Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Mor Armony, Itai Gurvich,

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 12, No. 3, Summer 2010, pp. 470–488
ISSN 1523-4614 | EISSN 1526-5498 | 10 | 1203 | 0470

informs®
DOI 10.1287/msom.1090.0281

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

Mor Armony
Stern School of Business, New York University, New York, New York 10012, marmony@stern.nyu.edu

Itai Gurvich
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208,
i-gurvich@kellogg.northwestern.edu

We study cross-selling operations in call centers. The following questions are 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 center while maintaining a prespecified service level target? We tackle these questions by characterizing control and staffing schemes that are asymptotically optimal in the limit, as the system load grows large. Our main finding is that a threshold priority 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 threshold priority reduces the staffing problem to a solution of a simple deterministic problem in one regime and to a simple search procedure in another. We show that our joint staffing and control scheme is nearly optimal for large systems. Furthermore, it performs extremely well, even for relatively small systems.

Key words: call centers; cross-selling; queueing systems; many-server queues; heavy traffic approximations; steady-state analysis
History: Received: April 29, 2008; accepted: August 24, 2009. Published online in *Articles in Advance* December 29, 2009.

1. Introduction

Call centers are in many cases the primary interaction channel of a firm with its customers. Historically, call centers have been mostly considered a service delivery channel. Such service-driven call centers typically plan their operations based on delay-related performance targets. Examples of such performance measures are average speed of answer, 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 firms, however, cannot be considered as purely service providers. Rather, customer service is a companion to one or several main products. For example, the core business of computer hardware companies such as 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 have call centers that give customer support, whereas their main business is selling financial products. For these firms, the inbound call center can be a natural 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. When the service part of the call is done, 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 the potential to become an ideal sales environment. Modern customer relationship management (CRM) systems have dramatically improved the information available to customer service representatives about the individual customer in real time. Specifically, in call centers, once the caller has been identified, the CRM system can inform the agent about 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 a 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 McKinsey report (Eichfeld et al. 2006), 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, Eichfeld et al. (2006) 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 are apparent, 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 customer waiting times. Does this imply that incorporating 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 a purely service-driven call center, the manager typically attempts to minimize the staffing level while maintaining a predetermined delay-related performance target. Hence, in this pure service context the operational tradeoff is clear: staffing cost versus service level. When sales or promotions are introduced, however, one should consider the potential revenue from these activities as a third component of 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 these cases, increased revenues from cross-selling need not come at the cost of service-level degradation. Rather, we show that the call center can simultaneously achieve high cross-selling rates and very small waiting times. There are cases, however, where the relation between staffing costs and potential revenues is more intricate and a more careful analysis is required.

In addition to staffing, the dynamic control of incoming calls and cross-sales offerings is another important component in the operations management of such centers. Specifically, the call center manager needs to determine when an agent should exercise a cross-selling opportunity. This decision should take into account not only the characteristics of the customer but also the effect on the waiting times of other customers. For example, to satisfy a delay target, it would be natural to stop all promotion activities in the presence of heavy congestion. Indeed, a heuristic used in some call centers to determine when to exercise a cross-selling opportunity is to cross-sell on service completion only when the number of callers in the queue is below a certain threshold. Optimal rules, however, are typically not as simple. As cross-selling of a customer can start only on his service completion, an optimal control is likely to use information about whether the busy agents are providing service or are engaged in cross-selling. In particular, a reduced state description that includes only information about the aggregate number of customers in the system appears to be insufficient. In reality, however, the agents may not signal when they move from the service phase of the interaction to the cross-selling phase; consequently, only the aggregate information is available to the system manager. Hence, a control scheme that relies on this information only is valuable in practice.

The staffing and control issues are codependent because even with seemingly adequate staffing levels, 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 typically have been addressed separately in the literature. To our knowledge, this paper and its follow up paper (Gurvich et al. 2009) are 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 the major operational tradeoffs in the cross-selling environment. This is done by specifying how to adjust the staffing level and how to choose the control 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 above-mentioned heuristic, we identify cases in which cross-selling should not be exercised, even when there are some idle agents in the system, in anticipation of future arrivals.

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

1. From a modeling perspective, this is the first paper (together with our followup paper Gurvich et al. 2009) to address the staffing and cross-selling control questions jointly in one single framework.

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 TP policy, together with a staffing rule and rigorously establish their near optimality.

(a) The qualities of the TP policy include:

(i) It is based only on the total number of customers in the system, rather than on the more elaborate two-dimensional description that distinguishes between agents providing service and those engaged in cross-selling.

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

(b) The staffing rule we propose is simple and easy to implement and reveals much about the regime at which the center should operate: The profit-driven (PD) versus the service-driven (SD) regime.

The rest of the paper is organized as follows: we conclude the introductory part with a literature review. Section 2 provides the problem formulation. In §3 we introduce the asymptotic framework and state our key result: the asymptotic optimality of the threshold policy and the corresponding regime driven staffing rule. A Markov decision process (MDP) approach to identify an optimal control given any fixed staffing level is described and explored in §4. The solution of this MDP is used as a benchmark in testing the performance of our asymptotic proposed scheme via an extensive numerical study in §5. This section also explores implementation issues relevant for call center managers. Section 6 concludes the paper with a discussion of our results and directions for future research. A table of main notations is available as the appendix immediately following §6.

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 discussion, while relegating the formal proofs to the e-companion and a technical appendix.

1.1. Literature Review

A successful and comprehensive treatment of cross-selling implementation in call centers would clearly require an interdisciplinary 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 cross-selling is quite rich, very little has been done from the operations point of view (the reader is referred to Akşin and Harker 1999 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 various coauthors. In Akşin and Harker (1999), 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. Örmeci and Akşin (2006), in contrast, 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, not only on its expected value. However, because of computational complexity, the results in Örmeci and Akşin (2006) are limited to multiserver loss systems (customers either hangup 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. Gunes and Akşin (2004) analyze the problem of providing incentives to agents 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 (2004, 2007) examine the value of customer identity information by comparing cross-selling revenues under several control schemes that differ with respect to the information they use. Exact analysis is performed for the single-server case in Byers and So (2007), and numerical results are given for the multi-server case in Byers and So (2004).

To position our paper in the context of the literature introduced above, note that previous models have considered cross-selling decisions that are made when a customer is assigned 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. Another distinctive feature of our model is our assumption that the system operates with many-servers and an infinite buffer. In contrast, single-server or loss-system assumptions are made in the existing literature for tractability purposes. 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 becomes simpler. Indeed, our solution provides conditions under which the staffing level that maximizes the expected profit from cross-selling simultaneously achieves extremely low waiting times.

In a follow-up paper (Gurvich et al. 2009), the authors use the results of the current paper to study the impact of a heterogeneous pool of customers on the structure of asymptotically optimal staffing and control schemes. Gurvich et al. (2009) also investigate the value of customer segmentation in such an environment.

Our solution approach follows the many-server asymptotic framework, pioneered by Halfin and Whitt (1981). In particular, we follow the asymptotic optimality framework approach first used by Borst et al. (2004) and adapted later to more complex settings (Armony and Mandelbaum 2009; Armony 2005; Armony and Maglaras 2004a, b; Gurvich et al. 2008; Mandelbaum and Zeltyn 2007). The asymptotic regime that we use has been shown to be extremely robust even in relatively small systems (see Borst et al. 2004). Consistent with this finding we give 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 for proofs in our framework. Instead, we introduce a proof methodology that was later formalized in Gurvich and Zeevi (2007) through the notion of constrained Lyapunov Functions.

To conclude this review, we mention that, although it is 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 (2008) and Netessine et al. (2006).

2. Problem Formulation

We consider a call center with calls arriving according to a Poisson process with rate λ. An agent-customer interaction begins with the service phase, whose duration is assumed to be exponentially distributed with rate μ_s. On 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 μ_{cs}. If cross-selling is not exercised, either intentionally or because the customer refuses to listen to a cross-selling offer, the customer leaves the system. It is assumed that all interarrival, service, and cross-selling times are independent and that the call center has an infinite waiting space.

Not all customers are viewed by the center as cross-selling candidates. We assume that the customer population is divided into two segments so that only a fraction $\bar{p}$ of the customers are potential cross-selling candidates. The remaining customers are not considered profitable and are never cross-sold to. Whether or not a specific customer is a profitable candidate is interpreted from a combination of the information available a priori via the CRM system and the information gathered by the agent during the customer interaction. It is important to note that, even if an agent attempts to cross-sell to a caller, the latter will not necessarily agree to listen to the cross-selling offer. We assume that a customer who 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, if faced with one. 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.

