

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

Mor Armony¹

Itay Gurvich²

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 large. 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 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 but performs extremely well even for relatively small systems.

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,

¹Stern School of Business, New York University, marmony@stern.nyu.edu

²Graduate School of Business, Columbia University, ig2126@columbia.edu

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 McKinsey report [11], 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, [11] 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 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 pre-determined performance target. Hence, in this pure service context the operational tradeoff is clear: Staffing cost Vs. Service Level. When sales and promotions are introduced, however, one should add the potential revenue from promotions 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 firm 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 careful analysis is required.

Closely related to staffing is the 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 rather involved. As cross-selling of a customer can start only upon his service completion, an optimal control likely to use information about whether an agent is providing service or is 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 part of the interaction to the cross-selling part and, consequently, it is only the aggregate information that 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 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 heuristic we identify cases in which cross-selling should not be exercised even when there are some 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 description that distinguishes between agents providing service and those engaged in cross-selling.
 - (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 Revenue versus the Service Driven 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. [14] and has been formalized by Gurvich and Zeevi [16] 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 also stated in §4. A Markov Decision Process (MDP) approach is described and explored in §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 relegating some of the formal proofs 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 cross-selling 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 various 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. Örmeci and Akşin [20], 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 [20] 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 [13] 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 [9, 10] 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 [10] and numerical results are given for the multi-server case in [9].

To position our paper in the context of the literature introduced above, note that 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 becomes simpler. Indeed, our solution provides conditions under which the staffing level that maximizes the expected revenue from cross-selling simultaneously achieves extremely low waiting times.

In a follow-up paper [14], 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. We 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 [17]. In particular, we follow the asymptotic optimality framework approach first used by Borst et al. [8], and adapted later to more complex settings ([3], [4], [5], [6], [15] and [18]). The asymptotic regime that we use has been shown to be extremely robust even in relatively small systems (see Borst et al. [8]); 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 for proofs in our framework. Instead, we introduce a proof methodology that was later formalized in Gurvich and Zeevi [16] 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 [7] and Netessine et. al. [19].

2. Problem Formulation

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 . 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 μ_{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 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 portion \bar{p} of the customers are potential cross-selling candidates. The remaining customers are not considered profitable and are never cross-sold. Whether or not a specific customer is a profitable candidate is recognized by the agent himself during his interaction with the customer. It is important to note that, even if an agent decides to cross-sell a caller, the latter 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 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. A schematic illustration of the system is given in Figure 1, in which N is the number of CSRs.

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 or not to exercise this cross-selling opportunity upon 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 , $i = 1, 2$ 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 $Y^\pi(t)$

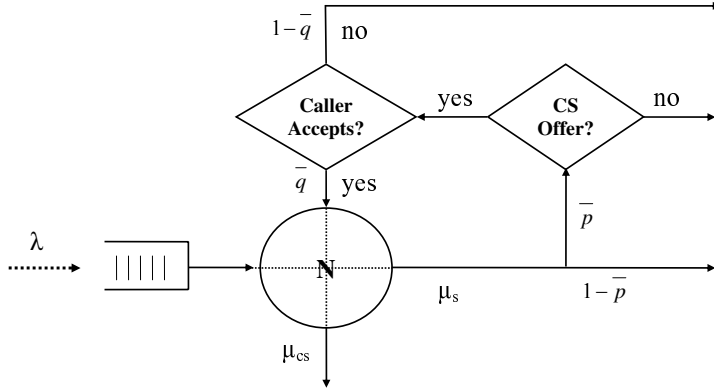


Figure 1: A Schematic Description of a Call Center with Cross-Selling

is the overall number of customers in the system at time t . That is, $Y^\pi(t) = Z^\pi(t) + Q^\pi(t)$. Finally, we let $W^\pi(t)$ be the virtual waiting time at time t . In all of 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, 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, $S^\pi(t) = \{Z_i^\pi(t), i = 1, 2; Q^\pi(t)\}$ is a Markov process under any stationary control. Let $A(t)$ be the number of calls by time t and x_k^π , $k = 1, 2, \dots$ be equal to 1 if the k^{th} arriving customer went through phase 2 and equal to 0 otherwise. Then, if steady-state exists under π , we let $P^\pi(cs)$ be the long-run proportion of customers that go through cross-selling, i.e.,

