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

Mor Armony\textsuperscript{1} 
Itay Gurvich\textsuperscript{2}

Submitted July 28, 2006; Revised August 31, 2007

Abstract

We study cross-selling operations in call centers. The following question is addressed: How many customer service representatives are required (staffing) and when should cross-selling opportunities be exercised (control) in a way that will maximize the expected profit of the firm while maintaining a pre-specified service level target. We tackle these questions by characterizing scheduling and staffing schemes that are asymptotically optimal in the limit, as the system load grows to infinity. Our main finding is that a threshold priority (TP) control, in which cross-selling is exercised only if the number of callers in the system is below a certain threshold, is asymptotically optimal in great generality. The asymptotic optimality of TP reduces the staffing problem to the solution of a simple deterministic problem, in some cases, and to a simple search procedure in others. Our asymptotic approach establishes that our staffing and control scheme is near-optimal for large systems. In addition, we numerically demonstrate that TP performs extremely well even for relatively small systems.

Acknowledgment

We thank Assaf Zeevi for his helpful suggestions on the proofs of asymptotic optimality.

\textsuperscript{1}Stern School of Business, New York University, marmony@stern.nyu.edu
\textsuperscript{2}Graduate School of Business, Columbia University, ig2126@columbia.edu
1 Introduction

Call Centers are in many cases the primary channel of interaction of a firm with its customers. Historically, call centers were mostly considered a service delivery channel. Typically, service driven call centers plan their operations based on delay related performance targets. Examples of such performance measures are average speed of answer (ASA), the fraction of customers whose call is answered by a certain time and the percentage of customer abandonment. These operational problems have gained a lot of attention in the literature.

Most companies, however, are not purely service providers. Rather, customer service is a companion to one or several main products. For example - the core business of computer hardware companies, like Dell, is to sell computers. They do, however, have a call center whose main purpose is to provide customer support after the purchase. Most banks nowadays have call centers that give customer support while their main business is selling financial products. For these companies, the inbound call center can be a natural and very convenient sales channel. As opposed to outbound tele-marketing calls, the interaction in the inbound call center is initiated by the customer. Once the customer calls the center, a sales opportunity is generated and the agent might choose to exercise this cross-selling opportunity by offering the customer an additional service or product.

From a marketing point of view, a call center has a potential of becoming an ideal sales environment. Modern Customer Relationship Management (CRM) systems have dramatically improved the information available to Customer Service Representatives (CSR’s) about the individual customer in real time. Specifically, in call centers, once the caller has been identified, the CRM system can inform the agent regarding this customer’s transaction history, her value to the firm and specific cross-selling opportunities. As a result, cross-sales offerings can be tailored to the particular customer, making modern call centers a perfect channel for customized sales. Many companies have identified the revenue potential of inbound call centers. Indeed, as suggested by a recent McKinsey report [13], call centers generate up to 25 percent of total new revenues for some credit card companies and up to 60 percent for some telecom companies. Moreover, [13] estimates that cross-selling in a bank’s call center can generate a significant revenue, equivalent to 10% of the revenue generated through the bank’s entire branch network.

Although the benefits of running a joint service and sales call center seem clear, there are various challenges involved in operating such a complex environment. An immediate implication of incorporating sales
is the increase in customer handling times caused by cross-sales offerings. Unless staffing levels are adjusted, the increased handling times will inevitably lead to service level degradation in terms of waiting times experienced by the customers. Does this imply that more cross-selling will necessarily lead to deterioration in service levels? What are the appropriate operational tradeoffs that one should examine in the context of a combined service and sales call center?

In purely service driven call centers, the manager typically attempts to minimize the staffing level while maintaining a pre-determined performance target. Hence, in this pure service context the operational tradeoff is clear: Staffing Vs. Service Level. When sales and promotions are introduced, however, one should add the potential revenue from promotions as a third component to this tradeoff. Clearly, if the potential revenue is very high in comparison to the staffing cost, it would be in the interest of the company to increase the staffing level and allow for as much cross-selling as possible. In this case, we show that with appropriate staffing, incorporating cross-selling actually reduces delays. There are cases, however, where the relation between staffing costs and potential revenues is more intricate and a careful analysis is required.

Beyond staffing, in a cross-selling environment there is also a component of dynamic control of incoming calls and cross-sales offerings. Specifically, the call center manager needs to determine when the agent should exercise a cross-selling opportunity. This decision should take into account not only the characteristics of the customer in service but also the effect on the waiting times of other customers. For example, in order to satisfy a waiting time target, it would be natural to stop all promotion activities in the presence of heavy congestion. Indeed, a common heuristic, which is used in practice to determine when to exercise a cross-selling opportunity, is to cross-sell upon service completion only when the number of callers in the queue is below a certain threshold. Optimal rules, however, are typically hard to find and to implement. The staffing and control issues are strongly related since even with seemingly adequate staffing levels, the actual performance might be far from satisfactory when one does not make a careful choice of the dynamic control. Yet, because of the complexity involved in addressing both issues combined, they have been typically addressed separately in the literature. To our knowledge, this paper is the first to consider the staffing and dynamic control in a cross-selling environment jointly, in a single, common framework.

The purpose of this work is to carefully examine operational tradeoffs that are critical in the cross-selling environment. This is done by specifying how to adjust the staffing level and how to choose the control in order to balance staffing costs and cross-selling revenue potential while satisfying quality of service constraints associated with delay performance. Specifically, we provide joint staffing and dynamic
control rules as explicit functions of the quality of service constraints, the potential value of cross-selling and the staffing costs. The control we propose is a Threshold Priority (TP) rule in which cross-selling is exercised only when the number of callers in the system is below a certain threshold. In contrast with the commonly used heuristics we identify cases in which cross-selling should not be exercised even if there are idle agents in the system, in anticipation for future arrivals.

To summarize, we contribute to the existing literature in several dimensions:

1. From a modeling perspective, we propose a realistic two-phase service model for cross-selling in call centers, in which cross-selling decisions are made at the end of the service phase, after gathering information about the caller.

2. From a practical perspective - with the objective of maximizing profits while satisfying commonly used quality of service constraints - we propose a simple and practical Threshold Priority (TP) policy together with a staffing rule and rigorously establish their near-optimality. The qualities of the TP policy include:

   (a) It is based only on the total number of customers in the system, rather than the more elaborate two-dimensional state-space of the system. This is useful in systems where the information about the phase of the service is available to each individual agent, but not to the system as a whole.

   (b) The simplicity of this policy has allowed us to significantly reduce the complexity of the staffing problem.

   The staffing rule we propose is simple, easy to implement and reveals much about the regime at which the center should operate: The Cross-Selling versus The Service regime.

3. Methodologically, we are the first to use a notion of constrained Lyapunov functions to establish steady-state convergence of queueing systems. This idea has been since applied in Gurvich et. al. [17] and has been formalized by Gurvich and Zeevi [19] where it is applied to queueing networks.

The rest of the paper is organized as follows: We conclude the introductory part with a literature review. §2 provides the problem formulation. §3 outlines the main results of the paper through an informal description of our proposed solution for the cross-selling problem. We formally introduce the cross-selling
problem in §4 where we define the asymptotic optimality framework. The asymptotic optimality results are stated in §4 and §5. In §6 we present some numerical results to support our proposed solution. The paper is concluded in §7 with a discussion of the results and directions for future research.

For expository purposes, our approach in the presentation of the results is to state them formally and precisely in the body of the paper, together with some supporting intuition, while leaving some of the formal proofs to the technical appendix [5]. Accordingly, all proofs that do not appear in the body of the paper are deferred to the technical appendix.

1.1 Literature Review

A successful and comprehensive treatment of cross-selling implementation in call centers would clearly require an inter-disciplinary effort combining knowledge from marketing and operations management as well as human resource management and information technology. An extensive search of the literature shows, however, that while the marketing literature on this subject is quite rich, very little has been done from the operations point of view (the reader is referred to Akşin and Harker [1] for a survey of some of the marketing literature).

Although the operations literature on this subject is scarce, the topic of cross-selling has received some attention. In the context of cross-selling in call centers, a significant contribution is due to Akşin with different co-authors. In Akşin and Harker [1] the authors consider qualitatively and empirically the problems of cross-selling in banking call centers. They also suggest a quantitative framework to evaluate the effects of cross-selling on service levels, using a processor sharing model, but they do not attempt to find optimal control or staffing levels. Örmeçi and Akşin [27], on the other hand, do pursue the goal of determining the optimal control, while assuming that the staffing level is given. In their framework, customers’ cross-selling value follows a certain distribution. The realization of this value can be observed by the call center before the cross-selling offer is made. Hence, the agent can base the decision on the actual realization of this value and not only its expected value. However, due to computational complexity, the results in [27] are limited to multi-server loss systems (customers either hang-up or are blocked if their call cannot be answered right away) and to structural results that are then used to propose a heuristic for cross-selling. Günes and Akşin [16] analyze the problem of providing incentives to agents in order to obtain certain service levels and value generation goals. This is indeed a critical issue in cross-selling environments where the decision of whether
to cross-sell or not is often made at the discretion of the individual agents.