We 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. We use the general notation π for a control policy that determines the actions in different decision epochs and, in particular, determines whether to exercise this cross-selling opportunity on a phase 1 completion of a cross-selling candidate. We let $Z_i^\pi(t)$ be the number of servers providing phase i service at time t and $i = 1, 2$ and $Z^\pi(t) = Z_1^\pi(t) + Z_2^\pi(t)$ be the total number of busy agents at time t under the control π. Given the number of agents, N, $I^\pi(t) := N - Z^\pi(t)$ is the number of idle agents at time t under the control π. The number of customers waiting in queue at time t is denoted by $Q^\pi(t)$ and $Y^\pi(t)$ is the overall number of customers in the system at time t $(Y^\pi(t) = Z^\pi(t) + Q^\pi(t))$. Finally, we let $W^\pi(t)$ be the virtual waiting time at time t (the waiting time that a virtual customer would experience if he arrived at time t). In all the above, we omit the time index t when referring to steady-state variables. Also, we omit the superscript π whenever the control is clear from the context. Note that under any stationary policy (as defined in Puterman 1994, p. 22), all transition rates in the system can be determined using the number of agents providing either phase of service and the queue length. In particular, $S^\pi(t) = \{Z_i^\pi(t),\ i = 1, 2;\ Q^\pi(t)\}$ is a Markov process under any stationary policy. Let $A(t)$ be the number of calls arriving by time t and let x_k^π, $k = 1, 2, \ldots$ be equal to 1 if the kth arriving customer ends up going through phase 2 and equal to 0 otherwise. Then, if a steady state exists under π, we let $P^\pi(cs)$ be the long-run proportion of customers that go through cross-selling, i.e.,

$$P^\pi(cs) := \lim_{t\to\infty} \frac{1}{A(t)} \sum_{k=1}^{A(t)} x_k^\pi.$$

2.1. Admissible Controls

The control policy π is picked from the following set of admissible controls $\Pi(\lambda, \mu_s, \mu_{cs}, N)$. Given a staffing level N and parameters λ, μ_s, and μ_{cs}, we say that π is an admissible policy if it is nonpreemptive and nonanticipative and it weakly stabilizes the system.[1] Nonanticipation should be interpreted here in the standard way; for a formal definition, see, e.g., Definition 4.1 in Gurvich and Whitt (2009). In a nutshell, a policy π is nonanticipative if a decision at a time t is only based on the information revealed by the evolution of the system up to that time point. When the parameters μ_s and μ_{cs} are fixed, we will omit them from the notation and use the notation $\Pi(\lambda, N)$ instead. The arrival rate λ and number of agents N will also be omitted whenever their values are clear from the context.

Finally, we note that by the *Poisson arrivals see time averages* (PASTA) property the steady-state virtual waiting time and the steady-state waiting time at arrival epochs coincide. With $C(\cdot)$ being the staffing cost function, the profit maximization formulation is then as follows:

$$\begin{aligned} &\text{maximize} && r\lambda P^\pi(cs) - C(N) \\ &\text{subject to} && E[W^\pi] \le \overline{W}, \\ &&& N \in \mathbb{Z}_+,\ \ \pi \in \Pi(\lambda, \mu_s, \mu_{cs}, N). \end{aligned} \tag{1}$$

Here the average steady-state waiting time $E[W^\pi]$ is constrained to be less than a predetermined bound $\overline{W}$. We assume that $C(\cdot)$ is convex and increasing in the

[1] Weak stability is defined as $\lim_{t\to\infty} E[Q^\pi(t)]/t = 0$. For there to exist a policy that stabilizes the system, we need, at the very least, $\lambda/\mu_s < N$.

staffing level N. Further assumptions are made on the cost function in §3, where we construct our asymptotic framework. Note that customers do not abandon or balk, nor are they being blocked.[2]

We used the maximization formulation (2) although the maximum need not exist. The word "maximize" should be formally interpreted as taking the supremum over all staffing levels and admissible control policies.

The following is an immediate consequence of Little's Law and Markov chain ergodic theory. Letting $R := \lambda/\mu_s$ be the offered (service) load, we have that, for any $\pi \in \Pi(\lambda, N)$ that admits a steady-state distribution,

$$E[Z_1^\pi] = R, \tag{2}$$

and

$$\begin{aligned}\lambda P^\pi(cs) &= \mu_{cs} \cdot E[Z_2^\pi] = \mu_{cs} \cdot (E[Z^\pi] - R)\\ &\le \mu_{cs} \cdot \left[(N-R) \wedge \frac{\lambda p}{\mu_{cs}}\right],\end{aligned} \tag{3}$$

where for two real numbers x and y, $x \wedge y = \min\{x, y\}$. Using observations (2) and (3), the problem (2) may be rewritten as

$$\begin{aligned}&\text{maximize} \quad r\mu_{cs}(E[Z^\pi] - R) - C(N)\\ &\text{subject to} \quad E[W^\pi] \le \bar{W},\\ &\qquad\qquad N \in \mathbb{Z}_+, \ \ \pi \in \Pi(N).\end{aligned} \tag{4}$$

We now introduce the TP control that will be shown to be nearly optimal for (4) when combined with an appropriate staffing rule.

Definition 2.1 (The TP Control). The TP control is defined as follows:

(1) **On customer arrival:** An arriving customer enters service immediately if there are any idle agents.

(2) **On phase-1 completion:** 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 $(Y(t) - N) \le K$ (where K is a predetermined integer).

(3) **On customer departure:** On a customer departure, the customer at the head of the queue will be admitted to service if the queue is nonempty.

For brevity, we use the notation TP[K] to denote TP with threshold K (where K may take negative or 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, on 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 $K \le 0$, TP[K] is a control that uses a threshold on the number of idle agents. Specifically, on service completion with a cross-selling candidate, the agent will exercise cross-selling if the number of idle agents is at least $|K|$.

As TP[K] uses only information on the overall number of customers in the system at the time of service completion, it is a stationary control. Furthermore, as TP disallows a positive queue when there are idle agents, we have that $Q(t) = [Y(t) - N]^+$ and $Z_1(t) + Z_2(t) = N - [Y(t) - N]^-$. Hence, the evolution of the system—queue length, customer in service, and customers in cross-selling—is captured by the Markov process $S(t) := \{Z_2(t), Y(t)\}$. Finally, we also establish that TP is an admissible control (see Lemma F.1, see e-companion). Roughly speaking, the system is stable under TP because of its self-balancing nature. Specifically, whenever the number of customers in the system exceeds the level K, all cross-selling activities are stopped. When this happens, and as $N > R$, the system has sufficient capacity to provide service to all incoming calls.

2.2. Customers Willingness to Listen to a Cross-Selling Offer

It is plausible that the probability that a customer is willing to listen to a cross-selling offer is not fixed, but rather depends on the customer experience up to that point (such as his waiting time, service time, service quality, etc.). This dependence introduces analytical complexity 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

[2] As an alternative to the average-waiting-time constraint, one might consider the commonly used quality of service (QoS) constraint of the form $P\{W > \bar{W}\} \le \delta$. This stipulates that at least a fraction $1-\delta$ of the customers will be answered within $\bar{W}$ units of time. One can verify that our analysis can be extended in a straightforward manner for constraints of this form under the additional assumption that customers are served in a first-come-first-served (FCFS) manner.

the customer service experience. This assumption is reasonable for systems in which waiting times are not too long and service quality is uniformly good. The independence assumption is relaxed in Gurvich et al. (2009), where a cruder form of analysis is performed.

Even with the above simplifying assumption, the problem of optimally staffing and controlling the call center requires keeping track of the two-dimensional process $S(t) = \{Z_2(t), Y(t)\}$. The two-dimensional structure makes the problem difficult to tackle with exact analysis. Instead, we try to identify the structure of nearly optimal solutions via asymptotic analysis. This is the subject of the next section.