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

Finally, we note that by the PASTA property the steady-state virtual waiting time and the steady-state waiting time at arrival epochs coincide. The profit maximization formulation is then as follows:

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

Here the average steady-state waiting time $E[W^\pi]$ is constrained to be less than a 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. We comment that, as an alternative to the average-waiting-time constraint, one might

consider the commonly used Quality of Service (*QoS*) constraint of the form $P\{W > \bar{W}\} \leq \delta$. This stipulates that at least a fraction $1 - \delta$ of the customers will be answered within \bar{W} units of time. All the insights of our analysis go through for constraints of this form under the additional assumption that customers are served in a First Come First Served (FCFS) manner.

The control policy π 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 λ, μ_s, μ_{cs} , we say that π is an admissible policy if it is non-preemptive, non-anticipative and*

$$\lim_{t \rightarrow \infty} \frac{E[Q_\pi(t)]}{t} = 0. \quad (2)$$

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 system through its stability conditions. When the parameters λ, μ_s and μ_{cs} are fixed, we will omit them from the notation and use instead the notation $\Pi(N)$. The number of agents, N , will be itself 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. The word “maximize” should be formally interpreted as taking the supremum over all staffing levels and admissible policies.

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

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

Using Lemma 2.1 we re-write (1) as

$$\begin{aligned} & \text{maximize} && r\mu_{cs}(E[Z^\pi] - R) - C(N) \\ & \text{subject to} && E[W] \leq \bar{W}, \\ & && N \in \mathbb{Z}_+, \pi \in \Pi(N). \end{aligned} \quad (3)$$

We now introduce the Threshold Priority (TP) control that we will show to be nearly optimal for (3) when combined with appropriate staffing levels.

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

- (1) **Upon customer arrival:** An arriving customer enters service immediately if there are any idle agents.
- (2) **Upon 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 $(Y(t) - N) \leq K$ (where K is a pre-determined integer).
- (3) **Upon customer departure:** Upon a customer departure, the customer at the head of the queue will be admitted to service if the queue is non-empty.

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 $|K|$.

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 and the resulting process $S(t) := \{Z_2(t), Y(t)\}$ is a Markov process. Furthermore, as TP disallows a positive queue when there are idle agents, we have that $Q(t) = [Y(t) - N]^+$ and $Z_1(t) + Z_2(t) = N - [Y(t) - N]^-$. The following Lemma shows that TP is an admissible control.

Lemma 2.2 *Fix λ and $N > R$ and assume that $TP[K]$ is used with some $K \geq -N$. Then, the Markov process $S(t)$ admits a unique steady-state distribution. Consequently, TP is admissible in the sense of Definition 2.1.*

The stability stated in Lemma 2.2 is a consequence of the self balancing nature of TP. The intuitive reasoning is as follow: 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.

We end this section with a brief comment on our modeling assumptions. It is plausible that, in reality, the probability that a customer will be willing to listen to a cross-selling offer is not

fixed, but rather depends on the customer experience up to that point, e.g., (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. [14] where a cruder form of analysis is performed.

3. The Proposed Solution

The purpose of this section is to give an informal description of our joint staffing-and-control scheme which is sufficient for practical purposes. A more rigorous description combined with an introduction to our asymptotic framework is given in §4.

