Customer Acquisition, Retention, and Queueing-Related Service Quality:
Optimal Advertising, Staffing, and Priorities for a Call Center

Philipp Afèche
Rotman School of Management, University of Toronto

Mojtaba Araghi
School of Business and Economics, Wilfrid Laurier University

Opher Baron
Rotman School of Management, University of Toronto

June 10 2015

We study the problem of maximizing profits for an inbound call center with abandonment by controlling customer acquisition, retention, and service quality via advertising, priorities, and staffing. This paper makes four contributions. First, we develop what seems to be the first marketing-operations model of a call center that captures the evolution of the customer base as a function of past demand and queueing-related service quality. Second, we derive metrics that drive the optimal decisions, the expected customer lifetime value of a base (i.e., repeat) customer and the expected one-time serving value of a new and base customer. These metrics link customer and financial parameters with operational service quality, reflecting the system load and the priority policy. Third, using a deterministic fluid approximation we provide new analytical prescriptions for the optimal advertising, staffing, and priority policies, as functions of these customer value metrics and the advertising and capacity costs. Fourth, we show via simulation that the fluid model prescriptions yield near-optimal performance for the stochastic model. This tractable modeling framework can be extended to further problems of joint customer relationship and call center management.

Key words: Abandonment; advertising; call centers; congestion; customer relationship management; fluid models; marketing-operations interface; promotions; priorities; service quality; staffing; queueing systems.

1 Introduction

Call centers are an integral part of many businesses. By some estimates, firms that use call centers conduct 80% of interactions with its customers through this channel, and 92% of customers base their opinion of a company on their call center service experiences (Anton et al. 2004). Moreover, these experiences can have a dramatic impact on customer satisfaction and retention. Long waiting times are cited by 67% of customers as a major cause of frustration, and a poor call center experience is cited by 40% of customers as the sole reason for terminating their relationship with a business
(Genesys Global Consumer Survey 2007). These findings underscore the key premise of customer relationship management (CRM), which is to view a firm’s interactions with customers as part of ongoing relationships, rather than in isolation. As Akşin et al. (2007, p. 682) point out, “firms would benefit from a better understanding of the relationship between customers’ service experiences and their repeat purchase behavior, loyalty to the firm, and overall demand growth in order to make better decisions about call center operations.” The importance of understanding this relationship has recently also been recognized in the marketing literature (cf. Sun and Li 2011).

This paper provides a starting point for building such understanding. Motivated by our conversations with managers of a credit card company, we develop an integrated marketing-operations model to study the problem of maximizing the profit of an inbound call center that serves new and existing (base) customers whose acquisition and retention depend on the queueing-related service quality. We address the following questions: What is the value of a call and the value of a customer? How do these metrics depend on the call center’s queueing-related service quality? How to optimally control customer acquisition, retention, and service quality via advertising, staffing and priorities? How should the optimal decisions account for the effects of service quality on customer value? Finally, how do traditional customer value metrics that ignore the service quality lead to suboptimal decisions? This paper makes four contributions.

1. **Modeling.** We develop a new model that links CRM with call center operations (§3). We consider a multi-server queueing system, new customer arrivals in response to advertising, their conversion to base customers of different types, the evolution and calls of the customer base, and call-related and call-independent profits and costs. Customers are impatient, and their abandonment adversely affects their acquisition and retention. This seems to be the first paper to model the impact of past demand and queueing-related service quality, that is, the probability of getting served, on the evolution and value of the customer base. In contrast, the customer base is independent of queueing in CRM models, and independent of past interactions in call center models.

2. **Call and Customer Value Metrics.** We derive new metrics that drive the optimal call center decisions (§4): The expected ***customer lifetime value*** (CLV) of base customers, the expected ***one-time serving value*** (OTV) of new and base customers, and the maximum new customer value per processing time. Unlike standard CLV metrics in the marketing literature, these metrics depend on operations through the service probabilities, and in turn on the system load and priority policy.

3. **Guidelines for Call Center Management.** Based on a deterministic fluid approximation of the underlying stochastic model, we provide new analytical prescriptions for the profit-maximizing advertising, staffing, and priority policies (§5). In contrast to standard call center models, in ours the optimal staffing and priority policies account for the effect of service levels on customers’ ***future*** calls and their financial impact, not only for the abandonment and/or waiting costs of
current calls. Specifically, we show that under mild conditions it is optimal to over\textit{load} the system, whereby the firm serves only new and lucrative base customers, but not unprofitable base customers. We illustrate how our results yield such a regime depending on customer attributes such as their call-independent profit rate and loyalty (§6.1). This result highlights the key role of the priority policy as a tool for controlling the customer base composition via differentiated retention efforts. This policy can be viewed as the service counterpart of targeted advertising, which controls the customer base via differentiated acquisition efforts (§6.2). Furthermore, we show that if decisions are based on marketing-focused policies that ignore the effect of service probability on the CLV, the advertising level is lower than optimal (§6.3). Finally, we discuss the effects of word of mouth about service quality on our results (§6.4). These findings not only confirm the assertion of Akşin et al. (2007) but also show specifically how firms can improve call center operations by accounting for the relationships between customers’ service experiences and their future behavior.

4. Fluid Model Accuracy. We show via simulation that the fluid model prescriptions yield near-optimal performance for the stochastic model (§7). These results suggest that the fluid model guidelines also apply to the stochastic model, and more generally, that our fluid model approach may prove effective in tackling further problems of joint CRM and call center management.

2 Literature Review

This paper bridges research streams on advertising, CRM, and call center management. It also relates to operations papers outside the call center context that link demand to past service.

Advertising. There is a vast literature on dynamic advertising policies. Feichtinger et al. (1994) offer an extensive review. In contrast to our study, the overwhelming majority of these papers ignore supply constraints. Some papers do consider supply constraints but study different settings from ours. Focusing on physical goods, Sethi and Zhang (1995) study advertising and production control, and Olsen and Parker (2008) study advertising and inventory control. Focusing on services, Horstmann and Moorthy (2003) study the relationship between advertising, capacity, and quality under competition, but in their model, unlike in ours, quality is independent of utilization.

CRM. Models of CRM and CLV are of central concern in marketing. For reviews see Rust and Chung (2006) and Reinartz and Venkatesan (2008) on CRM, and Gupta et al. (2006) on CLV. Studies that propose and empirically demonstrate the value of CLV-based frameworks for customer selection and marketing resource allocation include Rust et al. (2004) and Venkatesan and Kumar (2004). Based on the CLV components, CRM initiatives can be broadly classified as focusing on customer acquisition, growth, and/or retention. Studies that focus on the relationship between acquisition and retention spending include Blattberg and Deighton (1996), Reinartz et al. (2005),
Musalem and Joshi (2009), Pfeifer and Ovchinnikov (2011) and Ovchinnikov et al. (2014). Studies
that focus on the effects of marketing actions on customer growth include Li et al. (2005), Rust and
Verhoef (2005), and Günes et al. (2010). Papers that focus on explaining or predicting customer
retention and churn include Verhoef (2003), Braun and Schweidel (2011), and Ascarza and Hardie
(2013). Papers that study the links among service quality, customer satisfaction, retention and
other CLV components include Anderson and Sullivan (1993), Rust et al. (1995), Zeithaml et al.