3. Asymptotic Analysis

Consider a sequence of systems indexed by the arrival rate λ, which is assumed to grow without bound ($\lambda \to \infty$). The superscript λ denotes quantities associated with the λth system. We consider the following optimization problem, which is obtained from (4) by adding the dependence on λ where needed:

$$\begin{aligned} &\underset{N^\lambda,\,\pi^\lambda}{\text{maximize}} && r\mu_{cs}(E[Z^{\lambda,\pi}] - R) - C^\lambda(N^\lambda) \\ &\text{subject to} && E[W^{\lambda,\pi}] \le \overline{W}^\lambda \\ &&& N^\lambda \in \mathbb{Z}_+,\ \pi^\lambda \in \Pi(\lambda, N^\lambda). \end{aligned} \tag{5}$$

Here and for the rest of the paper, we omit the superscript λ from parameters that are not scaled with λ, such as the service rates, μ_s and μ_{cs}, the expected revenue per customer, r, and the probability, p. The superscript λ is also omitted from R, because R has a trivial dependency on λ given by its definition, $R = \lambda/\mu_s$. In terms of the cost functions, 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 functions via a deterministic relaxation for (6). Specifically, removing the waiting time constraint and replacing $E[Z^{\pi,\lambda}]$ with a free decision variable, a deterministic relaxation for the above problem is given by

$$\begin{aligned} &\underset{N^\lambda,\,z^\lambda}{\text{maximize}} && r\mu_{cs}(z^\lambda - R) - C^\lambda(N^\lambda) \\ &\text{subject to} && z^\lambda \le N^\lambda, \\ &&& N^\lambda \ge R, \\ &&& \mu_{cs}(z^\lambda - R) \le \lambda p, \\ &&& z^\lambda \ge 0,\ N^\lambda \in \mathbb{Z}_+. \end{aligned} \tag{6}$$

The constraints in (7) follow from some basic observations. First, $E[Z^{\pi,\lambda}] \le N^\lambda$ by definition. Also, as we consider only staffing and control solutions that make the system stable, we must have that $N^\lambda \ge R$. Finally, the last constraint follows from (3).

Note that any optimal solution to (7) provides an upper bound on the optimal solution of (6). Indeed, (7) is obtained from (6) by relaxing the restrictions on the waiting times and the fraction of cross-sales. We also note that any optimal solution $(z^{*\lambda}, N^{*\lambda})$ to (7) has $z^{*\lambda} = N^{*\lambda}$ and, in particular, the relaxation is equivalently given by

$$\begin{aligned} &\underset{N^\lambda}{\text{Maximize}} && r\mu_{cs} \cdot (N^\lambda - R) - C^\lambda(N^\lambda) \\ &&& R \le N^\lambda \le R + \frac{\lambda p}{\mu_{cs}} \\ &&& N^\lambda \in \mathbb{R}_+. \end{aligned} \tag{7}$$

The formulation (8) has a critical role in differentiating between the two operating regimes (see discussion following (22)). For each λ, let N_2^λ be the *smallest* optimal solution to (8). Then we make the following assumption:

Assumption 3.1.
(1) *There exist $\beta \ge 0$ and $\gamma \in \mathbb{R}$, such that*

$$\lim_{\lambda\to\infty} \frac{N_2^\lambda - R}{R} = \beta \quad \textit{and} \quad \lim_{\lambda\to\infty} \frac{N_2^\lambda - R(1+\beta)}{\sqrt{R}} = \gamma. \tag{8}$$

In particular, $N_2^\lambda = R + \beta R + \gamma\sqrt{R} + o(\sqrt{R})$.

(2) *There exists a "minimum wage" parameter $c > 0$ such that for all λ and all $N \ge R$,*

$$C^\lambda(N) - C^\lambda(R) \ge c(N - R).$$

Note that, by definition, we have that $N_2^\lambda \le R + \lambda p/\mu_{cs}$, and in particular that $\beta \le \mu_s p/\mu_{cs}$. Assumption 3.1 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, to consider a greater scope for our results, Assumption 3.1 also allows for functions that do scale with λ.

We impose the following assumption on the scaling of the waiting time constraint:

Assumption 3.2. *There exists a constant $\widehat{W} > 0$, such that for all λ, $\overline{W}^\lambda = \widehat{W}/\sqrt{R}$.*

It is important to note that Assumptions 3.1 and 3.2 are not used in any way in the staffing and control recommendation that we provide in §5. For these, one does not have to know the constants $\widehat{W}$, β, γ, and c. Rather, the staffing and control will be given explicitly in terms of cost function $C(\cdot)$ and the waiting time target $\overline{W}$. The scaling in Assumption 3.2 is consistent with many other models in which such square root approximations tend to perform extremely well (see, for example, Gurvich et al. 2008, Borst et al. 2004).

Definition 3.1 (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

$$\limsup_{\lambda\to\infty} \frac{E[W^\lambda]}{\overline{W}^\lambda} \le 1. \tag{9}$$

Let

$$V^\lambda(N^\lambda, \pi^\lambda) := r\mu_{cs}(E[Z^{\lambda,\pi}] - R) - (C^\lambda(N^\lambda) - C^\lambda(R)).$$

Definition 3.2 (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 $\{\widetilde{N}^\lambda, \tilde{\pi}^\lambda\}$ we have

$$\liminf_{\lambda\to\infty} \frac{V^\lambda(N^\lambda, \pi^\lambda)}{V^\lambda(\widetilde{N}^\lambda, \tilde{\pi}^\lambda)} \ge 1. \tag{10}$$

A natural notion of asymptotic optimality would stipulate that a sequence of pairs (N^λ, π^λ) is asymptotically optimal if it is asymptotically feasible and, for all λ large enough, it dominates any other solution; i.e.,

$$\liminf_{\lambda\to\infty} \frac{r\mu_{cs}(E[Z^{\lambda,\pi^\lambda}] - R) - C^\lambda(N^\lambda)}{r\mu_{cs}(E[Z^{\lambda,\tilde{\pi}^\lambda}] - R) - C^\lambda(\widetilde{N}^\lambda)} \ge 1$$

for any other asymptotically feasible sequence $(\widetilde{N}^\lambda, \tilde{\pi}^\lambda)$. This notion, however, is too weak, as it implies that when the optimal solution involves cross-selling to a small fraction of the customers, and, in particular, that

$$r\mu_{cs}(E[Z^{\lambda,\pi}] - R) = o(C(R)),$$

any staffing solution such that $N^\lambda = R + o(R)$ with any reasonable control would be asymptotically optimal. Yet we would like to be able to differentiate between different staffing and control rules even in cases when it is optimal to cross-sell to only a small fraction of the customers. Hence, we normalize around the base cost of $C(R)$, which constitutes a lower bound on feasible staffing levels.

Before stating our main asymptotic optimality result we need a few more definitions. Let

$$N_1^\lambda = \min\{N \in \mathbb{Z}_+ : E[W^{\text{FCFS}}_{\lambda,\mu_s}(N)] \le \overline{W}^\lambda\}. \tag{11}$$

Then N_1 is the minimal number of servers required to meet the service level target if no cross-selling is performed and, in particular, it serves as a lower bound on the number of servers required in the system with cross-selling. Also, define

$$\begin{aligned}\overline{N}_1^\lambda &= \min\{N \in \mathbb{Z}_+,\ N \ge R: \\ &\qquad E[W^{\text{FCFS}}_{\lambda,\mu_s}(N) \mid W^{\text{FCFS}}_{\lambda,\mu_s}(N) > 0] \le \overline{W}^\lambda\} \\ &= \min\left\{N \in \mathbb{Z}_+,\ N \ge R: \frac{1}{N\mu_s - \lambda} \le \overline{W}^\lambda\right\},\end{aligned} \tag{12}$$

where the last equality follows from the fact that, given a wait, the waiting time in an $M/M/N$ queue with arrival rate λ and service rate μ_s has an exponential distribution with rate $N\mu_s - \lambda$ (see, e.g., §§5–9 of Wolff 1989). Also, we will say that two sequences $\{x^\lambda\}$ and $\{y^\lambda\}$ satisfy $x^\lambda \gg y^\lambda$ if $x^\lambda/y^\lambda \to \infty$ 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 3.1. *Consider a sequence of systems, with $\lambda \to \infty$, that satisfies Assumptions* 3.1 *and* 3.2. *Then, with N_1^λ, $\overline{N}_1^\lambda$, and N_2^λ as defined in* (11), (12), *and* (8), *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 use any threshold value K^λ in the interval $[0, \lceil \lambda \overline{W}^\lambda \rceil]$ and $N^\lambda = N_2^\lambda$.*

(2) *Condition* 1 *fails but* $\liminf_{\lambda\to\infty}(N_2^\lambda - R)/(\overline{N}_1^\lambda - R) \ge 1$, *and $\mu_{cs} \ge \mu_s$, then it is asymptotically optimal to set $K^\lambda = 0$ and $N^\lambda = N_2^\lambda$.*