Simplicity of the dynamic control is clearly an important factor for a successful implementation of cross-selling. The simplicity of the control might result, however, in decreasing revenues from cross-selling. For example, it is intuitive that one can increase revenues by allowing the control to be based on the identity of the individual customer in addition to the number of customers in the system. Byers and So [11, 12] examine the value of customer identity information by comparing cross-selling revenues under several control schemes that differ by the amount of the information they use. Exact analysis is performed for the single server case in [12] and numerical results are given for the multi-server case in [11].

To position our paper in the context of the literature introduced above, note that our previous models have considered cross-selling decisions that are made upon customer assignment to an agent. Our two-phase service model allows this decision to be postponed until the end of the service phase when more information about the caller has been gathered. Thus, this model, reflects the reality in most call-centers where cross-selling decisions are not made a-priori in the beginning of the call. Our model is realistic also in terms of its relatively mild restriction on the queueing-system model. Indeed, while single-server or loss-system assumptions are made in the existing literature for tractability purposes, we consider a realistic model with many-servers, infinite buffers and commonly-used quality-of-service constraints. Also note that our paper is the first to consider how to optimally choose both the staffing level and the control scheme in a cross-selling environment. If the staffing decision is ignored and the staffing level is assumed to be fixed, the only relevant tradeoff is between service level (expressed in terms of delay) and the extent to which cross-selling opportunities are exercised. In this setting then, more cross-selling necessarily causes service level degradation. Moreover, the existing literature suggests that, when the staffing level is assumed fixed, it is difficult to come up with simple and practical control schemes for cross-selling. As we show in this paper, however, when one adds the staffing component along with asymptotic analysis, the solution may be simplified tremendously while emphasizing that a lot of cross-selling does not necessarily lead to low service levels.

A follow-up paper [17] uses the results of the current paper to study the impact of a heterogenous pool of customers on the structure of asymptotically optimal staffing and control schemes. It also investigates the value of customer segmentation in such an environment.

Our solution approach follows the many-server asymptotic framework, pioneered by Halfin and Whitt [20]. In particular, we follow the asymptotic optimality framework approach first used by Borst et al. [10],
and adapted later to more complex settings ([3], [4], [6], [7], [18] and [24]). The asymptotic regime that we use has been shown to be extremely robust also for relatively small systems (see Borst et al. [10]); Consistent with this finding we give a strong numerical evidence to support the claim that this robustness is also typical in our setting. We note however, that the existing methods of establishing steady-state convergence in this asymptotic framework were not sufficient to use in our framework. Instead, we introduce a proof methodology that was later formalized in Gurvich and Zeevi [19] through the notion of Constrained Lyapunov Functions.

To conclude this review, we mention that, while outside the context of call centers, there is a stream of operations management literature that deals with the implications of cross-selling on the inventory policy of a firm. Examples are the papers by Aydin and Ziya [8] and Netessine et. al. [26].

2 Problem Formulation

Consider a call center with calls arriving according to a Poisson process with rate \( \lambda \). An agent-customer interaction begins with the service phase, whose duration is assumed to be exponentially distributed with rate \( \mu_s \). Upon service completion, if cross-selling is exercised, this interaction will enter a cross-selling phase, whose duration is assumed to be exponentially distributed with rate \( \mu_{cs} \). If cross-selling is not exercised, either intentionally or due to the customer’s refusal to listen to a cross-selling offer, the customer leaves the system. It is assumed that all inter-arrival, service and cross-selling times are independent and that the call center has an infinite buffer.

Potentially, not all customers are cross-selling candidates. Indeed, the center might segment the customer population into two segments such that only a portion \( \bar{p} \) of the customers are cross-selling candidates. This segmentation might be updated based on information gathered during the service phase of the call. Moreover, even if an agent decides to cross-sell to a caller, she will not necessarily agree to listen to the cross-selling offer. We assume that a customer that is presented with the option to listen to a cross-selling offer will agree to do so with probability \( \bar{q} > 0 \). Assuming that all customers are statistically identical, we have that \( p = \bar{p}\bar{q} \) is the probability that a customer is a cross-selling candidate and agrees to listen to the cross-selling offer. The combined parameter \( p \) is sufficient for our analysis so that we will not make additional references to the parameters \( \bar{p} \) and \( \bar{q} \). We assume that a cross-selling offer has an expected revenue of \( r \), and revenues from different customers are independent. A schematic illustration of the system is given in
For simplicity we will say that a customer is in phase 1 of the customer-agent interaction if he is in the service phase and in phase 2 if he is in the cross-selling phase. Let $\pi$ be a control policy which determines upon a phase 1 completion of a cross-selling candidate whether or not to exercise this cross-selling opportunity. Accordingly, we let $Z^\pi_i(t)$ be the number of servers providing phase $i$ service at time $t$, $i = 1, 2$. Then, $Z^\pi(t) = Z^\pi_1(t) + Z^\pi_2(t)$ is the total number of busy agents (CSRs) at time $t$, and $I^\pi(t) = N - Z^\pi(t)$ is the number of idle agents at time $t$ in a system with $N$ agents. Also, let $Q^\pi(t)$ be the number of customers waiting in queue at time $t$ and $Y^\pi(t)$ be the overall number of customers in the system at time $t$, that is $Y^\pi(t) = Z^\pi(t) + Q^\pi(t)$. We denote by $W^\pi(t)$ the virtual waiting time encountered by a customer that arrives to the system at time $t$, and by $P^\pi(cs)(t)$ the probability that cross-selling will be exercised for a customer that arrives at time $t$ (that is, the probability that the customer will be asked to listen to a cross-selling offer and s/he will agree). In all of the above, we omit the time index $t$ when referring to steady state variables. Also, we omit the superscript $\pi$ whenever the control is clear from the context. Note that under any stationary policy, all transition rates in the system can be determined using the number of agents busy providing either phase of service and the queue length. In particular, $S^\pi(t) = \{Z^\pi_i(t), i = 1, 2; Q^\pi(t)\}$ is a sufficient state descriptor for a Markovian characterization of the system under any stationary control. The profit maximization problem formulation we consider is as follows:
maximize \( r\lambda P^\pi(cs) - C(N) \)

subject to \( E[W^\pi] \leq \bar{W} \), \( N \in \mathbb{Z}_+, \pi \in \Pi(\lambda, \mu_s, \mu_{cs}, N) \).

(1)

Here the average steady-state waiting time is constrained to be less than a pre-determined bound \( \bar{W} \). We assume that the staffing cost function, which we denote by \( C(\cdot) \), is convex increasing in the staffing level \( N \). Further assumptions are made on the cost function in §4, where we construct our asymptotic framework. Note that customers do not abandon, or balk, nor are they being blocked.

The control policy \( \pi \) is picked from the following set of admissible controls \( \Pi(\lambda, \mu_s, \mu_{cs}, N) \).

**Definition 2.1 Admissible Controls:** Given a staffing level \( N \), and parameters \( \lambda, \mu_s, \mu_{cs} \), we say that \( \pi \) is an admissible policy if it is non-preemptive, non-anticipative and

\[
\lim_{t \to \infty} \frac{E[Q_\pi(t)]}{t} \to 0.
\]

(2)

Loosely speaking, \( \Pi(\lambda, \mu_s, \mu_{cs}, N) \) is the set of stabilizing policies under the given parameters. Definition 2.1 takes into account the fact that the set of admissible policies depends on the parameters of the model through the stability conditions of the system. For simplicity of notation, when the parameters \( \lambda, \mu_s \) and \( \mu_{cs} \) are fixed, we will omit them from the notation and instead use just the notation \( \Pi(N) \). \( N \) will also be omitted whenever the staffing level is clear from the context. One should note that we used the maximization formulation (1) although the maximum need not exist. We choose the word “maximize” for convenience of presentation while formally referring to taking the supremum over all staffing levels and admissible policies.

Standard stability considerations imply that \( R := \lambda/\mu_s \) constitutes a lower bound on possible staffing levels. To allow for a more refined analysis it makes sense to normalize the cost around its lower bound. Hence, one may re-write (1) as follows:

maximize \( r\lambda P^\pi(cs) - (C(N) - C(R)) \)

subject to \( E[W^\pi] \leq \bar{W} \)

\( N \in \mathbb{Z}_+, \pi \in \Pi(N) \),

(3)

where the only change from (1) is the normalization of the cost around the constant \( C(R) \).

The following is an immediate consequence of Little’s Law and Markov Chain Ergodic theorems.
Lemma 2.1  For any $\pi \in \Pi(N)$ that admits a stationary distribution we have that

1. $E[Z_1^\pi] = R$, and

2. $\lambda P^\pi(cs) = \mu_{cs} \cdot E[Z_2^\pi] = \mu_{cs} \cdot (E[Z^\pi] - R) \leq \mu_{cs} \cdot (N - R) \wedge \frac{\lambda p}{\mu_{cs}},$

where for two real numbers $x$ and $y$, $x \wedge y = \min\{x, y\}$.