Our proposed solution includes a staffing recommendation N combined with the TP for dynamic control. For most cases, we will provide closed form rules that are nearly optimal in an appropriate sense. Furthermore, in cases in which we cannot provide closed form rules, the structural simplicity of the TP rule will reduce the choice of staffing and control rules to 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. Our mathematical results allow us to provide a few observations that simplify this search:

- **An upper bound for the threshold:** It suffices to consider threshold values $K \in [-N, \lambda\bar{W})$.
- **A lower bound for the number of agents:** It suffices to consider staffing levels N such that

$$N \geq N_1 := \arg \min \{N \in \mathbb{Z}_+ : E[W_{\lambda, \mu_s}^{FCFS}(N)] \leq \bar{W}\}, \quad (4)$$

where $W_{\lambda, \mu_s}^{FCFS}(N)$ is the steady state waiting time in an $M/M/N$ system with arrival rate λ , service rate μ_s and FCFS service discipline.

- **An upper bound for the number of agents:** It suffices to consider staffing levels N such that

$$N \leq R + \frac{\lambda p}{\mu_{cs}},$$

The second observation 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. The other two observations are consequences of our asymptotic analysis and, in particular, Theorem 4.1 and Theorem A.1 of the technical appendix.

In most cases, however, we can replace the search mechanism by closed form nearly optimal solutions that are based on the TP rule. In doing so, we identify two operational regimes that are conceptually different. In the first, the optimal staffing levels are determined to a great extent by revenue maximization considerations and are relatively insensitive to the service-level targets. In this regime, once the staffing level has been determined – independently of the target waiting time \bar{W} – the service-level constraint will be satisfied by the appropriate choice of the threshold level. We call this the **Revenue-Driven** (RD) regime. In the second case, the staffing decisions need to carefully take into account the service-level constraints. We call this the **Service-Driven** (SD) regime.

3.1 The Revenue Driven Regime

To introduce our proposed solution here, consider first the following deterministic relaxation of (3) (relaxing both the waiting time constraint and the integrality restriction on N):

$$\begin{aligned} & \text{Maximize} && r\mu_{cs}x - C(N), \\ & \text{s.t.} && x \leq (N - R) \wedge \frac{\lambda p}{\mu_{cs}}, \\ & && x, N \in \mathbb{R}_+, N \geq R, \end{aligned} \tag{5}$$

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 (5) will satisfy $N^* = R + x^*$, so that one could re-write (5) as

$$\begin{aligned} & \text{Maximize} && r\mu_{cs}(N - R) - C(N), \\ & \text{s.t.} && R \leq N \leq R + \frac{\lambda p}{\mu_{cs}}, \\ & && N \in \mathbb{R}_+. \end{aligned} \tag{6}$$

We denote the optimal value of N in (6) by N_2 . We will show that whenever

$$N_2 - R \gg N_1 - R, \tag{7}$$

then **it is nearly optimal to staff with $\lceil N_2 \rceil$ agents and use $TP[K]$, with any $K \in [0, \lambda\bar{W}]$. Here, N_1 is as defined in (4) and, informally, we say that $x \gg y$ if x is much greater than y . The**

cause for this useful simplification is as follows: when the value of cross-selling is high enough so that it is optimal to staff with many more agents than required for service-level feasibility (without cross-selling), the service-level constraint ceases to play a role in the staffing vs. revenue tradeoff. In these cases, the call center can achieve revenues that are almost as high as those that could be achieved in the absence of any service level considerations. Hence, whenever (7) holds, the staffing decisions are essentially driven by the revenues rather than by the service-levels.

N_2 is the recommended staffing level also when

$$N_2 \geq \bar{N}_1 := \inf\{N \in \mathbb{Z}_+, N \geq R : \frac{1}{N\mu_s - \lambda} \leq \bar{W}\}.$$

If this condition is satisfied but (7) is not, the proposed threshold is $K = 0$.

In cases in which $N_2 < \bar{N}_1$, more care is needed and the staffing levels will be more sensitive to the service-level target \bar{W} . This is the Service-Driven regime.