(3) $\mu_s = \mu_{cs}$ *and* $\limsup_{\lambda\to\infty}(N_2^\lambda - R)/(\overline{N}_1^\lambda - R) < 1$, *then it is asymptotically optimal to set $N^\lambda = N^{*\lambda}$ and $K^\lambda = K^{*\lambda}$, where $K^{*\lambda} := K^\lambda(N^{*\lambda})$ and $N^{*\lambda}$ are determined by*

choosing the smallest value of $N^{*\lambda}$ that satisfies

$$N^{*\lambda} := \arg\max_{N \geq N_1^\lambda} \{ r\mu_{cs}(E[Z^{FCFS}_{\lambda,\mu_s}(N) \mid Z^{FCFS}_{\lambda,\mu_s}(N) \geq (N + K^\lambda(N)) \wedge N]) - R - (C^\lambda(N - C^\lambda(R))) \}, \quad (13)$$

where $K^\lambda(N)$ in (13) satisfies

$$K^\lambda(N) := \max_{K \geq -N} \{ E[Q^{FCFS}_{\lambda,\mu_s}(N) \mid Z^{FCFS}_{\lambda,\mu_s}(N) \geq (N+K) \wedge N] \leq \lambda \bar{W}^\lambda \}. \quad (14)$$

Under Condition 1, the value of cross-selling drives the system to use a significant amount of extra staffing to allow for substantial cross-selling. Hence, whenever this condition is satisfied, we say that the system operates in the *PD regime.* In contrast, whenever Conditions 2 or 3 hold, the fraction of customers experiencing cross-selling is relatively small. Under Condition 2, staffing is still determined by profit considerations, and the system is still considered to be operating in the PD regime. Under Condition 3, the focus of the staffing becomes the satisfaction of the quality of service constraint. In that case, we say that the system operates in the *SD regime*. This regime characterization is further discussed in §5, where we provide a prescription for the practical implementation of a staffing and cross-selling rule.

4. The MDP Approach

When any of the conditions in Theorem 3.1 is satisfied, we manage to overcome the two-dimensional nature of the control problem through our asymptotic analysis and show that our proposed scheme that combines a TP threshold control with a regime-based staffing rule is asymptotically optimal. But how well does this scheme perform for a call center of moderate size? How well does it perform if neither of the sufficient conditions of Theorem 3.1 holds? To address these questions, we propose a solution approach to the control component of (2) that applies in great generality, beyond the cases covered by Theorem 3.1. Specifically, following the approach in Gans and Zhou (2003), we consider the solution to a related *MDP*. First, we truncate the state space to reduce the infinite-state space problem into a finite dimensional one. Specifically, we solve the MDP for a *finite buffer* system and show that the optimal solution to the finite-state-space MDP converges to the optimal solution of 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 §5 we show that TP is nearly optimal by numerically comparing its performance to that of the optimal control obtained from the solution of the LP. Note that in this section, λ is fixed and is omitted as a superscript from all notation.

We now turn to the formulation of the MDP 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, in Lemmas 4.1 and 4.2 we show that, if there exists an optimal control, then there always exists one that is *work conserving* and that serves customers *FCFS*. For a fixed N, we redefine $\Pi(N)$ to be the set of nonanticipative, nonpreemptive *feasible* policies. That is, given a staffing level $N \in \mathbb{Z}_+$, $\Pi(N)$ consists of all policies π for which (together with N) steady state exists and $E[W^\pi] \leq \bar{W}$. In particular, it is clear that $\Pi(N)$ will be empty unless $N \geq N_1$, where N_1 is defined in (11).

Lemma 4.1. *Fix λ, μ_s, μ_{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 π. In particular, π' admits a cross-selling rate that is at least as large as that admitted by π.*

Definition 4.1 (Work Conservation). We say that a policy π is work conserving if (i) whenever a customer arrives to find an idle agent, he is immediately admitted to service, and (ii) on a departure of a customer from the system, a waiting customer will be admitted to service if the queue is nonempty.

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 cross-selling even when customers are waiting.

Lemma 4.2 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 4.2. *Fix λ, μ_s, μ_{cs}, and N. Then for any feasible policy $\pi \in \Pi(N)$ that serves customers FCFS, there exists a work-conserving feasible policy $\pi' \in \Pi(N)$ that performs at least as well as π and serves customers FCFS. In particular, π' admits a cross-selling rate that is at least as large as that admitted by π.*

Within the set of work-conserving FCFS policies, we limit our attention to stationary policies. Stationary policies are practical and are typically optimal. Their optimality has not been established in general in a setting with infinite state space. Thus, we impose the restriction on stationary policies as an assumption and redefine the set of admissible policies $\Pi(N)$ accordingly.

Admissible Policies. For a fixed N, the admissible policies $\Pi(N)$ is the set of nonanticipative, nonpreemptive, *stationary, work-conserving, FCFS*, feasible policies.

Note that Lemmas 4.1 and 4.2 prove that the family of work-conserving FCFS policies is optimal, but we have not established that the family of *stationary* work-conserving FCFS policies is optimal. Instead, we henceforth restrict our attention to this new set of admissible policies as defined above.

We are now ready to construct the MDP and the associated LP for a cross-selling system with a *finite buffer*. We will later prove that, for a buffer that is large enough, and for any given admissible 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. This implies that the optimal solution of the constructed LP converges to an optimal solution of the original problem. 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 who find a full buffer on arrival are blocked and do not enter the system.

Because 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 fraction of steps that the corresponding DTMC spends in that state under its stationary distribution; see, e.g., Puterman (1994, pp. 562–563). 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 §4.2 of Altman 1999), one can solve the finite state MDP through an appropriate LP. Note that because of work conservation, in each state $\{Z_2, Y\}$ the action set consists of only two options—cross-sell on 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 because of the finite buffer). The corresponding LP for a system with L trunk lines and N agents is given by

$$\text{Max } \sum_{j=0}^{L} \sum_{i=0}^{j\wedge N} ri\mu_{cs}(\xi(i,j,0)+\xi(i,j,1)) \tag{15}$$

$$\begin{aligned}
\text{s.t } (\lambda + N\mu_s + N\mu_{cs}) \cdot (\xi(i,j,0)+\xi(i,j,1)) \\
= \lambda(\xi(i,j-1,0)+\xi(i,j-1,1))1_{\{j-1\geq 0\}} \\
+ \mu_s((j+1)\wedge N - i)(\xi(i,j+1,0) \\
+ (1-p)\xi(i,j+1,1))1_{\{j+1\leq L\}} \\
+ p\mu_s(j\wedge N-(i-1))\xi(i-1,j,1)1_{\{i-1\geq 0\}} \\
+ \mu_{cs}(i+1)(\xi(i+1,j+1,0) \\
+ \xi(i+1,j+1,1))1_{\{i+1\leq N\}}1_{\{j+1\leq L\}} \\
+ (\xi(i,j,0)+\xi(i,j,1)) \\
\cdot ((N-i)\mu_{cs} + (N-(j\wedge N - i))\mu_s + \lambda 1_{\{j=L\}}), \\
0 \leq j \leq L,\ 0 \leq i \leq j\wedge N,
\end{aligned} \tag{16}$$

$$\sum_{j=0}^{L} \sum_{i=0}^{j\wedge N} (\xi(i,j,0)+\xi(i,j,1)) = 1, \tag{17}$$

and

$$\sum_{j=N}^{L}(j-N)\sum_{i=0}^{N}(\xi(i,j,0)+\xi(i,j,1)) \le \lambda \bar{W}. \tag{18}$$

The system of equations in (16) 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 (16) lists the possible transitions from other states into $(i, j, 0)$ and $(i, j, 1)$, with the corresponding probabilities. Specifically, the first line on the right-hand side of (16) corresponds to transitions caused by arrivals. The second line corresponds to transitions from phase 1 service completions that are not followed by a cross-selling phase. The third line corresponds to transitions caused by phase 1 service completions that are followed by cross-selling. The fourth line corresponds to transitions caused by 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

$$N_1 = \arg\min\{N \in \mathbb{Z}_+ \colon E[W^{\mathrm{FCFS}}_{\lambda,\mu_s}(N)] \le \bar{W}\}. \tag{19}$$

Let $V^*_{\mathrm{LP}}(N, L)$ be the optimal solution of the LP corresponding to a system with N agents and L trunk lines (recall that λ is fixed). For a fixed N, let $V^*(N)$ be the optimal expected profit in (4), that is

$$V^*(N) = \sup_{\pi \in \Pi(N)\colon E[W^\pi] \le \bar{W}} r\mu_{cs}(E[Z^\pi] - R).$$

Then, we have the following:

Proposition 4.1. *Assume* $N \ge N_1$. *Then*

$$\lim_{L\to\infty} V^*{}_{LP}(N, L) = V^*(N). \tag{20}$$

We use the result of Proposition 20 in the next section to illustrate the 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 work-conserving FCFS policies, TP is close to optimal.