Remark 2.1  Lemma 2.1 implies that under any policy $\pi$, $E[Z_{2, \lambda, \pi}^\pi] \leq \frac{\lambda p}{\mu_{cs}}$. In particular, this implies that when performing a search for the optimal staffing level one can guarantee asymptotically optimal performance by considering only staffing levels that are less than or equal to $R + \frac{\lambda p}{\mu_{cs}}$.

Having Lemma 2.1 we can re-write (3) as

$$\begin{align*}
\text{maximize} & \quad r \mu_{cs} (E[Z^\pi] - R) - (C(N) - C(R)) \\
\text{subject to} & \quad E[W] \leq \bar{W}, \\
& \quad N \in \mathbb{Z}_+, \pi \in \Pi(N).
\end{align*}$$

(4)

Within the set of policies, we propose the following control:

Definition 2.2  The Threshold Priority (TP) control is defined as follows:

- An arriving customer will enter service immediately upon arrival if there are any idle agents.
- Upon a customer departure, the customer at the head of the queue will be admitted to service if the queue is non-empty.
- An agent that completes a phase 1 service with a customer at a time $t$ will exercise cross-selling if this customer is a cross-selling candidate and $(Y(t) - N) \leq K$ (where $K$ is a pre-determined integer).

For brevity, we use the notation $TP[K]$ to denote $TP$ with threshold $K$ (where $K$ may take negative as well as positive values). One should note the following: If $K > 0$, $TP[K]$ is a control that uses a threshold on the number of customers in queue. Specifically, upon service completion with a cross-selling candidate, the agent will exercise cross-selling if the number of customers in queue is at most $K$. Conversely, if $K \leq 0$, $TP[K]$ is a control that uses a threshold on the number of idle agents. Specifically, upon service completion
with a cross-selling candidate, the agent will exercise cross-selling if the number of idle agents is at least $|K|$.

Note that $TP[K]$ uses only information on the overall number of customers in the system at the time of service completion. In particular, $TP[K]$ is a stationary policy so that the system can be modelled as a Markov chain. Specifically, we claim that the state descriptor $\{Z_2(t), Y(t)\}$ where $Z_2(t)$ is the number of servers working on phase 2 service (cross-selling) and $Y(t)$ is the overall number of customers in system, is sufficient for a Markovian description of the system under $TP$. Indeed, since $TP$ disallows a positive queue when there are idle agents, we have that $Q(t) = \lfloor Y(t) - N \rfloor^+$ and $Z_1(t) + Z_2(t) = N - \lfloor Y(t) - N \rfloor^-$. Since our optimization problem is given in steady state terms it is necessary to have the following simple result for the Markov chain we have just constructed.

**Lemma 2.2** Fix $\lambda$. Then, for any staffing level, $N$, with $N > R$, and for any $K \geq -N$, steady-state for the Markov chain $\{Z_2(t), Y(t)\}$ exists under $TP[K]$.

Note that this lemma implies that (2) holds under $TP[K]$ and in particular that $TP[K]$ is admissible. Lemma 2.2 is rather intuitive. We claim that since all cross-selling is stopped whenever the number of customers in the system crosses the level $K$, the number of customers in the system cannot grow without bound. Indeed, if this number grows without bound, by the very definition of $TP[K]$, there will be a time after which no cross-selling will be exercised and in particular, there will be a time, starting from which, the system will behave like a FCFS $M/M/N$ system with arrival rate $\lambda$ and service rate $\mu_s$. However, since $N > R$ we know that from that time forward the number of customers in system cannot grow without bound, thus leading to a contradiction.

**Discussion of modeling assumptions:** With respect to the probability that a caller would be willing to listen to a cross-selling offer, it is plausible that in reality, this probability depends on the customer experience up to that point (such as his waiting time, service time, service quality, etc.). This dependence introduces analytical complications because the state space required to describe such a system is very large (in particular, it would need to include for each customer her current waiting time and service time). Given this complexity we assume in this paper that the probability of agreeing to listen to a cross-selling offering is independent of the customer service experience. This assumption is reasonable for systems in which waiting times are not too long and service quality is of uniform level. The independence assumption is relaxed in Gurvich et al. [17] where a cruder form of analysis is performed.
With respect to the problem formulation (3), as an alternative to the average-waiting-time constraint, one might consider the commonly used Quality of Service (QoS) constraint of the form \( P\{W > \bar{W}\} \leq \delta \). That is, one requires that at least a fraction \( 1 - \delta \) of the customers will be answered within \( \bar{W} \) units of time. Hence, it is worthwhile mentioning that all the insights of our analysis go through for constraints of this form under the assumption that customers are served in a First Come First Served (FCFS) manner. In particular, the structure of the asymptotically optimal staffing and control schemes we propose remains the same under both types of constraints.

3 The Proposed Solution

In §4 we present our asymptotic framework and a formal description of the proposed staffing and control schemes. The purpose of this section is to give an informal description of this scheme which is sufficient for practical purposes.

Our proposed solution includes a staffing recommendation \( N \) and the TP policy (recall Definition 2.2) with a threshold \( K \). We claim that TP performs extremely well in great generality. Moreover, focusing on this particular family of control policies reduces the problem of staffing and control into a simple search over two parameters; for each pair \((N, K)\) performance evaluation can be performed by solving for the steady-state distribution or via simulation. Since this task is not too computationally expensive, this approach is sufficient for all practical purposes. Nevertheless, we propose some asymptotic insights that simplify this search even further.

To simplify the search over the set of parameters \((N, K)\) we first note that it follows from Corollary B.2 that for every fixed \( N \) the range of threshold values to consider \([-N, \lambda \bar{W}/2]\). Secondly, a lower bound for the optimal staffing level \( N \) is

\[
N_1 = \arg \min \{ N \in \mathbb{Z}_+ : E[W_{\lambda,\mu_s}^{FCFS}(N)] \leq \bar{W} \},
\]

(5)

where \( W_{\lambda,\mu_s}^{FCFS}(N) \) is the steady state waiting time in an \( M/M/N \) system with arrival rate \( \lambda \), service rate \( \mu_s \) and FCFS service discipline. This is straightforward, since one cannot satisfy the constraint in the presence of cross-selling if the same constraint cannot be satisfied in the absence of cross-selling. Finally, Remark 2.1 provides us with \( R + \frac{\lambda p}{\mu_{cs}} \) as an upper bound for the optimal staffing level.
Thus far, we have observed that if one limits the set of controls to the set of \(TP\) controls, then the search for control and an appropriate staffing level are much simplified. It turns out, that in settings where the cross-selling is valuable enough, the staffing problem can be simplified even further, by replacing it with a solution of a single deterministic problem. Specifically, consider the following deterministic relaxation of (4) (relaxing both the waiting time constraint and the integrality restriction on \(N\)):

Maximize \(r\mu_{cs}x - (C(N) - C(R))\),

s.t. \(x \leq (N - R) \wedge \frac{\lambda p}{\mu_{cs}}\),

\(x, N \in \mathbb{R}_+, N \geq R\),

where \(x\) may be interpreted as the deterministic analog of the number of agents busy cross-selling, so that \(\mu_{cs}x\) is the cross-selling completion rate. Clearly, any optimal solution \((x^*, N^*)\) for (6) will satisfy \(N^* = R + x^*\), so that one could re-write (6) as

Maximize \(r\mu_{cs}(N - R) - (C(N) - C(R))\),

s.t. \(R \leq N \leq R + \frac{\lambda p}{\mu_{cs}}\),

\(N \in \mathbb{R}_+\).

We denote the optimal value of \(N\) in (7) by \(N_2\). We claim that (following Theorem 4.1) if

\[N_2 - R \gg N_1 - R,\]

then it is asymptotically optimal to staff with \([N_2]\) agents and use \(TP[K]\), with any \(K \in [0, \lambda \bar{W}/2]\). Here, \(N_1\) is as defined in (5) and, informally, we say that \(x \gg y\) if \(x\) is much greater than \(y\). The insight here is that if the value of cross-selling is such that it is worthwhile to staff with many more agents than required for feasibility, then feasibility ceases to be an issue and one can determine the staffing level based on the tradeoff between staffing cost and cross-selling value.

We will show that whenever (8) holds and the suggested staffing and control are used - cross-selling is exercised for a significant fraction of the customers and the expected revenue from cross-selling is substantial. Thus, whenever (8) holds, we say that the system operates in the Cross Selling Driven (CSD) Regime. Otherwise, we say that the system operates in the Service Driven (SD) Regime, reflecting the fact that when (8) does not hold, it is optimal for the system to cross-sell only to a small fraction of the customers, and the
focus is on service quality.

In this section, we have outlined a solution for the staffing and control problem (3). In settings where the two-dimensional phase information (the number of servers cross-selling and the number of servers giving service) is available to the system manager, an alternative to TP is to use the optimal control obtained by a solution to the appropriate Markov Decision Process (MDP) (see §5 for details). This will result in a slightly better performance (our paper’s results show that for large systems the two controls have similar performance), but at the cost of a more complicated control. In the following sections we give sufficient conditions for the approximate optimality of TP together with the suggested staffing procedure. For the cases that are not covered by the sufficient conditions, we devise a simple linear programming approach to obtain approximately optimal staffing and control schemes. Having the linear programming tool, we show numerically that TP performs extremely well, even in the cases in which it is not provably near-optimal.