3.2 The Service-Driven Regime

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 these case, even nearly optimal staffing and control solutions do not necessarily exhibit the decoupling between revenue and service-levels observed in the RD regime. This is a more challenging regime to handle. Whenever the service and cross-selling rates are identical ($\mu_s = \mu_{cs} =: \mu$) we can devise a nearly-optimal solution in closed form. This nearly optimal solution is based on a TP rule and staffing level N^* chosen as follows:

$$N^* = \arg \max_{N \geq N_1} r\mu (E [Z_{\lambda,\mu}^{FCFS}(N) | Z_{\lambda,\mu}^{FCFS}(N) \geq (N + K(N)) \wedge N] - R) - (C(N) - C(R)), \quad (8)$$

and

$$K(N) = \max_{K \geq -N} \{E[Q_{\lambda,\mu}^{FCFS}(N) | Z_{\lambda,\mu}^{FCFS}(N) \geq (N + K) \wedge N] \leq \lambda \bar{W}\}. \quad (9)$$

Here, $Z_{\lambda,\mu}^{FCFS}(N)$ and $Q_{\lambda,\mu}^{FCFS}(N)$ are, respectively, the steady-state number of busy servers and the number of customers in queue in an $M/M/N$ queue with arrival rate λ , service rate μ and N servers.

Thus, when the service-rates are equal, while the underlying dynamics are still complex, we are able to derive a nearly optimal solution. Together with the RD regime, we have provided closed-form solutions under different conditions but not in full generality. Our objective next is to prove the near optimality of our proposed solution under these conditions. But we are also interested in

going beyond these sufficient conditions. Accordingly, we will compare the performance of the TP-based search procedure, as outlined at the beginning of this section, to that of the true optimal solution obtained by the solution of a complex Markov Decision Process; see §5.

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 λ , which is assumed to grow without bound ($\lambda \rightarrow \infty$). The superscript λ is used to denote quantities associated with the λ^{th} system. We consider the optimization problem

$$\begin{aligned} & \text{maximize} && r\mu_{cs}(E[Z^{\lambda,\pi}] - R) - C^\lambda(N^\lambda) \\ & \text{subject to} && E[W^{\lambda,\pi}] \leq \bar{W}^\lambda \\ & && N^\lambda \in \mathbb{Z}_+, \pi^\lambda \in \Pi(\lambda, N^\lambda), \end{aligned} \tag{10}$$

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 , since R has a trivial dependency on λ given by its definition $R = \frac{\lambda}{\mu_s}$.

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 function through the solution to the deterministic relaxation (6). In particular, the scaled version of (6) is given by

$$\begin{aligned} & \text{Maximize} && r\mu_{cs} \cdot (N^\lambda - R) - C^\lambda(N^\lambda) \\ & && R \leq N^\lambda \leq R + \frac{\lambda p}{\mu_{cs}} \\ & && N^\lambda \in \mathbb{R}_+. \end{aligned} \tag{11}$$

For each λ , let N_2^λ be the **smallest** optimal solution to (11). Then, we make the following assumption:

Assumption 4.1

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

$$\lim_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{R} = \beta, \text{ and } \lim_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R(1 + \beta)}{\sqrt{R}} = \gamma. \tag{12}$$

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 $N \geq R$,

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

Note that, by definition, we have that $N_2^\lambda \leq R + \frac{\lambda p}{\mu_{cs}}$, and in particular that $\beta \leq \frac{\mu_{sp}}{\mu_{cs}}$. 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 greater scope for our results, Assumption 4.1 also allows for functions that do scale with λ . Such scaling allows to reflect the greater flexibility that large call centers have. For example, it is plausible that an addition of 20 agents to a shift would be cheaper for a call-center with thousands of agents than for a small call-center with only tens of agents in a shift. A simple example for such scalable cost functions can be given by a piecewise linear convex function with derivative:

$$C'^\lambda(x) = \begin{cases} c_0, & R < x < R + \tilde{c}\lambda \\ c_n, & R + \tilde{c}\lambda + (n-1)\sqrt{\lambda} < x < R + \tilde{c}\lambda + n\sqrt{\lambda}, 1 \leq n \leq M, \\ c_{M+1}, & R + \tilde{c}\lambda + M\sqrt{\lambda} < x < R + \frac{\lambda p}{\mu_{cs}}, \end{cases} \quad (13)$$

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 $c_n, n = 0, \dots, M+1$. Then, $C^\lambda(\cdot)$ can be easily seen to satisfy Assumption 4.1. Indeed N_2^λ will necessarily be at a break point of the λ^{th} cost function and will thus have the required form.