5. Discussion of Implementation and Performance Analysis

In this section we interpret the results of the asymptotic analysis in §3 to come up with a practical staffing and cross-selling prescription for call centers.

5.1. Prescription for Implementation

Our asymptotic optimality suggests that a TP control together with a proper staffing rule should perform close to optimality under a set of sufficient conditions, as listed in Theorem 3.1. In those cases, the TP rule offers significant simplifications toward the solution of the original problem. The results in Theorem 3.1 are, however, given in asymptotic terms (for sequences of systems) and it is not a priori clear how to verify that the conditions hold for a given system. Moreover, it is of interest to examine the performance of our proposed policy when the conditions of Theorem 3.1 do not hold. In these cases, the simplicity of the TP rule allows us to construct simple search-based procedures and to compare them to the performance of the optimal solution in cases that are not covered by Theorem 3.1.

First, we note that a key understanding from our asymptotic analysis is the distinction between the PD and the SD regimes. Letting N_2 be a staffing level that solves the optimization problem:

$$\begin{aligned} \text{Maximize}\quad & r\mu_{cs}(N-R) - C(N), \\ \text{s.t.}\quad & R \le N \le R + \frac{\lambda p}{\mu_{cs}}, \\ & N \in \mathbb{R}_+, , \end{aligned} \tag{21}$$

we showed that whenever

$$N_2 - R \gg N_1 - R, \tag{22}$$

it is nearly optimal to staff with $\lceil N_2 \rceil$ agents and use TP$[K]$, with any $K \in [0, \lceil \lambda \bar{W} \rceil]$. Here, N_1 is as defined in (11). The question is when is it reasonable to say that (22) holds. Our rule of thumb for regime characterization says that the PD regime is the correct regime provided that

$$\begin{aligned} &\text{there exists } x > 0 \text{ such that } C'(R(1+y)) < r\mu_{cs} \\ &\quad \text{for all } y \in [0, x]. \end{aligned} \tag{23}$$

Thus, (23) implies that the marginal capacity cost is less than the maximum potential marginal revenue from adding a server ($r\mu_{cs}$). In particular, we expect that in this case the optimal staffing level N_2 in (22) will satisfy $N_2 \gg R$. Moreover, if the function $C(\cdot)$ is convex, we will have that $N_2 = \inf\{x \geq R: C'(x) \leq r\mu_{cs}\}$.

We have implicitly assumed that the derivative exists, but if not (as staffing levels are discrete), the regime condition (23) can be rewritten as

$$\text{there exists } x > 0 \text{ such that } C(R(1+y)) - C(R(1+y)-1) < r\mu_{cs} \text{ for all } y \in [0, x].$$

Theorem 3.1 shows that, if (23) holds, the call center can obtain profit that is almost as high as the one that could be obtained in the absence of any service-level considerations. Hence, whenever (22) holds, the staffing decisions are essentially driven by the profit rather than by the service-level constraint. Even if (22) does not hold, the staffing level N_2 may still be nearly optimal, provided that

$$N_2 \geq \bar{N}_1 := \inf\left\{N \in \mathbb{Z}_+, N \geq R: \frac{1}{N\mu_s - \lambda} \leq \bar{W}\right\}, \quad (24)$$

and we have fast cross-selling, i.e., $\mu_{cs} \geq \mu_s$. In this case, Theorem 3.1 shows that it is nearly optimal to use N_2 with TP[K] with $K = 0$.

In some cases, the revenue from cross-selling is not high enough to justify staffing that is significantly above the basic staffing needed for service-level satisfaction (without cross-selling). In such cases, more care is needed and the optimal staffing level is more sensitive to the service-level target $\bar{W}$. This is the SD regime, which is characterized by the condition $N_2 \leq \bar{N}_1$. In these cases, even nearly optimal staffing and control solutions do not necessarily exhibit the decoupling between profit and service levels observed in the PD regime. Hence, it is a more challenging regime to handle. Whenever the service and cross-selling rates are equal ($\mu_s = \mu_{cs} =: \mu$), Theorem 3.1 shows that it is asymptotically optimal to use N^* for staffing and TP with threshold $K(N^*)$ for cross-selling, with N^* and $K(N^*)$ chosen as in (13) and (14). Thus, when the service rates are equal, while the underlying dynamics are still complex, we are able to derive a nearly optimal solution.

When the sufficient conditions of Theorem 3.1 do not hold, the simplicity of the TP-based rule still allows one to use a simple search-based procedure to determine the threshold and the staffing level. Indeed, focusing on a threshold control only reduces the problem into a simple search over these two parameters, where for each pair (N, K) performance measures may be obtained by computing the steady-state distribution or via simulation. Although the result is not provably optimal, we will show numerically that it performs extremely well. In designing the search procedure, we use the following observations: First, it suffices to consider staffing levels that satisfy $N \geq N_1$. Second, from Theorem 3.1, we know that it suffices to consider staffing levels N such that $N \leq R + \lambda p/\mu_{cs}$. Indeed, by definition we have that $N_2 \leq R + \lambda p/\mu_{cs}$ and we know that N_2 is nearly optimal if (22) holds. Finally, Theorem 3.1 shows that for near optimality it suffices to consider threshold values $K \in [-N, \lceil\lambda\bar{W}\rceil]$. Hence, the search procedure we propose is as follows:

1. Set $N = N_1$. Set $V_{\max} = -C(N_1)$.
2. While $N \leq \lceil R + \lambda p/\mu_{cs}\rceil$:
   (a) Set $K = \lceil\lambda\bar{W}\rceil$
   (b) while $K \geq -N$:
      (i) Evaluate $E[W]$ and $\mu_{cs}(E[Z] - R)$ (by simulation or by solving the balance equations).
      (ii) If $E[W] \leq \bar{W}$:
         (a) If $r\mu_{cs}(E[Z]-R) - C(N) > V_{\max}$, set $N^* = N$ and $K(N^*) = K$.
         (b) Set $N = N+1$ and go to back to 2.
      (iii) If $E[W] > \bar{W}$, set $K = K-1$ and go back to (b).

To summarize, we provide the following implementation prescription:

(1) If (23) holds, then use $\lceil N_2\rceil$ for staffing and TP[K] for cross-selling, with any $K \in [0, \lceil\lambda\bar{W}\rceil]$.

(2) If cross-selling is faster, on average, than service ($\mu_{cs} \geq \mu_s$) and (24) holds, then use $\lceil N_2\rceil$ for staffing and use TP[K] for cross-selling, with $K = 0$.

(3) If (24) does not hold but $\mu_s = \mu_{cs}$, use N^* and $K(N^*)$ as determined by (13) and (14), respectively.

(4) In all other cases, perform the search procedure above.

Of course, relying on the search procedure in the cases in which the conditions of Theorem 3.1 do not hold has no performance guarantees. The remainder of this section is dedicated to a detailed numerical study of the performance of our proposed two-parameter search procedure. Particularly, we demonstrate that our scheme performs extremely well, even when it is not *provably* nearly optimal.

5.2. Performance Evaluation

This section is divided to two parts. First, we analyze the performance of the TP rule for different (but fixed) staffing levels. This part can be regarded as dealing with the problem of finding optimal cross-selling policies for given staffing levels. In the second part we address the joint staffing and control problem and, in the process, provide numerical support for the distinction between the PD and the SD regimes. In both parts we compare the resulting performance against that of the optimal solutions (among all stationary work-conserving FCFS policies) as found via the explicit solution of the associated MDP with a large enough buffer.[3]

Performance of TP for Given Staffing Levels. Here we compare the value obtained through the solution of the original optimization problem via MDP versus the performance obtained using TP. We show that TP performs well beyond the scope covered by our sufficient conditions and that it also exhibits good performance in relatively small call centers.