4 Asymptotic Framework

In this section we introduce the asymptotic framework and establish our asymptotic optimality results. In our asymptotic analysis we consider a sequence of systems indexed by the arrival rate $\lambda$, which is assumed to grow without bound ($\lambda \to \infty$). The superscript $\lambda$ is used to denote quantities associated with the $\lambda$th system. Consider the problem (4) with respect to the $\lambda$th system:

\[
\begin{align*}
\text{maximize} & \quad r_{\mu cs}(E[Z^{\lambda,\pi}] - R) - (C^{\lambda}(N^{\lambda}) - C^{\lambda}(R)) \\
\text{subject to} & \quad E[W^{\lambda,\pi}] \leq \bar{W}^{\lambda} \\
& \quad N^{\lambda} \in \mathbb{Z}^+, \pi^{\lambda} \in \Pi(\lambda, N^{\lambda}),
\end{align*}
\]

(9)

where we omit the superscript $\lambda$ from parameters that are not scaled with $\lambda$, such as the service rates $\mu_s$ and $\mu_{cs}$, the expected revenue per customer, $r$, and the probability $p$. The superscript $\lambda$ is also omitted from $R$, since $R$ has a trivial dependency on $\lambda$ given by its definition $R = \frac{\lambda}{\mu_s}$. Let $V^{\lambda}(N^{\lambda}, \pi^{\lambda})$ be the objective function value in (9) when the staffing level and control scheme are $N^{\lambda}$ and $\pi^{\lambda}$. Clearly, one needs to define precisely how the different system parameters scale with $\lambda$.

We start by defining the cost function. We consider convex increasing functions $C^{\lambda}(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ with $C^{\lambda}(0) = 0$. We define the scaling of the cost function through the solution to the deterministic relaxation
In particular, the scaled version of (7) is given by

\[
\begin{align*}
\text{Maximize} \quad & r_{\mu cs} \cdot (N^\lambda - R) - (C^\lambda(N^\lambda) - C^\lambda(R)) \\
R \leq N^\lambda \leq R + \frac{\lambda p}{\mu cs} \\
N^\lambda \in \mathbb{R}_+.
\end{align*}
\]

(10)

For a fixed \( \lambda \), let \( N^\lambda_2 \) be the smallest optimal solution to (10). Then, we make the following assumption:

**Assumption 4.1**

1. There exist \( \beta \geq 0 \) and \( \gamma \in \mathbb{R} \), such that

\[
\lim_{\lambda \to \infty} \frac{N^\lambda_2 - R}{R} = \beta,
\]

and

\[
\lim_{\lambda \to \infty} \frac{N^\lambda_2 - R(1 + \beta)}{\sqrt{R}} = \gamma,
\]

In particular, \( N^\lambda_2 = R + \beta R + \gamma \sqrt{R} + o(\sqrt{R}) \).

2. There exists a 'minimum wage' parameter \( c > 0 \) such that for all \( \lambda \) and all \( N \geq R \), \( C^\lambda(N) - C^\lambda(R) \geq c(N - R) \).

Note that by the definition of \( N^\lambda_2 \) we have that \( N^\lambda_2 \leq R + \frac{\lambda p}{\mu cs} \), and in particular that \( \beta \leq \frac{\mu cs}{\mu cs} \). Assumption 4.1 is quite general in the sense that there is a large family of naturally occurring cost functions that satisfy it. For example, any sequence of identical linear functions, \( C^\lambda(x) = cx \), for some \( c > 0 \), trivially satisfies this assumption. The same holds for a sequence of identical convex functions. However, in order to allow a larger scope for our results, Assumption 4.1 also allows for functions that do scale with \( \lambda \) such as piecewise linear convex functions, in which the distances between consecutive break points in the \( \lambda^{th} \) cost function are either of order \( \lambda \) or of order \( \sqrt{\lambda} \). To be concrete, consider the piecewise linear convex function with derivative:

\[
C^\lambda(x) = \begin{cases} 
  c_0, & R < x < R + \tilde{c}\lambda \\
  c_n, & R + \tilde{c}\lambda + (n-1)\sqrt{\lambda} < x < R + \tilde{c}\lambda + n\sqrt{\lambda}, 1 \leq n \leq M, \\
  c_{M+1}, & R + \tilde{c}\lambda + M\sqrt{\lambda} < x < R + \frac{\lambda p}{\mu cs}.
\end{cases}
\]

(13)
for some constants $M, \tilde{c}$ satisfying $R + \tilde{c}\lambda + M\sqrt{\lambda} < R + \frac{\lambda p}{\mu cs}$ and an increasing sequence of positive numbers $c_n, n = 0, \ldots, M + 1.$ Then, $C^\lambda(\cdot)$ can be easily seen to satisfy Assumption 4.1. Indeed $N^x_2$ will be necessarily at a break point of the $\lambda^{th}$ cost function and will thus have the required form. While costs functions that scale with $\lambda$ are not necessary to use they are natural as they allow us to reflect economies of scale. Specifically, using a fixed cost function for all $\lambda$ is equivalent to assuming that the marginal costs of agents are identical whether the call center is of order of 1000's of agents or only 10's of agents.

With respect to the scaling of the waiting time constraint we assume:

**Assumption 4.2** There exists a constant $\hat{W} > 0$, such that for all $\lambda$, $\bar{W}^\lambda = \hat{W}/\sqrt{R}$.

Note that the choice of $1/\sqrt{R}$ as the order of $\bar{W}^\lambda$ is not arbitrary. It is consistent with many other models in which such square root approximations tend to perform extremely well (see for example [18] and [10]). It is also important to note that in order to apply our proposed solution, as presented in §3, one does not have to know the constants $\hat{W}, \beta, \gamma$ and $c$.

Next we formally state our notion of asymptotic optimality.

**Definition 4.1 Asymptotic Feasibility:** We say that a sequence of staffing levels and controls $\{N^\lambda, \pi^\lambda\}$ is asymptotically feasible, if when using $\{N^\lambda, \pi^\lambda\}$, we have

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

(14)

**Definition 4.2 Asymptotic Optimality:** We say that an asymptotically feasible sequence of staffing levels and controls $\{N^\lambda, \pi^\lambda\}$ is asymptotically optimal if for any other asymptotically feasible sequence $\{\tilde{N}^\lambda, \tilde{\pi}^\lambda\}$ we have

$$\liminf_{\lambda \to \infty} \frac{V^\lambda(N^\lambda, \pi^\lambda)}{V^\lambda(\tilde{N}^\lambda, \tilde{\pi}^\lambda)} \geq 1.$$  

(15)

**Remark 4.1** Note that since in the problem formulation (3) we have normalized the cost around $C(R)$, our sense of asymptotic optimality is stronger than first order asymptotic optimality. To wit, consider a sequence of cost functions $C^\lambda(\cdot)$ with $C^\lambda(N) - C^\lambda(R) \leq \tilde{c}(N - R)$ for any $R \leq N \leq R + \frac{\lambda p}{\mu cs}$ and for some positive constant $\tilde{c}$. Assume that an asymptotically optimal staffing sequence $\{N^\lambda, \pi^\lambda\}$ satisfies that $N^\lambda - R = O(\sqrt{\lambda})$. Then, under the first order asymptotic optimality conditions, $\{\tilde{N}^\lambda, \tilde{\pi}^\lambda\}$ with
\( \tilde{N}^\lambda = N^\lambda + k\sqrt{R} \) is also asymptotically optimal for any \( k > 0 \). Indeed, by Lemma 2.1

\[
\liminf_{\lambda \to \infty} \frac{\mu_{cs}(E[Z^\lambda,\pi^\lambda] - R)}{\mu_{cs}(E[Z^\lambda,\pi^\lambda] - R) + C^\lambda(\tilde{N}^\lambda)} = \liminf_{\lambda \to \infty} \frac{C(\tilde{N}^\lambda)}{C(N^\lambda)} = 1. \tag{16}
\]

Our notion of asymptotic optimality will differentiate between these two solutions and is, hence, much finer.

Before stating our main asymptotic optimality results we need a few more definitions. Re-define (5) as

\[
N_1^\lambda = \arg\min \left\{ N \in \mathbb{Z}^+ : E[W^{FCFS}_{\lambda,\mu_s}(N)] \leq \bar{W}^\lambda \right\}, \tag{17}
\]

and define

\[
\tilde{N}_1^\lambda = \inf\{ N \in \mathbb{Z}^+, N \geq R : E[W^{FCFS}_{\lambda,\mu_s}(N) | W^{FCFS}_{\lambda,\mu_s}(N) > 0] \leq \bar{W}^\lambda \}
= \inf\{ N \in \mathbb{Z}^+, N \geq R : \frac{1}{N\mu_s - \lambda} \leq \bar{W}^\lambda \}, \tag{18}
\]

where the last equality follows from well known results for \( M/M/N \) queues (see for example §5-9 of Wolff [29]). Also, we will say that two sequences \( \{x^\lambda\} \) and \( \{y^\lambda\} \) satisfy \( x^\lambda \gg y^\lambda \) if \( x^\lambda \to \infty \) as \( \lambda \to \infty \). In the following theorem, then, we state the sufficient conditions for asymptotic optimality of TP. For each condition we also specify how the staffing and the threshold level will be determined if the condition is satisfied.

