]}{\mu_{cs}(N^\lambda_2 - R)} \to 1 \text{ as } \lambda \to \infty.
\] (A17)

For each \(\lambda\), \(r_{\mu_{cs}}(N^\lambda_2 - R) - (C^\lambda(N^\lambda_2) - C^\lambda(R))\) constitutes an upper bound for the (centered) optimal value of the cross-selling problem given by \(V^+(\lambda) := \sup_{\pi \in \Pi, N \in \mathbb{Z}_+} V^\lambda(\pi, N)\); see formulation (6) and the discussion below it as well as Definition 3.2.

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

\[
\frac{r_{\mu_{cs}}E[Z^\lambda_0] - (C^\lambda(N^\lambda_2) - C^\lambda(R))}{r_{\mu_{cs}}(N^\lambda_2 - R) - (C^\lambda(N^\lambda_2) - C^\lambda(R))} \to 1 \text{ as } \lambda \to \infty.
\]

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

Hence, the sequence of pairs \(\{(N^\lambda_2, TP[K^\lambda])\}\) asymptotically achieves the upper bound and is, in particular, asymptotically optimal.

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

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

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

We first provide an informal outline of the proof:

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

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

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

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

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

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

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

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

Define the set

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

(A20)

We will be using the set $\Omega^\ast(\cdot, \cdot, \cdot, \cdot)$ in various proofs and set the parameters $\delta, \lambda, T, \Theta$ according to the need of the specific proof. In any of these cases, we will choose the parameters in a way that will guarantee that the set of sample paths that we consider has sufficiently large probability; see Lemma F.2. In the current proof it is important that we use $T/\lambda$ instead of $T$. That is we focus on a small time interval and characterize the behavior of the queue length there. To that end, plugging equation (A8) into equation (A6), and using the fact that $Z^\lambda_1(t) = N^\lambda - Z^\lambda_1(t)$ whenever $Q^\lambda(t) > 0$, we have that on $\Omega^\ast(\delta, \lambda, T/\lambda, \Theta/2)$

$$
Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \lambda(t \wedge \tau^\lambda) - \mu_1 \int_0^{t \wedge \tau^\lambda} Z^\lambda_1(u)du - \mu_{cs} \int_0^{t \wedge \tau^\lambda} N^\lambda - Z^\lambda_1(u)du + \delta\lambda t + \Theta/2.
$$

(A21)

The following lemma provides a handle on the process $Z^\lambda_1(t)$ that, in turn, allows us to characterize the behavior of $Q^\lambda(t)$.

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

$$
\sup_{0 \leq t \leq T/\lambda} \left( Z^\lambda_1(t) - \frac{\lambda}{\mu_1} \right)^- \leq 2\epsilon\lambda,
$$

(A22)

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

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

$$
\lambda - \mu_1 Z^\lambda_1(t) - \mu_{cs} Z^\lambda_2(t) = \mu_1 \left( Z^\lambda_1(t) - \frac{1}{\mu_1} \right)^- - \mu_1 \left( Z^\lambda_1(t) - \frac{1}{\mu_1} \right)^+ - \mu_{cs} \left( \frac{1+\beta}{\mu_2} - Z^\lambda_1(t) \right)
$$

$$
\leq \mu_1 \epsilon - \mu_1 \left( \frac{1+\beta}{\mu_2} - Z^\lambda_1(t) \right),
$$

(A23)

where $\mu = \mu_1 \wedge \mu_{cs}$. Hence,

$$
Q^\lambda(t \wedge \tau^\lambda) \leq Q^\lambda(0) + \left( \delta\lambda + \mu_1 \epsilon \lambda - \frac{\beta}{\mu_2} \lambda \right) (t \wedge \tau^\lambda) + \Theta/2.
$$

(A24)

Let $\eta := -\left( \mu_1 \epsilon + \delta - \frac{\beta}{\mu_2} \right)$, choose $\epsilon$ and $\delta$ small enough so that $\eta > 0$ and let $t^* := 2\Theta/\eta$. Then, as $\tau^\lambda$ is the first time that $Q^\lambda(t)$ goes below $Q^\lambda(0) - 3\Theta/2$ we must have that $\tau^\lambda \leq t^* / \lambda$ on $\Omega^\ast(\delta, \lambda, T/\lambda, \Theta/2)$.