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

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

It is imperative to note that Assumptions 4.1 and 4.2 are not used in any way in our proposed solution, as presented in §3. There, one does not have to know the constants \hat{W}, β, γ and c . Rather, it suffices to work directly with the cost function $C(\cdot)$ and the waiting time target \bar{W} . The scaling in Assumption 4.2 is consistent with many other models in which such square root approximations tend to perform extremely well (see for example [15] and [8]). We will show that this is the case also in our model.

We next 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

$$\limsup_{\lambda \rightarrow \infty} \frac{E[W^\lambda]}{\bar{W}^\lambda} \leq 1. \quad (14)$$

We now define our notion of asymptotic optimality. An intuitive notion of asymptotic optimality would stipulate that a sequence of pairs (N^λ, π^λ) is asymptotically optimal if it is asymptotically feasible and

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

for any other asymptotically feasible sequence $(\tilde{N}^\lambda, \tilde{\pi}^\lambda)$. This notion, however, is too weak as it implies that when the true 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 feasible routing rule 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. To that end, let

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

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

$$\liminf_{\lambda \rightarrow \infty} \frac{V^\lambda(N^\lambda, \pi^\lambda)}{V^\lambda(\tilde{N}^\lambda, \tilde{\pi}^\lambda)} \geq 1. \quad (15)$$

Before stating our main asymptotic optimality results we need a few more definitions. Re-define (4) as

$$N_1^\lambda = \arg \min \{N \in \mathbb{Z}_+ : E[W_{\lambda, \mu_s}^{FCFS}(N)] \leq \bar{W}^\lambda\}, \quad (16)$$

and define

$$\begin{aligned} \bar{N}_1^\lambda &= \inf \{N \in \mathbb{Z}_+, N \geq R : E[W_{\lambda, \mu_s}^{FCFS}(N) | W_{\lambda, \mu_s}^{FCFS}(N) > 0] \leq \bar{W}^\lambda\} \\ &= \inf \{N \in \mathbb{Z}_+, N \geq R : \frac{1}{N\mu_s - \lambda} \leq \bar{W}^\lambda\}, \end{aligned} \quad (17)$$

where the last equality follows from well known results for $M/M/N$ queues (see for example §5-9 of Wolff [21]). Also, we will say that two sequences $\{x^\lambda\}$ and $\{y^\lambda\}$ satisfy $x^\lambda \gg y^\lambda$ if $\frac{x^\lambda}{y^\lambda} \rightarrow \infty$ as $\lambda \rightarrow \infty$. 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 $\lambda \rightarrow \infty$, that satisfies Assumptions 4.1 and 4.2. Then, with $N_1^\lambda, \bar{N}_1^\lambda$ and N_2^λ defined in (16), (17) and (11) 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 δ in $[0, \hat{W})$ and $N^\lambda = N_2^\lambda$.
2. Condition 1 fails but $\liminf_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} \geq 1$, 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 $\limsup_{\lambda \rightarrow \infty} \frac{N_2^\lambda - R}{N_1^\lambda - R} < 1$, then it is asymptotically optimal to set $N^\lambda = N^{*\lambda}$ and $K^\lambda = K^{*\lambda}$, where $K^{*\lambda}$ and $N^{*\lambda}$ are determined by choosing the smallest value of $N^{*\lambda}$ that satisfies