**Theorem 4.1** Consider a sequence of systems, with \( \lambda \to \infty \), that satisfies Assumptions 4.1 and 4.2. Then, with \( N_1^\lambda, \tilde{N}_1^\lambda \) and \( N_2^\lambda \) defined in (17), (18) and (10) respectively, the following conditions are sufficient for asymptotic optimality of TP and the proposed staffing levels:

1. \( N_2^\lambda - R \gg N_1^\lambda - R \), then it is asymptotically optimal to set \( K^\lambda = \lceil \delta \sqrt{R} \rceil \) with any \( \delta \) in \( [0, \bar{W}/2] \) and \( N^\lambda = N_2^\lambda \).

2. Condition 1 fails but \( \liminf_{\lambda \to \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} \geq 1 \), and \( \mu_{cs} \geq \mu_s \), then it is asymptotically optimal to set \( K^\lambda = 0 \) and \( N^\lambda = N_2^\lambda \).

3. \( \mu_s = \mu_{cs} \) and \( \limsup_{\lambda \to \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} < 1 \), then it is asymptotically optimal to set \( N^\lambda = N^{*\lambda} \) and
\[ K^\lambda = K^{*\lambda}, \text{ where } K^{*\lambda} \text{ and } N^{*\lambda} \text{ are determined by choosing the smallest value of } N^{*\lambda} \text{ that satisfies} \]

\[ N^{*\lambda} = \arg \max_{N \geq N_1^\lambda} \tau \mu_{cs} \left( E \left[ Z_{\lambda,\mu_s}^{FCFS} (N) \mid Z_{\lambda,\mu_s}^{FCFS} (N) \geq (N + K^\lambda(N)) \wedge N \right] - R \right) - (C^\lambda(N) - C^\lambda(R)), \] (19)

and

\[ K^\lambda(N) = \max_{K \geq -N} \{ Q_{\lambda,\mu_s}^{FCFS} (N) \mid Z_{\lambda,\mu_s}^{FCFS} (N) \geq (N + K) \wedge N \leq \lambda \bar{W}^\lambda \}. \] (20)

Condition 1 corresponds to the Cross-Selling Driven regime. Intuitively, under this condition the value of cross-selling drives the system to use a significant amount of extra staffing to allow for substantial cross-selling. Under TP, whenever the queue length exceeds the threshold, all cross-selling activities are stopped and all capacity is dedicated to drain the queue. The significant extra capacity allows the system to drain the extra queue rather quickly thus preserving feasibility without causing the agents to idle. On the other hand, whenever the queue length is below the threshold every cross-selling opportunity is exercised. This causes all servers to be busy most of the time so that the cross-selling rate is close to its upper bound \( \mu_{cs}(N - R) \). Conditions 2 and 3 correspond to the Service Driven regime. The intuition behind condition 2 is also rather straightforward. Using the assumption that \( \mu_{cs} \geq \mu_s \), we have that whenever all agents are busy the depletion rate of the queue (with threshold equal to 0) is greater than or equal to \( N \mu_s \) (note that this is not true if \( \mu_{cs} < \mu_s \)). Consequently, feasibility is guaranteed by the definition of \( N_1^\lambda \) and the condition that \( N_2^\lambda \) is greater than \( N_1^\lambda \). The intuition behind Conditions 3 is based on the connection between the cross-selling model with equal rates and the model analyzed in Gans and Zhou [15], which is discussed in detail in \( \S 4.1 \).

Our proofs show that under Condition 1 the asymptotically optimal staffing is of the form \( N^\lambda = R + \beta R + o(R) \), for some \( \beta > 0 \). They also establish that \( \mu_{cs} E[Z_2^\lambda] = \mu_{cs}(N^\lambda - R) + o(N^\lambda - R) \). In particular, we have that under Condition 1 the cross-selling rate is approximately \( \mu_{cs} \beta R \), implying that the fraction of customers who end up listening to cross-selling offerings is non-negligible. Indeed, for all \( \lambda \), \( \frac{\mu_{cs} \beta R}{\lambda} = \frac{\mu_{cs} \beta}{\mu_s} > 0 \). This supports our characterization of the Cross-Selling Driven Regime as a regime where a non-negligible fraction of the customers listen to cross-sales offers. A similar argument justifies the name Service Driven Regime. Indeed, we prove that under either Conditions 2 or 3, \( N^\lambda - R \) is of order \( \sqrt{R} \), implying that the fraction of customers for which cross-selling is exercised is of order \( 1/\sqrt{\lambda} \) and, in particular, this fraction goes to zero as \( \lambda \) increases.

We now state a result that plays a critical role in the sufficiency of any of the conditions in Theorem 4.1
Theorem 4.2 Consider a sequence of systems, where the $\lambda^{th}$ system is staffed with $N^\lambda = R + \beta R + \gamma \sqrt{R} + o(\sqrt{R})$ for $0 \leq \beta \leq \frac{\mu_H}{\mu_{cs}}$, $\max(\beta, \gamma) > 0$, and operated with $TP[K^\lambda]$, where

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

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

$$E\left[\left((N^\lambda - Z^\lambda) - [K^\lambda]^+\right)\right] \to 0, \text{ as } \lambda \to \infty,$$

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

Observe that The requirement that $\max(\beta, \gamma) > 0$ is imposed to guarantee stability. Theorem 4.2 implies that whenever $K^\lambda < 0$ the number of idle agents is in some sense bounded by $|K^\lambda|$. Also, it suggests that whenever $K^\lambda \geq 0$, the number of idle agents is approximately 0.

The proofs of the three conditions are long. For brevity we focus on Condition 3 whose discussion illustrates some of the key ideas of our asymptotic analysis. The proofs of Conditions 1 and 2 are given in the Appendix.

4.1 The Service Driven Regime: Condition 3

Our optimality results under Condition 3 are closely related to the results of Gans and Zhou [15]. We start then with a description of the system analyzed in [15] as well as with some results comparing this system with the cross-selling system. While [15] 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 [15] 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 [15]: 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 4.1.

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 [15].

**Definition 4.3** 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 4.3 can be thought of as the TP control adapted to the G&Z model. Denote by $\overline{TP}^\lambda(N^\lambda, p^*)$ the randomized threshold policy of G&Z with threshold $K^\lambda$ determined through (23) and with a randomization probability $p^*$. The following is a version of the optimality result of [15] for the case $\mu_s = \mu_{cs}$. We only cite the parts of the Theorem that are relevant for our results.

**Theorem 4.3 (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 $\overline{W}^\lambda$. 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

$$K^\lambda(N^\lambda) = \max \left\{ k \in [-N^\lambda, 0] \left| \frac{\xi_k(N^\lambda)}{N^\lambda \mu_s - \lambda} \leq \overline{W}^\lambda \right. \right\}. \quad (23)$$

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 4.2** Note that under $\overline{TP}(N^\lambda, 0)$ (i.e. when setting $p^* = 0$), the steady state number of busy agents in the G&Z system, denoted by $E[\overline{Z}^\lambda]$, satisfies $E[\overline{Z}^\lambda] = E[Z_{\lambda,\mu_s}^{FCFS}(N^\lambda) \mid Z_{\lambda,\mu_s}^{FCFS}(N^\lambda) \geq N^\lambda + K^\lambda].$

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 $CS^n(\pi)(t)$ be the cumulative cross-selling completions up to time $t$ when the control $\pi$ is used. Also, let $TH^n(\pi')(t)$ 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 $\overline{Z}^{\lambda,\pi'}$ 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 $\mu_{cs}(E[\overline{Z}^{\lambda,\pi'}] - R)$. The following lemma does not assume $\mu_s = \mu_{cs}$.
Lemma 4.1 Fix $\lambda$, $\mu_s$, $\mu_{cs}$, $N$ and $\bar{W}^\lambda$. Let $\pi^*_giz$ 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
\[ TH_{\pi^*_giz}(t) \geq_{st} CS^\pi(t), \quad \forall t \geq 0. \tag{24} \]
In particular,
\[ \mu_{cs}(E[Z^{\lambda,\pi} - R]) \leq \mu_{cs}(E[\bar{Z}^{\lambda,\pi^*_giz} - R]). \tag{25} \]

Our proof uses sample path arguments and is given in the Appendix.