The CRM literature mostly focuses on modeling and estimation. In contrast, only a few pa-
pers study optimal decisions toward maximizing CLV, some for static policies (e.g., Blattberg and
Deighton 1996, Rust and Verhoef 2005, Ho et al. 2006) and others for dynamic policies (e.g., Bitran
and Mondschein 1996, Lewis 2005, Musalem and Joshi 2009, Günes et al. 2010, Sun and Li 2011,

Compared to the CRM literature, the key distinction of our model is that we explicitly link
customer acquisition and retention to the queueing-related service quality. Other studies that opti-
mize some notion of service quality (Ho et al. 2006 and Aflaki and Popescu 2014) ignore capacity
constraints (and customer acquisition). The two papers that consider a capacity constraint (to our
knowledge, only Pfeifer and Ovchinnikov 2011 and Ovchinnikov et al. 2014) ignore the effects of
queueing and service quality on customer acquisition, retention, and CLV.

Call center management. The vast call center literature focuses on operational decisions. See
Gans et al. (2003), Akşin et al. (2007), and Green et al. (2007) for surveys. Numerous papers
also consider marketing controls, in particular, cross-selling (cf. Akşin et al. 2007) and pricing
(typically for single-server queueing models, cf. Hassin and Haviv 2003 for a review). Traditionally,
these research streams study only queueing-related service quality metrics, i.e., waiting time and
abandonment, taking the quality delivered by the server(s) to be fixed and perfect. However, there
is growing interest in also considering service delivery failures (e.g., de Véricourt and Zhou 2005)
or trade-offs between the quality of service delivery and waiting time (e.g., Anand et al. 2011).

In contrast to this paper, these call center and queueing research streams model a firm’s customer
base as independent of the service quality of past interactions. That is, they take the arrival
process of customer requests as exogenous. (Models with balking take arrivals of potential requests
as exogenous, e.g., Anand et al. 2011. Models with retrials take arrivals of initial requests as
exogenous, e.g., de Véricourt and Zhou 2005, Randhawa and Kumar 2008.) Because these models
do not keep track of repeat customers, their service level prescriptions reflect only transaction-based
metrics such as waiting and abandonment costs. In contrast, the prescriptions in our model also
consider the effect of service levels on customers’ future calls and their CLV.
In a parallel effort, Farzan et al. (2012) consider a FIFO queue with repeat purchases that depend on past service quality; however, in contrast to our model, they take service quality as independent of queueing. In the marketing literature, Sun and Li (2011) empirically estimate how retention depends on customers’ allocation to on- vs. off-shore call centers, including on waiting and service time. Their results underscore the value of linking these service quality metrics to retention, CLV, and they point out the need to connect these metrics to queueing models.

Operations literature on the link between demand and past service. A number of papers consider the link between demand and past service for operations outside the call center context. Their models are fundamentally different from ours. Schwartz (1966) appears to be the first to consider how demand depends on past inventory availability. Gans (2002) and Bitran et al. (2008b) consider a general notion of service quality in the absence of capacity constraints, the former for oligopoly suppliers, the latter for a monopoly. Hall and Porteus (2000), Liu et al. (2007), Gaur and Park (2007), and Olsen and Parker (2008) study equilibrium capacity/inventory control strategies and market shares of firms that compete for customers who switch among them in reaction to poor service. The work of Olsen and Parker (2008) is distinct in that it considers nonperishable inventory, consumer backlogs, and firms that control not only inventory, but also advertising to acquire new/reacquire dissatisfied customers. Adelman and Mersereau (2013) study how a supplier of a physical good should dynamically allocate its limited capacity among a fixed portfolio of heterogeneous customers whose stochastic demands depend on their past fill rates.

3 Model and Problem Formulation

The credit card company that motivated this research uses call centers as the main customer contact channel. The firm also uses other channels such as the web, e-mail, and SMS, but these are much less costly and not suited for requests that require interaction with customer service representatives.

The focus of our model, one of the firm’s call centers, is both sales- and service-oriented. Potential new customers call in response to advertised credit card offers. Existing card holders call with service requests, e.g., to increase their credit limit, report a lost or stolen card, and so on.

Marketing and operations decisions are made each month according to the following procedure.

The marketing department sets the monthly advertising spending level and the volume of credit card offers to be mailed out, based on some measures of customer profitability. However, these decisions ignore operational factors, specifically, service level considerations and staffing cost. Moreover, the operations department is not involved in advertising decisions. Once the monthly advertising level is determined, credit card offers are mailed evenly over the course of that month. Offers have expiration dates and potential customers may call during this period to apply for a credit card.
Following the advertising budget decisions, the operations department forecasts the monthly call volume. The forecast of the new customer call volume considers the number of credit card offers to be mailed out as well as other advertising activities that increase brand awareness, and historical data on the demand response to these factors. The forecast of the base customer call volume considers the number of card holders and historical data on their calling pattern.

Based on the forecast of monthly call volume, the operations department determines the staffing level for that month to meet certain predetermined service level targets.

This decision procedure raises the following issues: 1. There is no clarity about the value of a call, the value of a customer, and how these measures depend on the call center’s service quality. 2. Advertising and staffing decisions either ignore service level considerations or make arbitrary assumptions about service level targets. 3. It is not clear whether to prioritize certain customers, and if so, on what basis. 4. Marketing and operations do not adequately coordinate their decisions.

These issues are pertinent not only for this particular company, but also for other firms such as insurance providers that use call centers as the main customer contact channel. To address these issues we develop and analyze an integrated marketing-operations model.

3.1 The Stochastic Queueing Model

Consider a firm that serves new and base (i.e., existing) customers through its inbound call center, modeled as a multi-server system. New customers are first-time callers and may turn into any of $m$ base customer types. Base customers make up the firm’s customer base and repeatedly interact with the call center. We index new customers by $i = 0$, base customers by $i \in \{1, 2, ..., m\}$, say type-$i$ customers when $i \in \{0, 1, 2, ..., m\}$, and type-$i$ base customers when $i \in \{1, 2, ..., m\}$. Calls arrive as detailed below. Customers queue if the system is busy upon arrival, but they are impatient. Type-$i$ customers' abandonment times are independent and identically distributed (i.i.d.) with mean $1/\tau_i$. Their service times are i.i.d. with mean $1/\mu_i$. Table 1 summarizes the notation.

We consider the system in steady-state under three stationary controls. The staffing policy sets the number of servers $N$ at a cost of $C$ per server per unit time. The advertising policy controls the new customer call arrival rate $\lambda_0$ as discussed below. The priority policy controls the service order of new and base customer calls. Let $q_i$ denote the steady-state service probability of type-$i$ customer calls. The service probabilities depend on the system parameters and controls as discussed below.

The following customer flows determine the evolution of the customer base. Figure 1 shows the flow of new (type-0) customers by dashed lines and that of type-$i$ base customers by solid lines.