We experiment with two different call centers. The first, with an offered load of $R=30$, represents a relatively small call center. The second, with an offered load of $R=100$, represents a medium-size call center. We assume that the service-level constraint is given by $E[W] \le 0.1/\mu_s$ which, in a setting with a mean service time of 5 minutes, corresponds to an upper bound of 30 seconds on the average waiting time. We fix $\mu_s = 1$ throughout but allow for different ratios of μ_{cs}/μ_s, ranging from 0.1 to 10, where the extreme cases $\mu_{cs} = 0.1$ and $\mu_{cs} = 10$ represent very slow and very fast cross-selling, respectively. In all the experiments r is assumed to be 1, so the expected cross-selling revenue is essentially equal to the cross-selling rate. We will initially also assume that $p=1$, but we will later consider values of $p<1$.

We use Equation (11) to find N_1, the fewest servers required to satisfy the service level in the absence of any cross-selling activity. We find that $N_1 = 34$ for $R = 30$ and $N_1 = 106$ for $R = 100$. We then vary the staffing level, N, in steps of four in the range $\{N_1, \dots, N_1+40\}$. The range of staffing levels that we examine is large enough to cover both operating regimes and a wide range of optimal thresholds—that are both negative when in the SD regime (see below) and positive when the staffing levels are sufficiently greater than N_1.

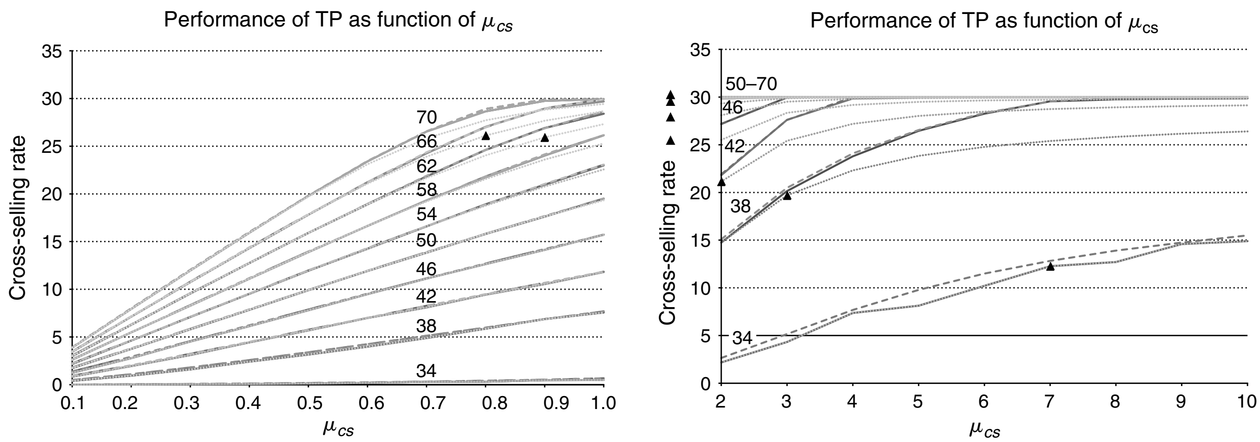
The experiments then proceed as follows: fixing the staffing level N, we vary the value of μ_{cs} and find the revenue under TP and the revenue obtained from the optimal solution of the MDP. For the former, given μ_{cs} and N, we fix the policy to be TP but vary the threshold, K, using two methods: (i) using our search procedure described in §5.1 (this search is limited to $K \in [-N, \lceil \lambda \bar{W} \rceil]$), and (ii) by an exhaustive search to find the threshold that maximizes the revenue among all possible thresholds. For each combination of μ_{cs}, N, and K, the revenue and the queue length are calculated by solving the balance equations of the CTMC, $S(t) := \{Z_2(t), Y(t)\}$.

The results are displayed in Figures 1 and 2 for $R=30$ and $R=100$, respectively. In each of these figures, the left-hand graph focuses on slow cross-selling ($\mu_{cs} \in \{0.1, 0.2, \dots, 1\}$) and the right-hand graph focuses on fast cross-selling ($\mu_{cs} \in \{2, \dots, 10\}$). In all graphs a separate curve is displayed for each staffing level, and the corresponding staffing level is displayed above the relevant curves. Note that for each value of staffing level we display three different series. The *dotted line* corresponds to the revenue obtained when using TP and searching for the threshold in the interval $[-N, \lceil \lambda \bar{W} \rceil]$. Whenever relevant, all values to the right of the marker (a triangle) correspond to values of cross-selling rate μ_{cs}, such that the staffing level N satisfies $N > R + \lambda p/\mu_{cs}$. Those particular staffing levels are never recommended by our scheme. The *solid line* corresponds to the revenue when using TP with the best threshold. Finally, the *dashed line* correspond; to the optimal revenue obtained from the MDP.

For almost all of the combinations of parameters examined, the solid and the dashed lines agree almost completely—the gaps are less than 2% (mostly, significantly less). The only case where there is a visible gap is the curve in Figure 1 that corresponds to fast cross-selling (the right-hand graph) and to the staffing level $N = N_1 = 34$. When the size of the call center increases (as in $R=100$), this gap virtually disappears. The gap between these two lines and the dotted line are more significant. However, this gap is particularly notable

[3] In each example we increased the buffer size until the effect on the profits was negligible. We found that a buffer size of 100 was sufficient for most examples.

Figure 1 **Performance of the TP Rule for a Small-to-Medium Call Center** ($R = 30$)



for staffing values that are greater than $R + \lambda p/\mu_{cs}$—these are staffing values that our scheme does not recommend using. This underlines the importance of jointly solving for the staffing and control problems. If staffing is determined according to our proposed rule, threshold values may be restricted to the interval $[-N, \lceil \lambda \bar{W} \rceil]$. If staffing levels are beyond the recommended range, a full search over all possible threshold values is needed.

The case of $R = 100$ also serves to emphasize another important point. In the introduction we mention that using a threshold on the queue length is a common heuristic; i.e., cross-selling is exercised on service completion only when the number of callers in the *queue* is below a certain threshold. Examining the optimal thresholds generated for this example through our numerical analysis reveals the need to use thresholds on the *number of idle agents* in some cases (as allowed by TP by setting $K < 0$), rather than on the number of customers in the queue. Focusing on the case $R = 100$ and $\mu_{cs} = 1/3$, the optimal thresholds for the different staffing levels are given in Table 1, 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, to satisfy feasibility.

Figure 2 **Performance of the TP Rule for a Medium-to-Large Call Center** ($R = 100$)

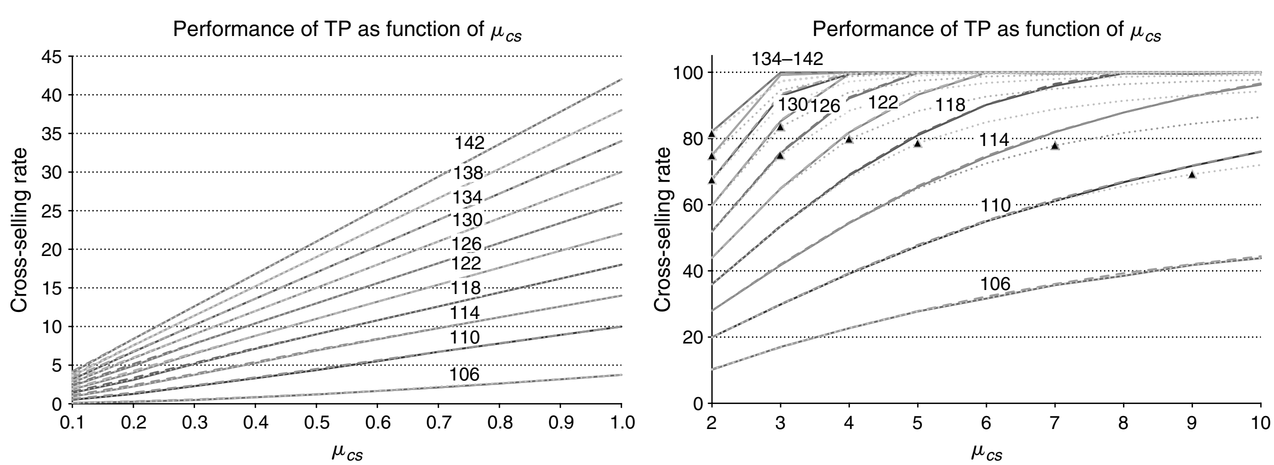


Table 1 Threshold Values for $R = 100$

Staffing	106	110	114	118	122	126	130	136	142
Threshold	−20	−10	−6	−3	−1	1	2	3	4

We next examine the performance of TP as a function of the fraction of cross-selling candidates, p. We show that TP also performs well for $p < 1$. For this experiment we focus on the smaller call center with $R = 30$. We also fix $\mu_{cs} = 0.5$ and, as before, $r = 1$. We then vary the staffing levels in $\{N_1 = 34, \ldots, 70\}$ and $p \in \{0.1, \ldots, 1\}$. Figure 3 has, for each staffing level, a curve that displays the revenues from cross-selling as a function of p under both TP and the optimal solution from the MDP.