For future reference let $\bar{V}(N^\lambda) = r\mu_{cs}(E[Z^{\lambda,\pi^*_giz} - R]) - (C^\lambda(N^\lambda) - C^\lambda(R))$, 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 $N^\lambda \geq N^\lambda_1$ for all $\lambda$ and
\[ \frac{N^\lambda - R}{\sqrt{R}} \rightarrow \hat{\gamma}, \quad \text{as } \lambda \rightarrow \infty, \tag{26} \]
for some $\hat{\gamma} > 0$. The existence of such a sequence is guaranteed since by §9 of [10] we have that
\[ \frac{N^\lambda_1 - R}{\sqrt{R}} \rightarrow \gamma, \tag{27} \]
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 $TP^\lambda(N^\lambda, p^*)$. Also, let $Y^{\lambda}$ be the steady state overall number of customers in a cross-selling system with $N^\lambda$ agents and using $TP[K^\lambda]$ with $K^\lambda$ determined through (23). Accordingly, we let $\bar{Z}^\lambda$ and $Z^\lambda$ be the number of busy agents in the above two systems. Define the scaled variables
\[ \bar{X}^{\lambda,p^*} = \frac{\bar{Y}^{\lambda,p^*} - N^\lambda}{N^\lambda - R}, \quad \text{and } X^\lambda = \frac{Y^\lambda - N^\lambda}{N^\lambda - R}. \]

Lemma 4.2 Assume that $\mu_s = \mu_{cs}$, $N^\lambda$ satisfies (26) and $K^\lambda$ is defined through (23). Then, there exists a function $\delta(\cdot, \cdot)$ such that
\[ \frac{K^\lambda}{\sqrt{R}} \rightarrow \delta(\hat{\gamma}, \bar{W}), \quad \text{as } \lambda \rightarrow \infty, \tag{28} \]
where $\delta(\cdot, \cdot)$ is finite for all positive and finite arguments.

Lemma 4.2 is a step in the proof of the following convergence result:
Proposition 4.1 With \( \mu_s = \mu_{cs} \) and \( N^\lambda \) satisfying (26), there exists a random variable \( \bar{X} \) such that for any \( p^* \in [0, 1] \)

\[
\bar{X}^{\lambda, p^*} \Rightarrow \bar{X}, \quad \text{as } \lambda \to \infty,
\]

and

\[
X^\lambda \Rightarrow \bar{X}, \quad \text{as } \lambda \to \infty.
\]

In both cases the convergence also holds in expectation.

The distribution of \( \bar{X} \) can be explicitly found and it is given in the appendix. Also, note that since the result is independent of \( p^* \) we have by Remark 4.2 that whenever \( N^\lambda \) satisfies (26)

\[
\lim_{\lambda \to \infty} \frac{E[Z_{FCFS}^{\lambda, \mu_s}(N^{*\lambda}) - R | Z_{FCFS}^{\lambda, \mu_s}(N^{*\lambda}) \geq N^{*\lambda} + K^\lambda] - (E[\bar{X}^{\lambda}] - R)}{N^\lambda - R} \to 0, \quad \text{as } \lambda \to \infty.
\]

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

Proposition 4.1 leads immediately to the following result, which establishes asymptotic optimality for a fixed sequence of staffing levels, \( N^\lambda \).

Corollary 4.1 Assume \( \mu_s = \mu_{cs} \) and consider a sequence of cross-selling systems such that \( N^\lambda \) satisfies equation (26), and for each \( \lambda \), \( TP[K^\lambda] \) is used with \( K^\lambda \) determined through equation (23). Then,

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

and

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

The following lemma will help us to translate the result of Corollary 4.1 to the more general asymptotic optimality result that we need.

Lemma 4.3 Assume \( \mu_s = \mu_{cs} \) in addition to Assumptions 4.1 and 4.2. Let \( N^{*\lambda} \) and \( K^\lambda \) determined through
(19) and (20) and assume that \( \limsup_{\lambda \to \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} < 1 \). Then,

\[
\liminf_{\lambda \to \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} > 0, \tag{34}
\]

and

\[
\limsup_{\lambda \to \infty} \frac{N^{*\lambda} - R}{\sqrt{R}} < \infty. \tag{35}
\]

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

- **Staffing:** Staff with \( N^{*\lambda} \) agents where \( N^{*\lambda} \) is given by equations (19) and (20).

- **Control:** Use \( TP[K^\lambda(N^{*\lambda})] \) where \( K^\lambda(N^{*\lambda}) \) is given by equation (20).

## 5 The Linear Programming (LP) Approach

When either of the conditions 1., 2., or 3. in Theorem 4.1 is satisfied, we managed to overcome the two-dimensional nature of the control problem through our asymptotic analysis. In this section we propose a solution approach to the control component of (3) which applies in great generality, beyond the cases covered by our sufficient conditions. Specifically, following the approach in [15], we consider the solution to a related Markov Decision Process (MDP). Without any restriction on the family of controls used, this MDP might require the solution of an infinite state space problem. Using asymptotic analysis, however, we are able to reduce the problem to finite dimensional one. Specifically, we solve the MDP for a finite buffer system and show that, when assuming stationary policies, this MDP is asymptotically equivalent to the original problem as the buffer size grows without bound. This finite buffer MDP is solved through a solution to a linear program (LP). The optimal control associated with the solution to the LP is generally not a TP control. Nevertheless, in \( \S \)6 we show that TP is nearly optimal by numerically comparing its performance to that of the asymptotically optimal control obtained from the solution of the LP. Note that this asymptotic approach is different from what we have done so far since we now fix \( \lambda \) and let the buffer size grow without bound. Accordingly, in this section, we fix \( \lambda \) and omit the superscript \( \lambda \) from all the
We now turn to the formulation of the Markov Decision Process and its reduction to a solution of an LP. We start by showing that it suffices to consider a subset of all admissible policies. Specifically, we show in Lemmas 5.1 and 5.2 that there always exists a work conserving policy that serves customers FCFS whenever there are customers in queue. For a fixed $N$, we re-define $\Pi(N)$ to be the set of non-anticipative non-preemptive feasible policies. That is, $\Pi(N)$ is the set of all admissible policies $\pi$ for which steady state exists and $E[W^\pi] \leq \bar{W}$ when there are $N$ agents in the system. In particular, it is clear that $\Pi(N)$ will be empty unless $N \geq N_1$ where $N_1$ was defined in (5). The following result applies with respect to $\Pi(N)$:

**Lemma 5.1** Fix $\lambda, \mu_s, \mu_{cs}$ and $N$. Then, for any $\pi \in \Pi(N)$ there exists a policy $\pi' \in \Pi(N)$ that serves customers FCFS and performs at least as well as $\pi$. In particular, $\pi'$ admits a cross-selling rate that is at least as large as that admitted by $\pi$.

**Definition 5.1 Work Conservation:** We say that a policy $\pi$ is work conserving if: (i) whenever a customer arrives to find an idle agent, s/he is immediately admitted to service, and (ii) upon a departure of a customer from the system, a waiting customer will be admitted to service if the queue is non-empty.

Note that work conservation implies that $Z(t) = N$ whenever $Q(t) > 0$. It does not imply, however, that the policy gives priority to the customers waiting in queue over cross-selling. In fact, work conservation does allow exercising cross-selling even when customers are waiting.

The following lemma shows that within the class of FCFS policies it is sufficient to consider work conserving policies. In turn, it suffices to consider FCFS work conserving policies.

**Lemma 5.2** Fix $\lambda, \mu_s, \mu_{cs}$ and $N$. Then, for any feasible policy $\pi \in \Pi(N)$ there exists a work conserving feasible policy $\pi' \in \Pi(N)$ that performs at least as well as $\pi$. In particular $\pi'$ admits a cross-selling rate that is at least as large as that admitted by $\pi$.

Within the set of work conserving FCFS policies we limit our attention to stationary policies. Stationary policies are not only very practical, due to their very definition as dependent only on the state of the system, but they are, in great generality, a sufficient family of policies for optimality. This, however, is not a trivial result to prove in a setting with infinite state space (although sufficient conditions do exist in the literature: 24
see, for example, Altman [2]). Thus, we impose the restriction to stationary policies as an assumption and re-define the set of admissible policies \( \Pi(N) \) accordingly.

We are now ready to construct the MDP and the associated Linear Program for a cross-selling system with a finite buffer. In this construction we limit ourselves to stationary work conserving FCFS policies. Note that the family of stationary work conserving FCFS policies need not be optimal for finite buffer systems. We will prove, however, that for a buffer size that is large enough and for any given stationary work conserving and FCFS policy for the infinite buffer system, there exists a stationary work conserving and FCFS policy for the finite buffer system that performs almost as well. Since the family of stationary work-conserving FCFS policies is optimal for the infinite buffer system, the constructed LP leads to an asymptotically optimal solution. For stationary work-conserving FCFS policies, the state descriptor \( \{Z_2(t), Y(t)\} \) suffices for a complete Markovian characterization of the system. Indeed, work conservation implies that the identities \( Z_1(t) = (Y(t) \wedge N) - Z_2(t) \) and \( Q(t) = [Y(t) - N]^+ \) hold, so that one can characterize the behavior of the whole system through the two-dimensional description \( \{Z_2(t), Y(t)\} \). Suppose that the number of customers in system is bounded above by a finite number of trunk lines, \( L \geq N \). Customers that find a full buffer upon arrival are blocked and do not enter the system.

Since all transition rates in the system are bounded by \( \lambda + N\mu_s + N\mu_{cs} \) we can replace the analysis of the underlying Continuous Time Markov Chain (CTMC) with the analysis of the associated Discrete Time Markov Chain (DTMC) which is obtained from the CTMC by uniformization. The construction by uniformization ensures that the steady state fraction of time that the CTMC spends in any given state corresponds exactly to the steady state fraction of steps that the corresponding DTMC spends in that state. Naturally, we let the uniformization rate equal the upper bound \( \lambda + N\mu_s + N\mu_{cs} \).