New customer calls arrive to the system following a stationary Poisson process with rate $\lambda_0$. The new customer call arrival rate depends on the firm’s advertising spending. Let $S(\lambda_0)$ denote the advertising spending rate per unit time as a function of the new customer arrival rate it
generates. We assume that the response of new customers to advertising spending follows the law of diminishing returns (Simon and Arndt 1980), so \( S(\lambda_0) \) is strictly increasing and strictly convex in \( \lambda_0 \). For analytical convenience we assume that \( S \) is twice continuously differentiable and \( S'(0) = 0 \). (In §6.4 we generalize this model so that new customers’ response to advertising depends on word of mouth from base customers about their service quality.)

A new customer who receives service joins the customer base as a type-i base customer with probability \( \theta_{0i} > 0 \), so new customers turn into base customers at an average rate of \( \lambda_0 \theta_{0i} \) per unit time. The times between successive calls of a type-i base customer are independent and exponentially distributed with mean \( 1/r_i \), so \( r_i \) is the call arrival rate per type-i base customer. Base customers may leave the company for call-related or call-independent reasons. A type-i base customer who abandons the queue remains in the customer base with probability \( \theta_i \), but immediately terminates her relationship with the company and leaves the customer base with probability \( 1 - \theta_i \). The customer base is also subject to attrition due to call-independent reasons, such as relocation, switching to a competitor that offers a preferred product, or death. The lifetimes of type-i base customers in the absence of abandonment are i.i.d. with mean \( 1/\gamma_i \), so \( \gamma_i > 0 \) is their average call-independent departure rate. We assume that the mean time between calls from any base customer and her mean sojourn time in the customer base are much larger than the mean service and abandonment times. That is, \( 1/r_i, 1/\gamma_i >> 1/\mu_0, 1/\mu_i, 1/\tau_0, 1/\tau_i \) for \( i \in \{1, 2, \ldots, m\} \).

Let \( x_i \) denote the average number of type-i base customers in steady-state, or simply the size of the type-i customer base. Therefore, in steady-state type-i base customers’ average call arrival rate is \( x_i r_i \), and their average call-independent departure rate is \( x_i \gamma_i \). The system is stable because customers are impatient, and the size of the customer base is finite since \( \lambda_0 < \infty \) and \( \gamma_i > 0 \).

The profit of the firm has call-dependent and call-independent components. On average the firm generates a profit of \( p_i \) per type-i customer call it serves (with a purchase or service request)
System Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Number of servers</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Service rate of type-$i$ customers</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>Call arrival rate of new customers</td>
</tr>
<tr>
<td>$r_i$</td>
<td>Call arrival rate per type-$i$ base customer</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>Call abandonment rate per type-$i$ customer</td>
</tr>
<tr>
<td>$\theta_{ni}$</td>
<td>Probability that a new customer joins as a type-$i$ base customer after service</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Probability that a type-$i$ base customer remains in the customer base after abandoning</td>
</tr>
<tr>
<td>$\gamma_i$</td>
<td>Call-independent departure rate per type-$i$ base customer</td>
</tr>
</tbody>
</table>

Economic Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i$</td>
<td>Profit per served call of type-$i$ customers</td>
</tr>
<tr>
<td>$c_i$</td>
<td>Cost per abandoned call of type-$i$ customers</td>
</tr>
<tr>
<td>$R_i$</td>
<td>Call-independent profit rate per type-$i$ base customer</td>
</tr>
<tr>
<td>$C$</td>
<td>Cost rate per server</td>
</tr>
</tbody>
</table>

Steady-State Performance Measures

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>Average number of type-$i$ base customers</td>
</tr>
<tr>
<td>$q_i$</td>
<td>Service probability of type-$i$ customer calls</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>Profit rate</td>
</tr>
</tbody>
</table>

Table 1: Summary of Notation.

and incurs a cost of $c_i \geq 0$ per type-$i$ customer call that abandons. In addition, the firm may generate a call-independent profit from type-$i$ base customers at an average rate of $R_i \geq 0$ per unit time. For example, in the case of the credit card company, this call-independent profit represents subscription fees and interest payments from card holders and transaction fees from merchants.

Let $\Pi$ denote the firm’s average profit rate in steady-state. It is given by

$$\Pi := \lambda_0 (p_0 q_0 - c_0 (1 - q_0)) + \sum_{i=1}^{m} x_i (R_i + r_i (p_i q_i - c_i (1 - q_i))) - C N - S (\lambda_0).$$

The first product is the new customer profit rate, the sum is over the base customer profit rates (both call-independent and call-dependent), the third term is the staffing cost rate, and the last is the advertising cost rate. The firm maximizes its profit by choosing the new customer arrival rate $\lambda_0$, the number of servers $N$, and the priority policy which controls the service probabilities $q_i$.

The profit rate (1) depends on two sets of stationary performance measures, the service probabilities $q_i$, and the customer base sizes $x_i$. The state-dependent customer flows and feedbacks through the system make it difficult to analyze these measures for the stochastic model, even under Markovian assumptions. We therefore approximate the stochastic model by a deterministic fluid model that we describe in §3.3. Before doing so we discuss the model assumptions in more detail.

3.2 Discussion of Model Assumptions

We discuss and validate our assumptions, focusing on the credit card company that motivated this work. We note that our model can be tailored to a range of businesses, such as insurance, mobile...
phone or catalog marketing companies, by appropriately choosing parameter values.

3.2.1 Service Quality

Call Center as the Bottleneck for Quality. Our model focuses on the call center as the bottleneck for quality, because for the credit card company, the call center is both the costliest customer contact channel and the most important one in terms of customer acquisition and retention. Therefore, the company’s key trade-off between the value and cost of service quality focuses on call center operations. In contrast, no such trade-offs arise for other attributes of the company’s product and service quality, such as the features of its credit products or the service quality that is determined by backoffice operations (network reliability, invoice accuracy, and so on). These attributes are either independent of the size of its customer base, easy to scale, or subject to clear standards. Indeed, the credit card company keeps these quality attributes relatively fixed, as discussed below.

Focus on Waiting-time-dependent Service Quality. Our model focuses on controlling the service probabilities, which are waiting-time-dependent measures of service quality. However, we assume that the call resolution quality is fixed and independent of congestion, as is common in the literature. Furthermore, for the call center of this credit card company, as for many financial institutions, it was clear that the quality of each service encounter must meet strict standards. It was less clear what waiting times are appropriate, supporting our focus on the waiting-time-dependent service quality. Nevertheless, our model is rich enough to accommodate optimization over call resolution quality. The call resolution quality for new customers is captured by the parameter $\theta_{0i}$, and similar parameters can be introduced for base customers to capture their defection probability after service.