Note that $\tau^\lambda$ is a random time but not necessarily a stopping time. Our arguments are sample-path arguments so that whether or not $\tau^\lambda$ is a stopping time is immaterial for our proofs.
The arguments that lead to equation (A24) can be modified to show that, on \( \Omega^* (\delta, \lambda, T/\lambda, \Theta/2) \) the queue will remain below \( Q^\lambda (0) - \Theta \) after it reaches \( Q^\lambda (0) - 3\Theta/2 \) for the first time. That is, that \( Q^\lambda (t) \leq Q^\lambda (0) - \Theta \) for all \( t^*/\lambda \leq t \leq T/\lambda \). In particular, on \( \Omega^* (\delta, \lambda, T/\lambda, \Theta/2) \),

\[
\sup_{\xi \in A^\lambda \cap \{q(\xi) > 2\Theta\}} \frac{Q^\lambda (2t^*/\lambda)^2 - q(\xi)^2}{q(\xi)} \leq \sup_{\xi \in A^\lambda \cap \{q(\xi) > 2\Theta\}} \frac{(q(\xi) - \Theta)^2 - q(\xi)^2}{q(\xi)} \leq -\Theta, \tag{A25}
\]

where \( A^\lambda \) is as in (A19). Since As \( Q^\lambda (t) \leq Q^\lambda (0) + A^\lambda (t) \), it is also the case that

\[
(Q^\lambda (t))^2 - (Q^\lambda (0))^2 \leq 2Q^\lambda (0) A^\lambda (t) + (A^\lambda (t))^2, \tag{A26}
\]

so that, removing the condition that \( \xi \in A^\lambda \), we have that

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

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

\[
\sup_{\xi \in A^\lambda \cap \{q(\xi) > 2\Theta\}} \frac{E_t [Q^\lambda (2t^*/\lambda)^2] - q(\xi)^2}{q(\xi)} \leq -\Theta^2. \tag{A28}
\]

The crude inequality (A26) also guarantees that

\[
\sup_{\xi \in A^\lambda \cap \{q(\xi) \leq 2\Theta\}} E [Q^\lambda (2t^*/\lambda)^2 - q(\xi)^2] \leq c_{10}, \tag{A29}
\]

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

\[
q(\xi)^2 - E_t [Q^\lambda (2t^*/\lambda)^2] \geq \frac{\Theta}{2} q(\xi) - c_{11} + \left( c_{11} - \frac{\Theta}{2} q(\xi) - E_t [Q^\lambda (2t^*/\lambda)^2 - q(\xi)^2] \right) 1_{\{\xi \notin A^\lambda\}}, \tag{A30}
\]

where \( c_{11} = \Theta^2 + c_{10} \). By definition of stationarity we have that \( E_{\nu^\lambda} [Q^\lambda (0)^2] = E_{\nu^\lambda} [Q^\lambda (2t^*/\lambda)^2] \) where \( \nu^\lambda \) is the steady-state distribution of the process \( (\Xi^\lambda (t), t \geq 0) \), and in particular,

\[
0 = E_{\nu^\lambda} \left[ q(\xi)^2 - E_t [Q^\lambda (2t^*/\lambda)^2] \right] \nu^\lambda (d\xi).
\]

When applied to (A30), this yields

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

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

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

\[
P_{\nu^\lambda} \left\{ \Xi^\lambda (0) \notin A^\lambda \right\} \leq c_3 e^{-c_4 \lambda / \log (2\sqrt{cT})}, \tag{A32}
\]

for all \( \lambda \) large enough.
Lemma B.3: Under the conditions of Theorem B.1

$$\lim_{\lambda \to \infty} \sup \lambda^{-m} E_{\nu^\lambda} \left[ \left( \frac{Q^\lambda(0)}{\lambda} \right)^m \right] < \infty,$$

for any integer \(m \geq 1\).

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

$$\lim_{\lambda \to \infty} \sup \lambda^{-m} E_{\nu^\lambda} \left[ (Q^\lambda(0))^m 1 \{ \xi \notin \mathcal{A}_k^\lambda \} \right] = 0. \quad (A33)$$

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

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

This concludes the proof of Proposition B.1.