By results for constrained long run-average MDP’s (see for example section 4.2 of [2]) one can solve the finite state MDP through an appropriate LP. Note that due to work conservation, in each state \( \{Z_2, Y\} \) the action set consists of only two options - cross sell upon service completion (1) or do not cross-sell (0). Let \( \xi(i, j, k) \) be the steady state probability of being in state \( \{Z_2, Y\} = (i, j) \) and taking the action \( k \in \{0, 1\} \) (note that a stationary distribution will always exist for this model due to the finite buffer). The corresponding LP for a system with \( L \) trunk lines and \( N \) agents is then formulated as follows:

\[25\]
Max \[ \sum_{j=0}^{L} \sum_{i=0}^{J \wedge N} ri\mu_{cs}(\xi(i, j, 0) + \xi(i, j, 1)) \]  

\[ \text{s.t} \quad (\lambda + N\mu_s + N\mu_{cs}) \cdot (\xi(i, j, 0) + \xi(i, j, 1)) = \lambda(\xi(i, j - 1, 0) + \xi(i, j - 1, 1))1_{\{j-1 \geq 0\}} \]
\[ + \mu_s((j + 1) \wedge N - i)(\xi(i, j + 1, 0) + (1 - p)\xi(i, j + 1, 1))1_{\{j+1 \leq L\}} \]
\[ + p\mu_s(j \wedge N - (i - 1))\xi(i - 1, j, 1)1_{\{i-1 \geq 0\}} \]
\[ + \mu_{cs}(i + 1)(\xi(i + 1, j + 1, 0) + \xi(i + 1, j + 1, 1))1_{\{i+1 \leq N\}}1_{\{j+1 \leq L\}} \]
\[ + (\xi(i, j, 0) + \xi(i, j, 1))(\lambda(\xi(i - 1, j, 0) + (N - (j \wedge N - i))\mu_s + \lambda 1_{\{j=L\}}), \]
\[ 0 \leq j \leq L, \quad 0 \leq i \leq j \wedge N, \]

\[ \sum_{j=0}^{L} \sum_{i=0}^{J \wedge N} (\xi(i, j, 0) + \xi(i, j, 1)) = 1, \] \hspace{1cm} (37)

and

\[ \sum_{j=N}^{L} (j - N) \sum_{i=0}^{N} (\xi(i, j, 0) + \xi(i, j, 1)) \leq \bar{W}^\lambda \] \hspace{1cm} (39)

The system of equations in (37) represents the balance equations of the underlying DTMC, keeping in mind that the action choice only affects the chain transitions immediately following a phase 1 service completion. In particular, for any fixed \((i, j)\) the right hand side in (37) lists the possible transitions from other states with the associated probabilities. Specifically, the first line on the right hand side of (37) corresponds to the transitions due to arrivals. The second line corresponds to transitions due to phase 1 service completions that are not followed by a cross-selling phase. The third line corresponds to transitions due to phase 1 service completions that are followed by cross-selling. The fourth line corresponds to transitions due to phase 2 service completions and the last line corresponds to transitions from the state to itself.

Recall that any feasible staffing level for the original system (with infinite number of lines) must be greater than or equal to \(N_1\) where

\[ N_1 = \arg\min \left\{ N \in \mathbb{Z}_+ : E[W_{\lambda,\mu_s}^{FCFS}(N)] \leq \bar{W}^\lambda \right\}. \] \hspace{1cm} (40)

Let \(V^*_{LP}(N, L)\) be the optimal solution of the LP corresponding to a system with \(N\) agents and \(L\) trunk lines (recall that \(\lambda\) if fixed). For a fixed \(N\), let \(V^*(N)\) be the optimal expected revenue in (3) when \(N\) is
fixed and assuming stationary policies, that is

\[ V^*(N) = \sup_{\pi \in \Pi(N): E[W^*] \leq \bar{W} \in \mathbb{R}^+} r \mu_{cs} (E[Z^\pi] - R). \]

Then, we have the following:

**Proposition 5.1** Assume \( N \geq N_1 \). Then,

\[ \lim_{L \to \infty} V^*_{LP}(N, L) = V^*(N). \] (41)

We use the result of Proposition 5.1 in the next section to illustrate the remarkably good performance of TP. Specifically, we show numerically that TP achieves a cross-selling rate that is almost identical to the one obtained through the LP with a large buffer size. This indicates that, within the set of stationary policies, TP is close to optimal.

### 6 Numerical Experiments

In this section we present the results of our numerical experiments. Through a comparison of the value obtained through the asymptotically optimal LP against the performance obtained using TP, we show that TP performs well beyond the scope covered by our sufficient conditions and that its good performance is preserved also for relatively small call centers. Specifically, we experiment with two different systems, one with an offered load of \( R = 30 \), representing a relatively small call center, and the other with \( R = 100 \), representing a medium-large call center. With respect to TP, for each of the values of \( R \), we vary the staffing levels and for each of those staffing levels we search for the best threshold level that satisfies an average waiting time that is equal to 1/10 of the average service time and calculate the corresponding cross-selling rate. Also, for each staffing level we calculate the cross-selling rate obtained from the asymptotically optimal LP. Recall that the asymptotically optimal LP corresponds to a system with a finite number of trunk lines. For both the asymptotically optimal LP and the calculation of the best threshold under TP we assume the same finite buffer. Specifically, we fix the number of lines to be 100 and 200 for \( R = 30 \) and \( R = 100 \) respectively. As suggested by the previous section, and in particular by Proposition 5.1, when the number
of lines is large enough, performance should close the corresponding system with an infinite number of lines. Indeed, repeating the same experiments with larger number of lines, hardly changes the results. We fix $\mu = 1$ throughout but allow for different ratios of $\mu_{cs}/\mu_s$. Specifically, we consider both $\mu_{cs} = 3$ (fast cross-selling) and $\mu_{cs} = 1/3$ (slow cross-selling). In all the experiments $p$ and $r$ are assumed to be 1, so that the cross-selling revenue is essentially equal to the cross-selling rate.

As will be shown shortly, TP does perform extremely well, even in cases that are not covered by Theorem 4.1 (including slow cross-selling with a number of servers that does not exceed $R$ significantly). The performance of TP can be improved even further by introducing randomization. Intuitively, the non-randomized version of TP is likely to lead to an average waiting time that is strictly less than imposed by the constraint. With the appropriate randomization, however, the constraint can be satisfied as an equality. Using this intuition, it seems reasonable to add randomization to TP (along the lines of the randomized policy in [15]). In particular, if the threshold value we chose is $K$, we use the following randomized version of TP: Whenever there is phase 1 service completion and $(Y(t) - N) \leq K$, the agent will exercise cross-selling on a cross-selling candidate whose service has just been completed. Whenever there is phase 1 completion and the number of customers in system is strictly above $K + 1$, the agent does not exercise cross-selling. Otherwise, if the number of busy agents is $K + 1$, cross-selling is attempted on cross-selling candidates with probability $p^*$. As expected, our experiments suggest that the improvement obtained through this randomization becomes negligible as the system size increases.

Our experiments were performed as follows: Fixing the load $R$, the number of lines $L$, and the staffing level $N \geq N_1$, the following was done:

- Calculate the approximately optimal cross-selling rate associated with the asymptotically optimal LP.
- Search for the optimal threshold $K(N)$.
- Fixing $N$ and $K(N)$, find the maximum possible randomization probability $p^*$ by using increments of size 0.2. (Precision may obviously be improved at the cost of more computation).

We vary the staffing level and plot the results of TP, randomized TP and the asymptotically optimal LP on a single graph.

Figure 2 displays the comparison results for $R = 30$. The graph on the left hand side of the figure (graph 2(a)) displays the result for $\mu_{cs} = 1/3$ while the one on the right (graph 2(b)) displays the result
when $\mu_{cs} = 3$. Note that due to the different cross-selling rates the y-axis is scaled differently in these two graphs.

![Performance Comparison](image1)

**Figure 2**: Cross Selling Rates Comparison for $R = 30$: (a) $\mu_{cs} = 1/3$ (b) $\mu_{cs} = 3$

Figure 3 displays the comparison results for $R = 100$. Again the graph on the left hand side corresponds to $\mu_{cs} = 1/3$ while the graph on the right hand side corresponds to $\mu_{cs} = 3$. Note that in this case, due to the size of the system, the randomization hardly makes any difference. In particular, when $\mu_{cs} = 3$ the lines of TP and Randomized TP coincide almost completely. The case of $R = 100$ also serves well to emphasize another important point. In the introduction to this paper we have mentioned that using a threshold on the queue length is a common heuristic, i.e, cross-selling is exercised upon service completion only when the number of callers in the queue is below a certain threshold. Looking into the optimal thresholds generated for this example through our numerical analysis reveals, however, the real need to use thresholds on the number of idle agents (as allowed by TP by setting $K < 0$), rather than on the number of customers in queue. Focusing on the case $R = 100$ and $\mu_{cs} = 1/3$, the optimal thresholds for the different staffing levels are given in Table 6 which shows that, for all staffing levels, 106 – 122, the optimal threshold is negative. Intuitively, it is optimal to have a threshold on the number of idle agents whenever the waiting time constraint will otherwise be violated. In particular, if staffing level is sufficiently low, one has to reserve idle agents for future arriving calls, in order to satisfy feasibility.