Assumptions on Waiting-time-dependent Service Quality. We assume that (i) a customer’s transitions into and out of the customer base depend on waiting times only through abandonment, and (ii) a base customer’s reaction to the waiting time of her current call is independent of her service history. Therefore, customer flows depend on waiting times only through the service probabilities $q_i$. Assumption (i) is consistent with the notions of “critical incidents” and “end effects” in the literature. The prevalent assumption in the service literature is that customers think about terminating their relationship with providers only when some critical incident occurs (Keaveney 1995, Gremler 2004). In our setting the abandonment is the natural critical incident – this is how customers signal that their waiting times are too long. The discussion in Bitran et al. (2008a) (see section 4.2.1 and references therein) suggests the presence of an “end effect”, whereby the outcome of a service encounter may significantly affect the memory of the preceding waiting experience. In our setting this means that a customer’s future behavior (captured by $\theta_{0i}$ for new customers, and by $R_i$, $r_i$, $\theta_i$ and $\gamma_i$ for base customers) only depends on whether she got served or not, but conditional
on this outcome, is independent of her waiting time. Assumption (ii) suggests that customers exhibit a “recency effect”, that is, their decision to remain in or leave the customer base only depends on the outcome of their current service encounter. This assumption is common in models that link demand to past service levels (e.g., Hall and Porteus 2000, Ho et al. 2006, Liu et al. 2007). Furthermore, although this assumption may not perfectly represent individual customers’ defection behavior, it does capture the aggregate effect of abandonment on defection. The parameter $\theta_i$ can be interpreted as a “loyalty coefficient” (cf. Reichheld 1996, Hall and Porteus 2000) that measures the fraction of service failures, i.e., abandonments, from which the firm gracefully recovers.

**Fixed Waiting-time-independent Quality.** We model the parameters $\theta_0$, $\theta_i$, $r_i$ and $\gamma_i$ as fixed. They can be viewed as aggregate measures of how customers evaluate the call resolution quality, the firm and the quality of its products and services in relation to competitors. Customers’ aggregate reaction to service or abandonment, their calling frequency and call-independent defection behavior are summarized by these parameters. These parameters can be, and indeed are, used to segment base customers into different types. In the credit card company the call center manager had little effect on these parameters and therefore treated them as constant. Our model can accommodate optimization over these parameters, for example, to study the effect of training or product improvement on customer satisfaction and loyalty, but these decisions are outside the scope of this paper. We do consider the sensitivity of our results to some of these parameters in §6.1 and §7.3.

### 3.2.2 Service Capacity

**Focus on a Single Channel.** We assume a single channel, because as noted above, for the credit card company the call center is both the costliest and the most important channel. However, one can also view our model as capturing a premium high-cost channel explicitly and a self-service low-cost channel such as the Web implicitly and approximately: According to the Yankee Group (2006) the average cost of serving a customer through the Web is an order of magnitude cheaper than serving her via call centers. Our model can be extended to multiple channels by considering multiple capacity pools with different costs, service rates, interaction rates, and so on.

**Pooled versus Dedicated Servers.** Our model assumes that servers are flexible, that is, the firm can pool capacity across different queues. Capacity pooling is common in practice, including in the credit card company discussed above. Another practice is to dedicate capacity to specific types of calls. The fluid models for these different capacity allocation models yield the same results.

### 3.2.3 Service Demand

**Stationary Demand.** We assume that the new customer arrival rate is a concave, stationary, and deterministic function of the advertising spending rate. Under these assumptions the optimal
advertising policy is to spend continuously at an even rate (Sasieni 1971, Feinberg 2001). Such a policy yields a constant demand rate in that it eliminates the demand response decays that arise between advertising pulses under pulsing policies. Indeed, although the credit card company sets the advertising level at the beginning of each month, it mails out the offers gradually over the month. This steady advertising pattern yields a steady demand response. Our model ignores the typical predictable short-term arrival rate fluctuations over the course of the day and week, in order to focus on the first-order relationship between the stationary average advertising spending and demand response levels. This focus is in line with the main objective of this paper, which is to derive strategic-level guidelines for optimal advertising, staffing, and priority policies based on customer value metrics that account for the queueing-related service quality. Our stationary model captures the essential effects of these decisions on profitability. Nevertheless, for call centers where staffing levels can be flexibly adjusted over time (this holds for the credit card company, and for a growing number of other companies, as a result of flexible workforce contracts and outsourcing providers), our results can also be used at the tactical level to adjust capacity to both predictable and unpredictable short-term arrival rate fluctuations, so as to maintain the optimal service levels.

Independence of Base Customer Demand from Advertising. We assume that card holders’ calling rate $r_i$ is fixed. This reflects the fact that advertisements and promotions to base customers aim to increase their credit card usage (e.g., by introducing reward points), not their calling rate. For simplicity and to focus on the effect of service quality on retention, we do not model promotion decisions to base customers. However, our model implicitly captures the costs and benefits of such promotions through the call-independent profit $R_i$, the per-call-profit $p_i$, the new customer conversion probability $\theta_{0i}$, the loyalty coefficient $\theta_i$, and the call-independent defection rate $\gamma_i$.

3.3 The Approximating Fluid Model

In this section we formulate the fluid approximation and the resulting profit-maximization problem. In steady-state, the size of the customer base of type $i$ must be constant in time:

$$\lambda_0 q_0 \theta_{0i} = x_i (\gamma_i + r_i (1 - q_i) (1 - \theta_i)),$$

where the left-hand side is the rate at which new customers generate type-$i$ base customers, and the right-hand side is the type-$i$ customer base decay rate which is proportional to its size. As discussed in §3.1, a type-$i$ base customer’s departure rate has two components, the call-independent rate $\gamma_i$ and the call-dependent rate, which equals the product of her calling rate $r_i$, abandonment probability $1 - q_i$, and probability of leaving the customer base after abandonment $1 - \theta_i$.

From (2), the mean number of type-$i$ base customers in steady state is

$$x_i = \frac{\lambda_0 q_0 \theta_{0i}}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)}.$$
For the system to operate, the new customer service probability must be positive, i.e., \( q_0 > 0 \).

Let \( q = (q_0, q_1, ..., q_m) \), where \( q_i \) denotes the steady-state service probability for calls of type \( i \).

From (1) and (3) the profit-maximization problem simplifies to the following nonlinear program:

\[
\max_{q \geq 0, \lambda_0 \geq 0} \Pi = \lambda_0 \left( p_0 q_0 - c_0 (1 - q_0) \right) + \sum_{i=1}^{m} x_i \left( R_i + r_i [p_i q_i - c_i (1 - q_i)] \right) - CN - S (\lambda_0)
\]

subject to

\[
x_i = \frac{\lambda_0 q_0 \theta_{0i}}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)}, \quad i = 1, 2, ..., m,
\]

\[
q_i \leq 1, \quad i = 0, 1, 2, ..., m,
\]

\[
\frac{\lambda_0 q_0}{\mu_0} + \sum_{i=1}^{m} \frac{x_i r_i q_i}{\mu_i} \leq N.
\]

The left-hand side of the capacity constraint (6) expresses the total processing time allocated to (and consumed by) new and base customer calls.

One obvious caveat of the deterministic fluid model is that it does not account for queueing effects and customer impatience in evaluating the service probabilities. However, the great advantage of the fluid model is its analytical tractability. It yields clear results on the optimal decisions, as shown in §5 and §6. Moreover, our simulation results in §7 show that the optimal decisions that the fluid model prescribes yield near-optimal performance for the stochastic system it approximates.

## 4 Customer Value and Queueing-Related Service Quality

In §4.1 we derive customer value metrics that are novel in that they depend on the queueing-related service quality. In §4.2 we reformulate the problem (4)-(6) in terms of these metrics. In §4.3 we identify priority-dependent measures of new customer lifetime value that drive the optimal policy.