$$N^{*\lambda} = \arg \max_{N \geq N_1^\lambda} r \mu_{cs} (E [Z_{\lambda, \mu_s}^{FCFS}(N) | Z_{\lambda, \mu_s}^{FCFS}(N) \geq (N + K^\lambda(N)) \wedge N] - R) - (C^\lambda(N) - C^\lambda(R)), \quad (18)$$

and

$$K^\lambda(N) = \max_{K \geq -N} \{E[Q_{\lambda, \mu_s}^{FCFS}(N) | Z_{\lambda, \mu_s}^{FCFS}(N) \geq (N + K) \wedge N] \leq \lambda \bar{W}^\lambda\}. \quad (19)$$

Condition 1 corresponds to the Revenue 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 \bar{N}_1^λ and the condition that N_2^λ is greater than N_1^λ .

Condition 3, which stipulates that the service and cross-selling rates are equal, allows us to show that for large enough systems, the dynamics of the call-center are approximately equal to those of a simpler model that is, in turn, easier to analyze. In this simple model the cross-selling phase is replaced by service to an infinitely long queue of low priority customers. This latter model was studied by Gans and Zhou [12]. We rely on their results in establishing the optimality of the TP policy and our proposed staffing for their model. The rigorous mathematical approximation then allows us to establish the asymptotic optimality results for our cross-selling model.

5. The MDP 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 (1) which applies in great generality, beyond the cases covered by Theorem 4.1. Specifically, following the approach in [12], 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 §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 λ and let the buffer size grow without bound. Accordingly, in this section, we fix λ and omit the superscript λ from all the notation.

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 π 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 (4). The following result applies with respect to $\Pi(N)$:

Lemma 5.1 *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 5.1 Work Conservation: *We say that a policy π 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 λ, μ_s, μ_{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 π . 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 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: 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:

$$\text{Max} \quad \sum_{j=0}^L \sum_{i=0}^{j \wedge N} r i \mu_{cs} (\xi(i, j, 0) + \xi(i, j, 1)) \tag{20}$$

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

(21)

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

(22)

and

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

(23)

The system of equations in (21) 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 (21) 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 (21) 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

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

(24)

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 λ is 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

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

Then, we have the following:

Proposition 5.1 *Assume $N \geq N_1$. Then,*

$$\lim_{L \rightarrow \infty} V^*_{LP}(N, L) = V^*(N). \quad (25)$$

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. 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. We fix $\mu_s = 1$ throughout but allow for different ratios of μ_{cs}/μ_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.

For each of the values of R , we vary the staffing levels. Then, for each of the staffing levels we search for the greatest threshold level that satisfies an average waiting time that is equal to $1/10$ of the average service time. Under TP, the process $S(t) := \{Z_2(t), Y(t)\}$ is a Continuous Time Markov Chain (CTMC) and we can find the cross-selling rate by solving the balance equations and calculating $\mu_{cs}(E[Z] - R)$.

For each of the staffing levels we also calculate the cross-selling rate obtained from the asymptotically optimal LP. We assume the same finite buffer for the asymptotically optimal LP and the steady-state calculations under TP. The values we use are 100 and 200 lines for $R = 30$ and $R = 100$ respectively. By Proposition 5.1, when the number of lines is large enough, the result of the asymptotically optimal LP provides a good approximation for the optimal values in the infinite buffer case. The differences when experimenting with larger number of lines are negligible.

We will show that TP performs extremely well, even in cases that are not covered by Theorem 4.1. The performance 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 [12]). 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 the same 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.