<table>
<thead>
<tr>
<th>Staffing</th>
<th>106</th>
<th>110</th>
<th>114</th>
<th>118</th>
<th>122</th>
<th>126</th>
<th>130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>-20</td>
<td>-10</td>
<td>-6</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 1**: Threshold values for $R = 100$

To summarize our numerical experiments, in all the cases we have analyzed above, the performance of
TP (and in particular the randomized version of TP) is extremely good, and can be improved even further by refining the search of \( p^* \). The results of these experiments support the claim that TP exhibits a remarkable performance in great generality for large as well as moderate size call centers and beyond the scope covered by our sufficient conditions.

To conclude this section, we illustrate the significance of taking cross-selling into account in the staffing decision. We focus on the example of Figure 2(b) with \( R = 30 \) and \( \mu_{cs} = 3 \), and consider the staffing decision we consider the following piecewise linear staffing cost function: Each agent up to the level of 35 agents costs $2 per unit of time. Beyond that, each additional agent costs $4 per unit of time. We now compare two approaches towards the staffing problem. First, we take the naive approach of making the staffing decision ignoring cross-selling, i.e., we choose the number of agents to be the minimal number that would satisfy the waiting time constraint in an \( M/M/N \) model with the corresponding parameters. The corresponding number of agents is 34. With 34 agents our experiments above show that the maximal revenue from cross-selling rate is approximately $5.2, corresponding to an expected profit (calculated as in (4)) of approximately $ - 2.9 per unit of time. Now, assume that we allow ourselves to adjust the staffing levels. Then, it is easily verified that the optimal staffing level would actually be 35 agents leading to an expected profit of approximately $0.6 per unit of time. Hence, even in this setting where potential revenues from cross-selling are relatively low (recall \( r = 1 \)) considering cross-selling in the staffing decision makes a big difference with respect to profits and losses. The significance of cross-selling considerations in the staffing decision becomes even more pronounced as cross-selling value increases.
7 Conclusions and Future Research

The practice of cross-selling in call centers is becoming prevalent and many organizations recognize its revenue potential. Yet, operational aspects of cross-selling have so far attracted little attention in the literature. In particular, very few papers address the control problem of determining when to exercise cross-selling opportunities, and (to the best of our knowledge) our paper is the first one to address the staffing problem of determining how many customer service representatives are needed. Those papers that have dealt with the control problem all illustrate that solving this problem is difficult, which could indeed be the reason why no simple solutions have been proposed so far. In this paper we have tackled the joint problem of determining staffing and control by using an asymptotic approach, in which we look for a staffing level and a control which might not be optimal for each particular problem instance, but they are asymptotically optimal in the sense that they perform extremely well, in the limit, as the arrival rate grows large.

Our approach has allowed us to not only identify a simple control rule (the Threshold Priority (TP) rule), but to also propose a corresponding staffing rule. Together, the staffing and control rule are provably asymptotically optimal in the limit as the system size grows large, under very general assumptions. We have also shown numerically, that they perform well even for systems with relatively small arrival rate.

The managerial implications of our results (beyond the tactical level of determining how many servers are needed and when to exercise cross-selling opportunities) are as follows:

1. Cross-selling must be taken into account when determining the staffing level: A naive approach could be to determine the staffing level ignoring the existence of cross-selling, taking into account only staffing costs, service level constraints and service time, and then handle cross-selling by determining what control to use. We have shown that this approach can lead to far from optimal solutions, and that in fact the value of cross-selling and the associated additional handling time must be taken into account in the staffing decision as well as the control.

2. The incorporation of cross-selling into a service-focused call center might improve overall customer experience (and not only the revenue of the firm): For any fixed staffing level, increasing handling times by introducing cross-selling will certainly increase the delays experienced by all customers. However, if staffing levels are appropriately adjusted those delays will be significantly reduced, if the value of cross-selling is high enough. Taking into account also that cross-selling offerings can be
tailored to customer needs and finally, that customers can always refuse to listen to a cross-selling offering, customers indeed gain from the incorporation of cross-selling.

Many questions remain unanswered with respect to the operational aspects of cross-selling in call centers. Particularly, it is unclear how the customers‘ experience prior to the cross-selling offering affects their tendency to a) listen to the offer and b) purchase the product. Clearly, though, if customers‘ experience has a significant effect on these two tendencies, then one must take this dependence into account when determining the staffing and control. Empirical and experimental research can be helpful in determining how callers actually respond to cross-selling offerings depending on factors such as their delay, service time and overall quality of service. Another interesting question is how to utilize the customer identity when determining whether to exercise a cross-selling opportunity and what products to attempt to sell. Our follow-up paper [5] (co-authored with Costis Maglaras) addresses some of these questions by studying the effect of customer heterogeneity on operational and economic controls emphasizing the impact of the firm’s ability to customize its decisions based on individual customer characteristics.

We conclude by commenting on the connection between our staffing proposal and the well known square-root safety staffing rule, commonly used in the literature.

### 7.1 The Square Root Safety Staffing Rule in a Cross-Selling Environment

In pure service call centers, in which no cross-selling activities are performed, a common rule of thumb for staffing is the Square Root Safety staffing (SRSS) rule. Specifically, with \( R \) defined as before, SRSS suggests using \( N = R + \gamma \sqrt{R} \), for some \( \gamma > 0 \). SRSS was theoretically supported by Halfin and Whitt [20], Borst et. al. [10], Armony [3], and Gurvich et. al. [18], among others.

Our analysis in the current paper suggests that a direct implementation of SRSS in a cross-selling environment may be far from optimal. In particular, we have shown that under certain conditions the safety staffing \( (N - R) \) is orders of magnitude greater than \( \sqrt{R} \). Specifically, we have shown that if it is deterministically optimal (referring to (6)) to cross-sell a fraction \( f^* > 0 \) of the customers, a staffing level of the form

\[
N = R + f^* \frac{\lambda}{\mu_{cs}} = R + f^* \frac{\mu_s}{\mu_{cs}} R
\]

(42)

is asymptotically optimal (and the addition of an extra \( \gamma \sqrt{R} \) will not affect the asymptotic optimality).
A direct implementation of SRSS is, hence, inappropriate in a cross-selling environment. The validity of SRSS holds, however, if we define it differently. Indeed, defining $R' = R + \frac{f^* \lambda}{\mu_{cs}}$, we have by (42) that it is asymptotically optimal to use $N = R' + \gamma \sqrt{R'}$, for any $\gamma \geq 0$. There is however, a crucial difference between the SRSS rule for a pure service call center and the one we have just suggested for the cross-selling call center. While the square root term is critical to ensure short delays in pure service systems, it would be of little importance in cross-selling systems where the capacity dedicated to cross-selling is significant, i.e, in the cross-selling driven regime. In particular, in the cross-selling system one may ignore the square root component, since the service level is easily guaranteed by fine tuning the amount of cross-selling (and the waiting time) by adjusting the threshold level associated with the TP control.

How do these simple observations relate to call center practice? In practice, call center managers might regard the observed handling times as consisting of a single phase and ignore the fact that the observed handling times are not only often composed of two phases but are actually highly dependent on the cross-selling control used. In particular, higher handling times will be observed when the control leads to increased cross-selling. Basing the staffing decision on a naive estimate of the handling times might then lead to inappropriate staffing levels. Interestingly, if a call center is already cross-selling to its optimal fraction $f^*$, its naive estimate of the mean handling time will be $\frac{1}{\mu_s} + \frac{f^* \mu_{cs}}{\mu_{cs}}$. In particular, the estimate of the offered load will be $R' = \frac{\lambda}{\mu_s} + \frac{f^* \mu_{cs}}{\mu_{cs}}$, so that using SRSS will most likely perform rather well under a reasonable control rule. If, on the other hand, the call-center starts by operating away from its optimal fraction of cross-sold customers, this fraction will remain sub-optimal regardless of the control used. Indeed, assume that the call center uses $N = R' + \gamma \sqrt{R'}$ agents with $R'$ now equal to $R + \frac{f \lambda}{\mu_{cs}}$ for some $f \neq f^*$. Then, since an appropriately chosen square-root term is sufficient to guarantee service level satisfaction, the call center will - under any reasonable policy (and in particular under TP) - cross-sell very close to its maximum capability which is given by $\mu_{cs}(N - R) = \lambda f + O(\sqrt{R'})$ (see Lemma 2.1). The new estimate of the average service time (which is obtained by averaging over all customers) will then be $\frac{1}{\mu_s} + \frac{f \mu_{cs}}{\mu_{cs}} + o(1)$. Consequently, the call center will continue performing sub-optimally. Observe that while the $o(1)$ component in the service time might have some effect on staffing, its cumulative effect will only become significant in the very long-run.
References