### 4.1 Customer Value Metrics

Let \( L_i (q_i) \) denote the mean type-\( i \) base customer lifetime value (CLV), i.e., her total profit during her sojourn time in the customer base, as a function of the service probability \( q_i \):

\[
L_i (q_i) := \frac{R_i + r_i (p_i q_i - c_i (1 - q_i))}{\gamma_i + r_i (1 - q_i) (1 - \theta_i)}.
\]

The CLV of a type-\( i \) base customer is the product of her profit rate per unit time, the numerator in (7), by her average sojourn time in the customer base, the reciprocal of the denominator in (7).

Let \( V_0 \) denote the mean one-time service value (OTV) of a new customer, i.e., the value of serving a new customer’s current call, but not any of her future calls:

\[
V_0 := p_0 + c_0 + \sum_{i=1}^{m} \theta_{0i} L_i (0).
\]
Serving a new customer yields an instant profit of $p_0$, plus a call-independent but type-dependent profit stream, where with probability $\theta_{0i}$ the new customer turns into a type-$i$ base customer with CLV $L_i(0)$ (given a zero service probability). Not serving a new customer yields a loss of $c_0$.

Similarly, let $V_i$ denote the mean OTV of a type-$i$ base customer:

$$V_i := p_i + c_i + (1 - \theta_i) L_i(0), \quad i = 1, 2, \ldots, m.$$  \hspace{1cm} (9)

Serving a type-$i$ base customer's current call, but not any of her future calls, yields the profit $p_i + L_i(0)$, not serving her current call yields $-c_i + \theta_i L_i(0)$, so the difference yields (9).

From (7) and (9), the base customer OTV and CLV satisfy the following intuitive relationship:

$$V_i = \frac{L_i(1) - L_i(0)}{r_i/\gamma_i}, \quad i = 1, 2, \ldots, m,$$  \hspace{1cm} (10)

where $r_i/\gamma_i$ is the mean number of calls during a type-$i$ base customer's lifetime if all her calls are served. For $i \geq 1$ we assume $V_i \mu_i \geq V_{i+1} \mu_{i+1}$ without loss of generality and $V_i \mu_i > V_{i+1} \mu_{i+1}$ for simplicity (cases with $V_i \mu_i = V_{i+1} \mu_{i+1}$ add cumbersome detail but no insight to the analysis).

### 4.2 The Profit-Maximization Problem in Terms of Customer Value Metrics

We reformulate the problem (4)-(6) in terms of the OTVs (8)-(9), by formalizing the priority policy in terms of the capacity allocation to customer types, instead of their service probabilities $q_i$.

Let $N_i$ denote the capacity allocated to (and consumed by) type-$i$ customers. We have

$$N_0 := \frac{\lambda_0 q_0}{\mu_0},$$  \hspace{1cm} (11)

$$N_i := \frac{x_i r_i q_i}{\mu_i},$$  \hspace{1cm} (12)

Let $N := (N_0, N_1, N_2, \ldots, N_m)$. The numerators of (11) and (12) represent, respectively, the throughput rates of new and type-$i$ base customer calls as functions of their service probabilities, and the denominators are their service rates. It follows from (3), (11), and (12) that

$$x_i = \frac{N_0 \mu_0 \theta_{0i} + N_i \mu_i (1 - \theta_i)}{\gamma_i + r_i (1 - \theta_i)}.$$  \hspace{1cm} (13)

Write $s_0$ for the mean service time of a new customer call. Then

$$s_0 := \frac{1}{\mu_0}$$  \hspace{1cm} (14)

and $\lambda_0 s_0$ is the offered load of new customer calls. Let $s_i$ denote the expected total service time of all type-$i$ base customer calls that may be generated as a result of serving a new customer. Then

$$s_i := \theta_{0i} \frac{r_i}{\gamma_i \mu_i} + \frac{1}{\mu_i},$$  \hspace{1cm} (15)

where $\theta_{0i}$ is the probability that a new customer who is served joins as a type-$i$ base customer, $r_i/\gamma_i$ is the mean number of calls during a type-$i$ base customer’s lifetime if all her calls are served,
and $1/\mu_i$ is the mean service time of a type-$i$ base customer call. That is, $s_i$ is the offered load of type-$i$ base customer calls that a new customer generates, and $\lambda_0s_i$ is their total offered load.

Let $\Pi (N, N, \lambda_0)$ denote the profit rate as a function of the capacity allocation vector $N$, the total capacity $N$, and the new customer arrival rate $\lambda_0$. By using (7)-(9) and (11)-(15), we transform (4)-(6) into the following profit-maximization problem (refer to the Online Supplement for details):

$$
\max_{N \geq 0, N \geq 0, \lambda_0 \geq 0} \Pi (N, N, \lambda_0) = \sum_{i=0}^{m} N_i V_i \mu_i - \lambda_0 c_0 - CN - S (\lambda_0) \quad (16)
$$

subject to

$$
N_0 \leq \lambda_0 s_0, \quad (17)
$$

$$
N_i \leq N_0 \mu_0 s_i, \quad i = 1, 2, ..., m, \quad (18)
$$

$$
\sum_{i=0}^{m} N_i \leq N. \quad (19)
$$

In the profit rate (16), the terms $N_i V_i \mu_i$ capture the value generated by allocating $N_i$ units of capacity to serve type-$i$ customer calls. The capacity allocation constraints (17)-(18) correspond to the service probability constraints in (5). By (17) the capacity consumed by new customer calls must not exceed their offered load. By (18) the capacity consumed by type-$i$ base customer calls must not exceed the product of new customer throughput by the offered load of type-$i$ base customer calls that a new customer generates. Finally, (19) captures the capacity constraint.

### 4.3 Service-Policy-Dependent New Customer Lifetime Value

Because our model captures the conversion of new to base customers, the lifetime value of a new customer depends on the priority policy, through the service probabilities of base customers. To capture this dependence and determine the maximum value of a new customer, we define priority-dependent new customer life time value metrics. These metrics include the weighted new customer OTV, $V_0 \mu_0$, which corresponds to the policy that serves no base customers.

Let $\bar{s}_i$ denote the total processing time requirement that a new customer generates under the policy that serves her first call and her subsequent calls as a base customer of type $j \leq i$. Then

$$
\bar{s}_i := \sum_{j=0}^{i} s_j, \quad i = 0, ..., m, \quad (20)
$$

$\bar{s}_m$ is the total offered load per new customer, and $\lambda_0 \bar{s}_m$ is the total offered load. We call a system underloaded if $\lambda_0 \bar{s}_m \leq N$, balanced if $\lambda_0 \bar{s}_m = N$, and overloaded if $\lambda_0 \bar{s}_m > N$.

Let $\bar{V}_i$ denote the expected total value per processing time of a new customer under this policy:

$$
\bar{V}_i := \frac{\sum_{j=0}^{i} s_j V_j \mu_j}{\bar{s}_i}, \quad i = 0, ..., m. \quad (21)
$$
Note that $\nabla_0 = V_0 \mu_0$. The quantity $\nabla_i$ is a service-policy-dependent and capacity-normalized measure of a new customer’s lifetime value. It follows from (8), (10), (15), and (21) that

$$\nabla_i = \frac{p_0 + c_0 + \sum_{j=1}^{i} \theta_{0j} L_j (1) + \sum_{j=i+1}^{m} \theta_{0j} L_j (0)}{s_i}, \quad i = 0, \ldots, m. \quad (22)$$

The numerator of $\nabla_i$ captures the value of a new customer under the policy that serves base customers of type $j \leq i$ (so their CLV is $L_j (1)$) but not those of type $j > i$ (so their CLV is $L_j (0)$), and the denominator of $\nabla_i$ captures the capacity required per new customer under this policy.