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,

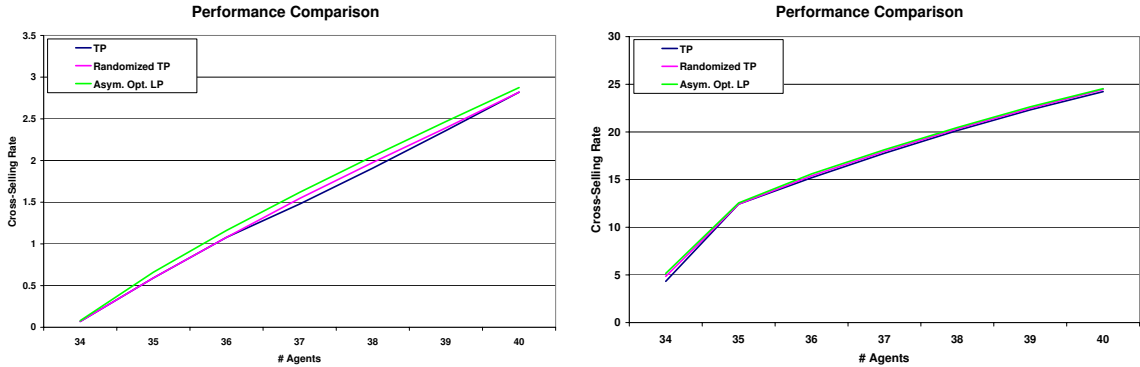


Figure 2: Cross Selling Rates Comparison for $R = 30$: (a) $\mu_{cs} = 1/3$ (b) $\mu_{cs} = 3$

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.

Staffing	106	110	114	118	122	126	130
Threshold	-20	-10	-6	-3	-1	1	2

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 remarkably good performance in great generality for large as well as moderate size call centers and beyond the scope covered by our sufficient conditions of Theorem 4.1.

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 under the following piecewise linear staffing cost function: Each agent up to

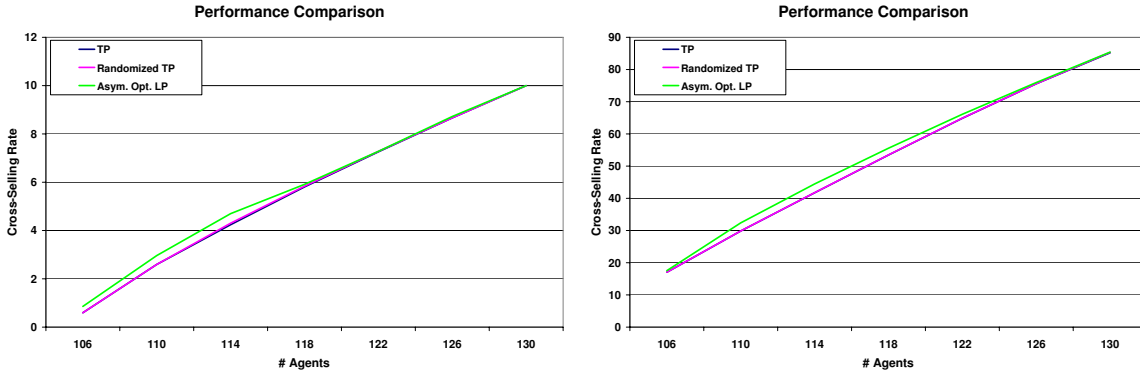


Figure 3: Cross Selling Rates Comparison for $R = 100$: (a) $\mu_{cs} = 1/3$ (b) $\mu_{cs} = 3$

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 resulting 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 (3)) 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 rules 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.

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. 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 the managers to easily incorporate this data in addition to pre-specified service-level targets into the staffing and control decision.

This more comprehensive approach towards 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 very 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 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. A follow-up paper [14] 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 studied 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 [17], Borst et. al. [8], Armony [4], and Gurvich et. al. [15], 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 (5)) to cross-sell to a fraction $f^* > 0$ of the customers, a staffing level of the form

$$N = R + \frac{f^*\lambda}{\mu_{cs}} = R + f^* \frac{\mu_s}{\mu_{cs}} R \quad (26)$$

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, $R' := R + \frac{f^*\lambda}{\mu_{cs}}$, 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 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 revenue 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 $\frac{1}{\mu_s} + \frac{f^*}{\mu_{cs}}$. In particular, the estimate of the offered load will be $R' = \frac{\lambda}{\mu_s} + \frac{\lambda f^*}{\mu_{cs}}$, 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 $R + f\lambda/\mu_{cs}$ 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 $\frac{1}{\mu_s} + \frac{f}{\mu_{cs}} + o(1)$. 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