In all the cases we have analyzed, the performance of TP is remarkably good. The results of these experiments support the claim that TP exhibits good performance in great generality for large as well as moderate size call centers and beyond the scope covered by the sufficient conditions of Theorem 3.1. In the above experiments we have considered exogenous staffing levels. We now consider the joint staffing and control problem with a focus on regime selection.

Staffing Optimization and Regime Selection. We now provide a numerical example that illustrates the staffing selection procedure and relate it to the regime choice—profit driven versus service driven. Consider the setting $R = 30$, $\mu_s = 1$, and $\mu_{cs} = 1/3$, $p = 1$, and $r = 1$. We initially assume that the cost function is linear, i.e, that $C(N) = C \times N$. In this case, the optimization problem (22) has a simple solution (as a function of the cost coefficient C): If $C \geq r\mu_{cs} = 1/3$, then $N_2 = R = 30$. If $C < r\mu_{cs}$, then $N_2 = R + \lambda p/\mu_{cs} = 120$. We calculate N_1 as defined in (11) to find that $N_1 = 34$. Hence, if $C < r\mu_{cs}$, we have $N_2 - R \gg N_1 - R$ and Theorem 3.1 asserts that the system is in the PD regime and that the asymptotically optimal staffing level would be N_2^λ. If $C \geq r/\mu_{cs}$, we expect the system to be operating in the SD regime.

Figure 3 Performance of the TP Rule as a Function of p ($R = 30$)

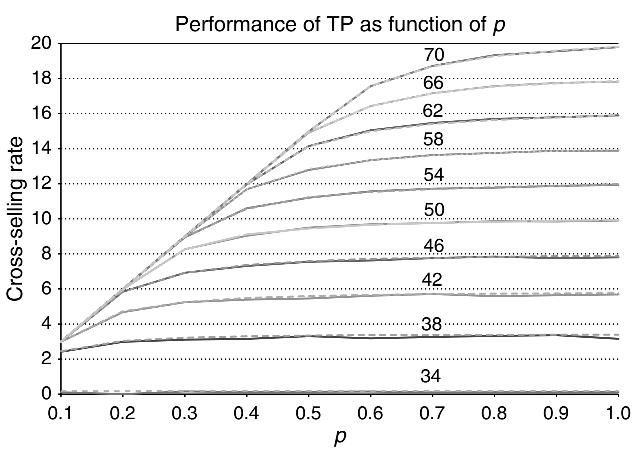
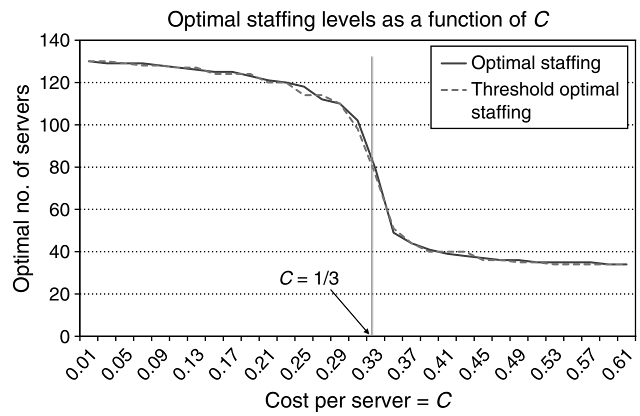


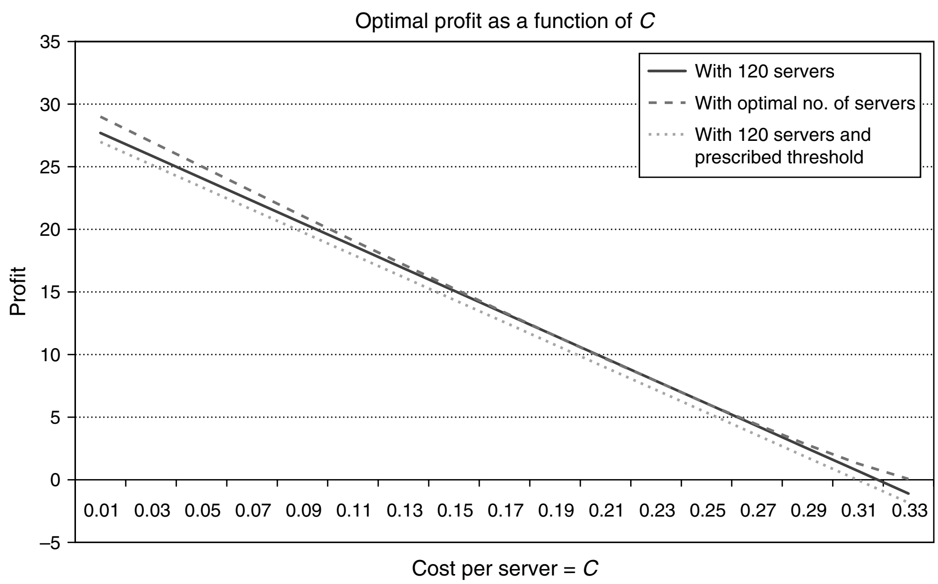
Figure 4 Optimal Number Servers as a Function of the Per Server Staffing Cost



We now perform a numerical experiment to illustrate the strength of this regime characterization result. First, for each value of C we find the optimal staffing by solving the MDP for each staffing level and choosing the optimal one. The series that we obtain corresponds to the solid-line series in Figure 4. In addition, for each value of C we find the optimal staffing and threshold levels, assuming a threshold policy and performing a full search to find the best threshold value and the corresponding optimal staffing level. The results are depicted in the dashed line in Figure 4. The vertical line represents the critical value, $C = 1/3$.

We make the following two observations with respect to Figure 4. First, it shows that the TP-based staffing works extremely well for various cost figures. In particular, the staffing that we obtain by using TP and a search for staffing is close to the optimal

Figure 5 **Comparison of Asymptotically Optimal Staffing vs. Optimal Staffing for $C < 1/3$**



one that we obtain from the MDP.[4] Second, we see that there is, indeed, a very sharp transition from the PD regime to the SD regime when C approaches 1/3, as predicted by our theory. The same theorem also predicts that in the region $C < 1/3$ the staffing level $N_2 = 120$ should be *nearly* optimal (note that the graph proposes staffing levels in the range (80, 130)). In Figure 5 we plot the profit (as a function of the coefficient C) when fixing the staffing level at 120 for all $C < 1/3$ (with threshold control obtained by fully searching for the best threshold value) versus the profit obtained when using the optimal staffing level for each value of C (with control as determined by the MDP). As can be seen, these profits are close, so that using 120 servers, as proposed by our scheme for the PD regime, results in a nearly optimal solution. Moreover, even if one does not search for the best threshold value, but uses an arbitrary value of $K \in [0, \lceil \lambda \overline{W} \rceil]$, the performance with 120 servers is still close to optimal. This can be seen as the dotted line in Figure 5 that corresponds to the expected profit when using 120 servers and a fixed threshold $K = 2 < \lambda \overline{W} = 3$.

We now provide a numerical illustration of this point using the setting with $R = 30$ and $\mu_{cs} = 1/3$.

[4] The maximal deviation in the graph is 4 servers, obtained at the value of $C = 0.31$. At the same time the difference in profits is less than 3%.

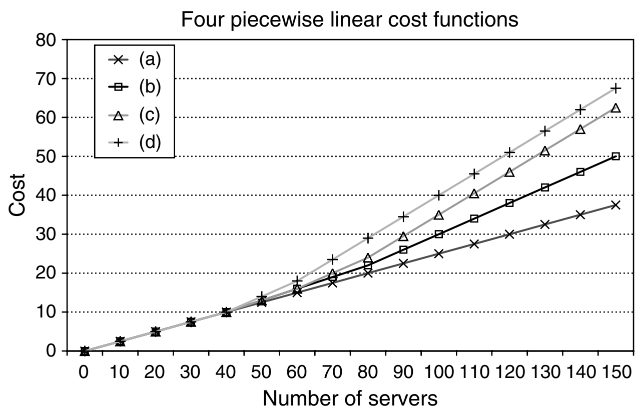
We assume a piecewise linear cost function with three break points. Specifically, we assume that

$$C(x) = \begin{cases} c_1 \times x & x \le 40, \\ c_1 \times 40 + c_2 \times (x - 40) & 40 \le x \le 60 \\ c_1 \times 40 + c_2 \times 20 + c_3 \times (x - 60) & 60 \le x \le 80 \\ c_1 \times 40 + c_2 \times 20 + c_3 \times 20 & \\ \quad + c_3 \times (x - 80) & x \ge 80. \end{cases}$$

The convexity is imposed by assuming that $c_1 \le c_2 \le c_3 \le c_4$.