**Lemma 1** Define

$$k := 0 \text{ if } \nabla_0 > \nabla_1 \text{ and } k := \max \{1 \leq i \leq m : \nabla_{i-1} \leq \nabla_i\} \text{ if } \nabla_0 \leq \nabla_1. \quad (23)$$

Then $\nabla_k$ is the maximum value of a new customer per processing time:

$$\nabla_k = \max_{0 \leq i \leq m} \nabla_i. \quad (24)$$

Moreover,

$$\nabla_0 < \ldots \nabla_{k-1} \leq \nabla_k > \nabla_{k+1} > \ldots \nabla_m, \quad (25)$$

$$V_i \mu_i \geq \nabla_i \text{ for } i \leq k, \text{ and } \nabla_i > V_i \mu_i \text{ for } i > k. \quad (26)$$

When the advertising level is a decision variable, it is optimal to serve every new customer, so the value of serving a new customer is reduced by the abandonment cost $c_0$. Therefore, define

$$\tilde{\nabla}_i := \nabla_i - \frac{c_0}{s_i}, \quad i = 0, \ldots, m. \quad (27)$$

Note that $\tilde{\nabla}_0 = (V_0 - c_0) \mu_0$. The quantity $\tilde{\nabla}_i$ is the counterpart of $\nabla_i$ in settings where the advertising level is a decision variable rather than fixed. Lemma 2 parallels Lemma 1.

**Lemma 2** Define

$$k^* := 0 \text{ if } \tilde{\nabla}_0 > \tilde{\nabla}_1 \text{ and } k^* := \max \{1 \leq i \leq m : \tilde{\nabla}_{i-1} \leq \tilde{\nabla}_i\} \text{ if } \tilde{\nabla}_0 \leq \tilde{\nabla}_1. \quad (28)$$

Then $\tilde{\nabla}_{k^*}$ is the maximum value of a new customer, net of abandonment cost, per processing time:

$$\tilde{\nabla}_{k^*} = \max_{0 \leq i \leq m} \tilde{\nabla}_i. \quad (29)$$

Moreover, $k \leq k^*$,

$$\tilde{\nabla}_0 < \ldots \tilde{\nabla}_{k^*-1} \leq \tilde{\nabla}_{k^*} > \tilde{\nabla}_{k^*+1} > \ldots \tilde{\nabla}_m, \quad (30)$$

$$V_i \mu_i \geq \tilde{\nabla}_i \text{ for } i \leq k^*, \text{ and } \tilde{\nabla}_i > V_i \mu_i \text{ for } i > k^*. \quad (31)$$

We show in §5 that the optimal policy hinges on these service-policy-dependent measures of a new customer’s lifetime value, as well as on the weighted base customer OTVs, $V_i \mu_i$. Furthermore, we show in §6.3 that profitability may suffer if one treats customer value as independent of service levels, as is common in the marketing literature, and assumes that all calls must be served.
5 Optimal Priority Policy, Staffing, and Advertising

In this section we characterize the solution of the profit-maximization problem (16)-(19). In §5.1 we characterize the optimal priority for fixed capacity and advertising level, in §5.2 the optimal priority and staffing policies for fixed advertising level, and in §5.3 the optimal solution over all three controls. The results of §§5.1-5.2 can also be used at the tactical level to adjust the priority and capacity to predictable and/or unpredictable short-term arrival rate fluctuations.

5.1 Optimal Priority Policy for Fixed Staffing and Advertising Level

Consider the problem of choosing the priority policy for fixed capacity and new customer arrival rate. This situation may arise, for example, due to hiring lead times, time lags between advertising and demand response, unplanned demand bursts, or poor marketing-operations coordination.

Proposition 1 establishes that customers should be prioritized according to a simple index policy.

**Proposition 1** Fix the new customer arrival rate $\lambda_0$ and the number of servers $N$.

1. It is optimal to prioritize base customers of type $i \leq k$ over new customers, new customers over base customers of type $i > k$, and base customers in decreasing order of their $V_i \mu_i$-index.

2. The optimal capacity allocation satisfies

$$N_i^* = \begin{cases} 
\min (\lambda_0 \bar{s}_k, N) \frac{\bar{\sigma}_i}{\bar{\sigma}_k}, & i \leq k, \\
\min (\lambda_0 \bar{s}_i, (N - \lambda_0 \bar{s}_{i-1})^+) , & i > k. 
\end{cases}$$

3. The optimal profit rate is concave in the capacity $N$ and satisfies

$$\Pi^*(N, \lambda_0) := \min (\lambda_0 \bar{s}_k, N) \bar{\nu}_k + \sum_{i=k+1}^{m} \min (\lambda_0 \bar{s}_i, (N - \lambda_0 \bar{s}_{i-1})^+) V_i \mu_i - \lambda_0 c_0 - CN - S(\lambda_0).$$

The optimal priority policy is intuitive: It allocates the capacity to maximize the value generated per processing time. By Lemma 1, this requires serving all base customer calls of type $i \leq k$. Therefore, if $N < \lambda_0 \bar{s}_k$, it is optimal to turn away enough new customers to ensure service for all base customers of type $i \leq k$, and to serve none of type $i > k$. However, if $N \geq \lambda_0 \bar{s}_k$, it is optimal to serve all calls of type $i \leq k$, and those of type $i > k$ in decreasing order of their weighted OTVs.

*Effect of the Priority Policy on New Customer Value and Profits.* In the profit rate (33), under the optimal priority policy the value generated by a new customer per unit of processing time equals $\bar{\nu}_k$. If $k = 0$, it is optimal to prioritize new over all base customers, so the value of serving a new customer call is limited to its own OTV, and $\bar{\nu}_k = \bar{\nu}_0$. In contrast, if $k > 0$, it is optimal to prioritize some base customers over new ones, so the value of serving a new customer call also includes the OTVs of subsequent base customer calls of type $i \leq k$, and $\bar{\nu}_k = \sum_{j=0}^{k} s_j V_j \mu_j / \bar{s}_k$. 

16
Controlling Customer Retention through the Priority Policy. Proposition 1 underscores the key role of the priority policy, not only to differentiate the service level across members of the customer base, but also to control customer retention. Specifically, (3) implies that the optimal customer bases of types $l < j$ with optimal service levels $q^*_l = 1$ and $q^*_j = 0$, satisfy

$$\frac{x^*_l}{x^*_j} = \frac{\theta_{0l} \gamma_j + r_j (1 - \theta_j)}{\theta_{0j} \gamma_l}.$$ 

For equal customer base joining probabilities ($\theta_{0l} = \theta_{0j}$) and attrition rates ($\gamma_l = \gamma_j$),

$$\frac{x^*_l}{x^*_j} = 1 + \frac{r_j}{\gamma_j} (1 - \theta_j) > 1.$$ (34)

That is, the optimal service policy yields a larger customer base for types that are served.