- [1] Akşin O.Z., Harker P.T., “To sell or not to sell: Determining the trade-offs between service and sales in retail banking phone centers”. *Journal of Service Research*, 2(1), pp. 19-33, 1999.
- [2] Altman E., “Constrained Markov Decision Processes”, Chapman & Hall/CRC, London, 1999.
- [3] Armony M., Mandelbaum A., “Routing and staffing in large-scaled service systems with heterogeneous servers and impatient customers”, Working Paper, NYU, New York, NY, 2006.
- [4] Armony M., “Dynamic routing in large-scale service systems with heterogeneous servers”, *Queueing Systems*, 51(3-4), pp. 287-329, 2005.
- [5] Armony M., Maglaras C., “On customer contact centers with a call-back option: customer decisions, routing rules and system design”, *Operations Research*, 52(2), pp. 271-292, 2004.
- [6] Armony M., Maglaras C., “Contact centers with a call-back option and real-time delay information”, *Operations Research*, 52(4), pp. 527-545, 2004.
- [7] Aydin G. and Ziya S., “Upselling a promotional product using customer purchase information”, *M&SOM*, forthcoming, 2008.
- [8] Borst S., Mandelbaum A. and Reiman M., “Dimensioning Large Call Centers”, *Operations Research*, 52(1), pp. 17-34, 2004.

- [9] Byers R.E. and So K.C., “The value of information-based cross-sales policies in telephone service centers”, Working Paper, Graduate School of Management, University of California, Irvine, CA, 2004.
- [10] Byers R.E. and So K.C., “A mathematical model for evaluating cross-sales policies in telephone service centers”, *Manufacturing & Service Operations Management*, 9(1), 1-8. 2007.
- [11] Eichfeld A., Morse T.D. and Scott K.W., “Using call centers to boost revenue”, McKinsey Quarterly, May 2006.
- [12] Gans N. and Zhou Y.-P., “A call-routing problem with service-level constraints”, *Operations Research*, 51(2), pp. 255-271, 2003.
- [13] Gunes E.D. and Akşin O.Z., “Value creation in service delivery: Relating market segmentation, incentives, and operational performance”, *Manufacturing & Service Operations Management*, 6(4), pp. 338-357, 2004.
- [14] Gurvich I., Armony M. and Maglaras C., “Cross-Selling in call centers with a heterogeneous population”, *Operations Research*, forthcoming, 2008.
- [15] Gurvich I., Armony M. and Mandelbaum A., “Service level differentiation in call centers with fully flexible servers”, *Management Science*, 54(2), pp. 279-294, 2008.
- [16] Gurvich I. and Zeevi A., “Validity of heavy-traffic steady-state approximations in open queueing networks: Sufficient conditions involving state-space collapse”, Working Paper, Columbia University, New York, NY, 2007.
- [17] Halfin S. and Whitt W., “Heavy-traffic limits for queues with many exponential servers”, *Operations Research*, 29, pp. 567-587, 1981.
- [18] Mandelbaum A. and Zeltyn S., “Staffing many-server queues with impatient customers: constraint satisfaction in call centers”, Working Paper, Technion - Israel Institute of Technology, Haifa, Israel 2007.
- [19] Netessine S., Savin S. and Xiao W., “Dynamic revenue management through cross-selling in E-commerce retailing”, *Operations Research*, 54(5), pp. 893-913, 2006.
- [20] Örmeci E.L. and Akşin O.Z., “Revenue management through dynamic cross-selling in call centers”, Working Paper, Koç University, Istanbul, Turkey, 2006.
- [21] Wolff R.W., “Stochastic Modeling and the Theory of Queues”, Prentice Hall, Englewood Cliffs, NJ, 1989.