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

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

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

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

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

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

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

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

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

Lemma B.4: Fluid Limits) Consider a finite interval \([0, T]\) and suppose that

$$\left( \frac{Q^\lambda(0)}{\lambda}, \frac{Z_1^\lambda(0)}{\lambda}, \frac{Z_2^\lambda(0)}{\lambda} \right) \Rightarrow (\bar{Q}(0), \bar{Z}_1(0), \bar{Z}_2(0)).$$
Then, under the assumptions of Theorem B.1, the sequence \( \left( \frac{Q^\lambda(t)}{\lambda}; \frac{Z_1^\lambda(t)}{\lambda}, \frac{Z_2^\lambda(t)}{\lambda}; \frac{T_i^\lambda(t)}{\lambda}, i = 1, \ldots, 5 \right) \) is tight in \( D[0, T] \) and every subsequence \( \{\lambda^k\}_{k \geq 1} \) contains a further subsequence that converges in probability to some limit uniformly on compact sets. Moreover, any such limit process

\[
(\tilde{Q}(t); \tilde{Z}_1(t); \tilde{Z}_2(t); \tilde{T}_i(t), i = 1, \ldots, 5),
\]

satisfies the following equations:

\[
\begin{align*}
\dot{\tilde{Z}}_1(t) + \tilde{Q}(t) &= \tilde{Q}(0) + \tilde{Z}_1(0) + t - \int_0^t \mu_s \tilde{Z}_1(u)du, \\
\dot{\tilde{Z}}_2(t) &= \tilde{Z}_2(0) - \mu_s + \int_0^t \mu_s \tilde{Z}_1(u)du - \tilde{T}_1(t), \\
\dot{\tilde{Z}}_1(t) &= \tilde{Z}_1(0) + \tilde{T}_2(t) - \mu_s \int_0^t \tilde{Z}_1(u)du + \tilde{T}_3(t) + \tilde{T}_4(t) - \tilde{T}_5(t), \\
\dot{\tilde{T}}_1(t)1_{(\tilde{Q}(t) > 0)} &= p \mu_s \tilde{Z}_1(t), \\
\dot{\tilde{T}}_2(t)1_{(\tilde{Z}_1 + \tilde{Z}_2 < \frac{1 + \beta}{\mu_s})} &= 1, \\
\dot{\tilde{T}}_3(t)1_{(\tilde{Q}(t) > 0)} &= \mu_s \tilde{Z}_1(t), \\
\dot{\tilde{T}}_4(t)1_{(\tilde{Q}(t) > 0)} &= \mu_c s \tilde{Z}_2(t), \\
\dot{\tilde{T}}_5(t)1_{(\tilde{Q}(t) > 0)} &= 0.
\end{align*}
\]

**Lemma B.5** Fix \( \epsilon > 0 \) and assume \( 0 < \beta \leq \frac{\mu_s}{\mu_c} \). Let \( (\tilde{Q}(t); \tilde{Z}_1(t); \tilde{Z}_2(t); \tilde{T}_i(t), i = 1, \ldots, 5) \) be a non-negative process that satisfies equations (A36)-(A43). Then, there exists \( t^0(\epsilon) \) (independent of \( \tilde{Z}_1(0) \) and \( \tilde{Z}_2(0) \)), such that for all \( t \geq t^0(\epsilon) \),

\[
\left| \tilde{Z}_1(t) - \frac{1}{\mu_s} \right| \leq \epsilon.
\]

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

\[
\tilde{I}(t) := \frac{1 + \beta}{\mu_s} - \tilde{Z}_1(t) - \tilde{Z}_2(t) \leq \epsilon,
\]

for all \( t \geq t^* \).

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

\[
\limsup_{\lambda \to \infty} P_{\nu^\lambda}\left( \frac{I^\lambda(t)}{\lambda} > 2\epsilon \right) = 0.
\]

Since \( I^\lambda(t) \leq N^\lambda \leq \lambda/\mu_s + \lambda p/\mu_c \) we have that

\[
\limsup_{\lambda \to \infty} \frac{E_{\nu^\lambda}[I^\lambda(t)]}{\lambda} \leq 3\epsilon,
\]

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

\[
\limsup_{\lambda \to \infty} \frac{E_{\nu^\lambda}[I^\lambda(0)]}{\lambda} \leq 3\epsilon.
\]
Since \( \epsilon \) was arbitrary, this concludes the proof of the theorem for the case \( \beta > 0 \). The case \( \beta = 0 \) is proved in \( \S F \). ■