Relationship to Standard Policies. The optimal policy in Proposition 1 bears some resemblance to standard priority policies, in that it is an index policy, like the $c\mu$ rule for systems without abandonment, or the $c\mu/\theta$ rule (Atar et al., 2010) and static $c\mu$-type policies (Tezcan and Dai 2010) for systems with abandonment. However, it has two important differences to these policies: (1) the $V_i\mu_i$ indices that are the basis for priorities in our model consider the effect of service levels on customers’ future calls and their financial impact. (2) The index $\nabla_k$ that determines the priority of one type – the new customers – may depend on the $V_i\mu_i$ indices of several other types. These differences reflect the key novelty of our model: It captures the conversion of new to base customers and the sojourn time in the customer base as functions of the queueing-related service quality.

Optimal Policy in Stochastic Systems. The profit depends on the priority policy only in conditions that result in throughput loss. In the fluid model, the system loses throughput if and only if it is overloaded (i.e., $N < \lambda_0 s_m$). However, a stochastic system may experience throughput loss even if it is underloaded. Therefore, prioritizing customers in line with Proposition 1 improves profits also in underloaded stochastic systems, as our simulation results in §7 confirm.

5.2 Jointly Optimal Priority and Staffing Policy for a Fixed Advertising Level

Consider the problem of setting the priority and staffing policies for fixed advertising (and new customer arrival rate). This problem arises because the advertising policy is often a strategic decision that affects the more tactical operational decisions, and also in response to unplanned bursts of arrivals. The optimal number of servers $N^*$ is the largest capacity level with nonnegative marginal profit. The marginal profit of an extra server equals the maximum value of the additional throughput it generates, minus its cost $C$. The optimal profit rate (33) in Proposition 1 implies the following optimal priority and staffing policies (we ignore integrality constraints).

**Proposition 2** Fix the new customer arrival rate $\lambda_0$. Under the optimal priority and staffing policies it is profitable to operate if, and only if, $\nabla_k > C$. In this case:
1. It is optimal to serve new customers, all base customers of type $i \leq k$, and base customers of type $j > k$ if and only if $V_j \mu_j \geq C$.

2. The optimal staffing is

$$N^* = \lambda_0 \left( \tilde{s}_k + \sum_{i=k+1}^{m} s_i 1\{V_i \mu_i \geq C\} \right),$$

and the system is overloaded if and only if $C > V_m \mu_m$.

3. The optimal profit rate is

$$\Pi^*(\lambda_0) := \lambda_0 \left( \tilde{s}_k (\tilde{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C)^+ \right) - S(\lambda_0).$$

An important implication of Proposition 2 is that operating an overloaded system is optimal under practically plausible conditions, whenever it is optimal to operate but some base customer types are unprofitable. Mathematically, this holds if and only if $\nabla_k > C > V_m \mu_m$, that is, the server cost is smaller than the maximum value of a new customer per processing time, but larger than the lowest weighted base customer OTV. As we show in §6.1, these conditions are easily met, which suggests that they may commonly hold in practice. Under these conditions, the optimal policy achieves two goals: (1) It serves all “high-value” base customers (of type $i \leq k$) and all new customers because they generate these base customers. Note that, taken on their own, these calls of new customers may not be profitable, that is, their weighted OTV may be lower than the server cost (i.e., $V_0 \mu_0 < C < \nabla_k$). (2) It serves “low-value” base customers (of type $i > k$) if and only if their weighted OTV exceeds the capacity cost. Importantly, the optimal policy deliberately does not serve unprofitable base customers who were acquired along with high-value customers.

5.3 Jointly Optimal Priority, Staffing, and Advertising Policy

Proposition 3 summarizes the solution to the profit-maximization problem (16)-(19).

**Proposition 3** Under the optimal advertising, priority and staffing policies it is profitable to operate if, and only if, $\tilde{V}_{k^*} > C$. In this case:

1. It is optimal to serve new customers, all base customers of type $i \leq k^*$, and base customers of type $j > k^*$ if and only if $V_j \mu_j \geq C$.

2. The optimal staffing is

$$N^* = \lambda_0 \left( \tilde{s}_{k^*} + \sum_{i=k^*+1}^{m} s_i 1\{V_i \mu_i \geq C\} \right),$$

and the system is overloaded if and only if $C > V_m \mu_m$.
3. The optimal new customer arrival rate and profit rate satisfy, respectively,

$$\lambda_0^* = \arg \left\{ \lambda_0 \geq 0 : \bar{s}_k (\bar{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C)^+ = S' (\lambda_0) \right\} > 0,$$

(38)

$$\Pi^* = \lambda_0^* \left( \bar{s}_k (\bar{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C)^+ \right) - S (\lambda_0^*).$$

(39)

The profitability condition and the optimal priority and staffing policies in Proposition 3 parallel their counterparts in Proposition 2, adjusted for the abandonment cost effect discussed in §4.3. In particular, Proposition 3 suggests that operating an overloaded system is sensible in practice; it is optimal if and only if $$\bar{V}_k > C > V_m \mu_m$$. These conditions are easily met as shown in §6.1.

The optimal advertising spending specified in (38) balances the maximum value that a new customer generates with the marginal cost of acquiring this customer. Like the optimal staffing level, the optimal advertising level depends on the optimal priority policy and the resulting CLV.

6 Implications and Extensions

In §6.1 we illustrate how the optimal service policy depends on customer attributes. §§6.2-6.4 focus on the case of homogeneous base customers, both for simplicity (our discussion also applies for heterogeneous base customers), and because this case arises naturally when the firm can perfectly target its customer acquisition by type. We discuss the case of homogeneous base customers and this connection to targeted new customer acquisition in §6.2, the effects of ignoring the service probability in the CLV in §6.3, and the effects of word of mouth about service quality in §6.4.

6.1 Effects of Customer Attributes on Optimal Service Policy: Examples

We illustrate Proposition 3 with two examples that show how the optimal service policy depends on customer attributes. In both examples we assume two base customer types (i.e., $$m = 2$$), $$p_0 = -10, c_0 = 0$$ for new customers; $$p_i = -10, c_i = 10, \theta_{0i} = 0.2$$, and $$r_i / \gamma_i = 10, i = 1, 2$$, for base customers, $$\mu_i = 1$$ for $$i \in \{0, 1, 2\}$$, and $$C = 50$$. By Lemma 2 and Proposition 3 the optimal system is overloaded if, and only if, $$\max (\bar{V}_0, \bar{V}_1) > C > V_2 \mu_2$$. The condition $$\max (\bar{V}_0, \bar{V}_1) > V_2 \mu_2$$ is equivalent to $$k^* < 2$$, that is, it is optimal to prioritize new customers over type-2 base customers. The capacity cost condition implies that it is optimal not to serve type-2 base customers.