We now examine different scenarios for values of c_i, $i = 1, \ldots, 4$ and compare the optimal profit (obtained through the MDP and a staffing-level search) versus the one obtained using the asymptotically optimal staffing obtained from the solution to (8) (denoted by N_2) together with TP (the optimal threshold is found through search). In all these scenarios we assume that $c_1 < 1/3$. The four staffing-cost functions that we consider are depicted in Figure 6, and their parameters and resulting profits appear in Table 2. We note that one of the main differences between the various parameter choices in Table 2 is that the switching point from a marginal cost that is less than 1/3 to one that is greater than 1/3 is obtained at a different value. The implication of having different switching points is that N_2 from (22) obtains a different value

Figure 6 **The Piecewise Linear Functions**



for each cost function. This allows us to compare the performance of our solution procedure for different staffing levels. The staffing and profit comparisons in Table 2 illustrate that in the PD regime staffing using N_2 with TP for cross-selling control results in performance that is close to optimal.

5.3. 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 (1981), Borst et al. (2004), Armony (2005), Gurvich et al. (2008), and 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 (22)) to cross-sell to a fraction $f^* > 0$ of the customers, a staffing level of the form

$$N = R + \frac{f^*\lambda}{\mu_{cs}} = R + f^*\frac{\mu_s}{\mu_{cs}}R \tag{25}$$

is asymptotically optimal.

A direct implementation of SRSS is, hence, inappropriate in a cross-selling environment. If, however, we define a different notion of offered load, $R' := R + f^*\lambda/\mu_{cs}$, that takes into account the optimal amount of cross selling, then we have that it is asymptotically optimal to use $N = R'$ agents. The asymptotically optimal staffing could then be regarded as a special case of SRSS with the coefficient of the square-root term being equal to 0. This observation underscores a crucial difference between the SRSS rule for a pure service call center and the one we propose for the cross-selling call center. Although 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 PD regime. In particular, in the cross-selling system one may ignore the square root component, because 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 TPy 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 $1/\mu_s + f^*/\mu_{cs}$. In particular, the estimate of the offered load will be $R' = \lambda/\mu_s + \lambda f^*/\mu_{cs}$, so that use of the SRSS rule 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

Table 2 **Staffing Comparison with Piecewise Linear Cost**

	c_1	c_2	c_3	c_4	Optimal staffing	N_2	Optimal profit	Profit for N_2
(a)	0.25	0.25	0.25	0.25	118	120	6.0835	6.0544
(b)	0.25	0.30	0.30	0.40	80	80	2.0581	2.0581
(c)	0.25	0.30	0.40	0.55	60	60	1.4425	1.4425
(d)	0.25	0.40	0.55	0.55	40	40	0.3812	0.3812

will remain suboptimal regardless of the control used. Indeed, assume that the call center uses $N = R' + \gamma\sqrt{R'}$ agents with R' now equal to $R + f\lambda/\mu_{cs}$ for some $f \neq f^*$. Then, because 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 $\mu_{cs}(N - R) = \lambda f + O(\sqrt{R'})$. The new estimate of the average service time (which is obtained by averaging over all customers) will then be $1/\mu_s + f/\mu_{cs} + o(1)$. Consequently, the call center will continue to perform suboptimally. 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.

6. 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) this work (together with Gurvich et al. 2009) is the first 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 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 that 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 refine the commonly used TP control rule but to also propose a corresponding staffing rule. Under a set of assumptions on system parameters, the staffing and control rules are asymptotically optimal in the limit as the system size grows. We have also shown numerically that they perform well in various settings for systems with a relatively small arrival rate.

A naive approach would be to determine the staffing level ignoring the existence of cross-selling, taking into account only staffing costs, service-level constraints, and service time. This approach can lead to far from optimal solutions. To properly manage cross-selling, one should take into account the value of cross-selling and the associated additional handling time when making staffing and control decisions. The simple structure of our solution allows managers to easily incorporate these data in addition to prespecified service-level targets into the staffing and control decision.

This more comprehensive approach toward staffing and control of call centers with cross-selling can prevent service-level degradation when transforming a pure-service call center into one that combines service and cross-selling activities. We have shown that call centers with valuable cross-selling have the capability to provide very short waiting times and, at the same time, obtain revenues that are close to the optimal revenues obtained in the absence of waiting-time constraints.

Many questions remain unanswered with respect to the operational aspects of cross-selling in call centers. Particularly, it is unclear how customers' experience prior to the cross-selling offering affects their tendency to (i) listen to the offer and (ii) 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 sell. A follow-up paper (Gurvich et al. 2009) 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.

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Acknowledgments

This work was first submitted when I. Gurvich was a graduate student at Columbia University.

Appendix. Table of Main Notation

λ	arrival rate
μ_s	service rate
μ_{cs}	cross-selling rate
$R = \lambda/\mu_s$	offered load
N	number of servers
p	fraction of callers who are cross-selling candidates
r	expected revenue from cross-selling
π	control policy
Z_1	number of servers providing service
Z_2	number of servers performing cross-selling
$Z = Z_1 + Z_2$	total number of busy servers
$I = N - Z$	total number of idle servers
Q	queue length
$Y = Z + Q$	total number of customers in the system
W	virtual waiting time
$\bar{W}$	upper bound on the expected wait
$P\{cs\}$	long run proportion of customers that go through cross selling
$C(\cdot)$	staffing cost function
TP$[K]$	threshold priority policy with threshold level K

References

Akşin, O. Z., P. T. Harker. 1999. To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers. *J. Service Res.* **2**(1) 19–33.

Altman, E. 1999. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, London.

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

Armony, M., C. Maglaras. 2004a. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* **52**(2) 271–292.

Armony, M., C. Maglaras. 2004b. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., A. Mandelbaum. 2009. Routing and staffing in large-scaled service systems with heterogeneous servers and impatient customers. *Oper Res.* Forthcoming.

Aydin, G., S. Ziya. 2008. Pricing promotional products under upselling. *Manufacturing Service Oper. Management* **10** 360–376.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.

Byers, R. E., K. C. So. 2004. The value of information-based cross-sales policies in telephone service centers. Working paper, Graduate School of Management, University of California, Irvine.

Byers, R. E., K. C. So. 2007. A mathematical model for evaluating cross-sales policies in telephone service centers. *Manufacturing Service Oper. Management* **9**(1) 1–8.

Eichfeld, A., T. D. Morse, K. W. Scott. 2006. Using call centers to boost revenue. *McKinsey Quart.* (May).

Gans, N., Y.-P. Zhou. 2003. A call-routing problem with service-level constraints. *Oper. Res.* **51**(2) 255–271.

Gunes, E. D., O. Z. Akşin. 2004. Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing Service Oper. Management* **6**(4) 338–357.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* **11**(2) 237–253.

Gurvich, I., A. Zeevi. 2007. Validity of heavy-traffic steady-state approximations in open queueing networks: Sufficient conditions involving state-space collapse. Working paper, Columbia University, New York.

Gurvich, I., M. Armony, A. Mandelbaum. 2008. Service level differentiation in call centers with fully flexible servers. *Management Sci.* **54**(2) 279–294.

Gurvich, I., M. Armony, C. Maglaras. 2009. Cross-selling in call centers with a heterogeneous population. *Oper. Res.* **57**(2) 299–313.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–587.

Mandelbaum, A., S. Zeltyn. 2007. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res*. **57**(5) 1189–1205.

Netessine, S., S. Savin, W. Xiao. 2006. Dynamic revenue management through cross-selling in E-commerce retailing. *Oper. Res.* **54**(5) 893–913.

Örmeci, E. L., O. Z. Akşin. 2006. Revenue management through dynamic cross-selling in call centers. *Production Oper. Management*. Forthcoming.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues.* Prentice Hall, Englewood Cliffs, NJ.