Example 1: Effect of Service-Independent Profit on Optimal Service Policy. This example considers the case where type-1 generate a higher service-independent profit than type-2 customers: We fix $$R_1 / \gamma_1 = 1000$$ and vary $$R_2 / \gamma_2 \in [0, 1000]$$. We assume equally loyal base customers: $$\theta_1 = \theta_2 = 0.3$$. As shown in Figure 2, it is optimal to serve type-2 customers if and only if their service-independent profit is sufficiently large, i.e., $$R_2 / \gamma_2 \geq 680$$. At this threshold their
weighted OTV equals the capacity cost, $V_2\mu_2 = C$. This structure follows because type-2 customers’ OTV increases in their service-independent profit, and they are not very loyal in response to service failure. For $R_2/\gamma_2 \in [680, 820]$ it is optimal to prioritize new over type-2 customers ($k^* = 1$), the opposite holds for $R_2/\gamma_2 \geq 820$ ($k^* = 2$), but in both ranges the type-2 customers are profitable enough to be served at optimality. Finally, for $R_2/\gamma_2 \in [300, 680]$, it is optimal not to serve type-2 customers although their OTV exceeds that of new customers, i.e., $V_2\mu_2 > \bar{V}_0$.

As discussed in §5.1, the optimal priority policy also controls the customer base composition. Here, if both types have the same attrition rate ($\gamma_1 = \gamma_2$), then (34) implies that $x_1^* = 7x_2^*$ when it is optimal not to serve type 2 (i.e., for $R_2/\gamma_2 < 680$), whereas $x_1^* = x_2^*$ when both types are served.

**Example 2: Effect of Loyalty on Optimal Policy.** This example considers the case where after abandoning, type-1 are less loyal than type-2 customers: We set $\theta_1 = 0.3$ and vary $\theta_2 \in [0.3, 1.0]$. We assume equally lucrative base customers: $R_1 = R_2 = 800$. As shown in Figure 3, it is optimal to serve type-2 customers if and only if their loyalty coefficient is not too large, i.e., $\theta_2 \leq 0.75$. At this threshold their weighted OTV equals the capacity cost, $V_2\mu_2 = C$. This structure holds because base customers have a positive CLV even if they are not served (i.e., $R_i > r_ic_i$) and their OTV decreases in their loyalty coefficient. As in Example 1, there is a parameter range, in this case $\theta_2 \in [0.66, 0.75]$, where it is optimal to serve type-2 customers with lowest priority ($k^* = 1$).

### 6.2 Homogeneous Base Customers and Targeted New Customer Acquisition

Proposition 3 specializes as follows to the case of homogeneous base customers.
Figure 3: Optimal Policy as Function of Type-2 Loyalty. $\theta_1 = 0.3$. ($\tilde{V}_2 \geq \tilde{V}_1 \iff V_2\mu_2 \geq \tilde{V}_1$)

**Proposition 4** Suppose $m = 1$. Under the optimal advertising, priority and staffing policies it is profitable to operate if, and only if, $\max(\tilde{V}_0, \tilde{V}_1) > C$. In this case:

1. If $\tilde{V}_0 > C > V_1\mu_1$, it is optimal to prioritize and serve only new customers, the system is overloaded, $N^* = N^*_0 = \lambda^*_0 s_0$, and the optimal new customer arrival rate satisfies
   \[ \lambda^*_0 = \arg \left\{ \lambda_0 \geq 0 : s_0(\tilde{V}_0 - C) = S'_0 (\lambda_0) \right\} > 0. \]  
   \[ (40) \]

2. If $V_1\mu_1 \geq C$, it is optimal to serve all customers, the system is balanced, $N^* = \lambda^*_0 \tilde{s}_1$, and the optimal new customer arrival rate satisfies
   \[ \lambda^*_0 = \arg \left\{ \lambda_0 \geq 0 : \tilde{s}_1(\tilde{V}_1 - C) = S'_0 (\lambda_0) \right\} > 0, \]
   \[ (41) \]

where $\tilde{s}_1(\tilde{V}_1 - C) = s_0(\tilde{V}_0 - C) + s_1(V_1\mu_1 - C)$.

By Proposition 3, the optimal policy hinges on the maximum value of a new customer per processing time, $V_{i\mu_i}$, and on the weighted OTVs of lower-value base customers, $V_i\mu_i$ for $i > k^*$. By Proposition 4, for homogeneous base customers these conditions yield one of two optimal operating regimes. By Part 1, serving only new customers is optimal if $\tilde{V}_0 > C > V_1\mu_1$. That is, a new customer generates more value per processing time if base customers are not served (i.e., $\tilde{V}_0 > \tilde{V}_1$) and, taken on their own, base customer calls are not profitable ($C > V_1\mu_1$). In this case, by (40) the maximum value of a new customer net of capacity cost, $s_0(\tilde{V}_0 - C)$, accounts only for her initial call, and at optimality this value equals the marginal customer acquisition cost. By Part 2, serving
all customers is optimal if base customer calls are profitable. In this case, by (41) a new customer’s maximum value net of capacity cost, \( \bar{\gamma}_1(\bar{V}_1 - C) \), accounts for her initial and all her future calls.

A striking feature of our results is that it may be optimal to overloaded the system and deny service to some base customers. There are noteworthy connections among the optimality conditions for this regime, the heterogeneity of base customers, and the degree of targeting in new customer acquisition. By Proposition 3, for an overloaded system to be optimal, the maximum value of a new customer per processing time must be strictly larger than the lowest weighted base customer OTV, that is, \( \bar{V}_{k^*} > V_m \mu_m \). From (30)-(31) in Lemma 2, this condition is equivalent to

\[
\bar{V}_{m-1} > V_m \mu_m. \tag{42}
\]

Clearly for homogeneous base customers \( (m = 1) \), (42) can only hold if the weighted new customer OTV exceeds that of base customers (i.e., \( \bar{V}_0 > V_1 \mu_1 \), so that \( k^* = 0 \)).

In contrast, for heterogeneous base customers \( (m > 1) \) condition (42) is easily satisfied, even if the weighted new customer OTV is lower than all weighted base customer OTVs, that is \( \bar{V}_0 < V_m \mu_m \). (Refer to Example 1 where for some parameters \( \bar{V}_0 < V_2 \mu_2 \) yet it is optimal not to serve type 2.) Intuitively, with heterogeneous potential base customers the company easily finds itself in a situation where new customer calls are unprofitable, taken on their own, but generate some high-value base customer types, along with unprofitable types who are also acquired in the process. In this case, the optimal policy limits the loss from these customers through discriminatory service.

This connection between base customer heterogeneity and the optimality conditions for service denial also points to the value and effects of targeting new customer acquisition: If advertising could be perfectly targeted by base customer type, by identifying a priori the customer attributes of advertising targets, then the problem is separable in base customer types. That is, Proposition 4 applies separately to each base customer type and its corresponding new customer attributes. This suggests that the ability to target customer acquisition by type would yield a higher-value customer base that is composed of fewer types but with high service levels for more (or all) of them.

6.3 Ignoring the Effect of Service Probability on Customer Lifetime Value

Standard measures of customer lifetime value in the marketing literature ignore the service probability. In this section we demonstrate the suboptimal performance that may result if decision makers take customer value as independent of service levels and assume that all calls must be served. We focus on the case \( \bar{V}_0 > V_1 \mu_1 \), so by (22) and (27) we have \( \bar{V}_0 > \bar{V}_1 \). The CLV of a new customer decreases if all her calls as base customer are served. By Proposition 4 the optimal policy is to serve only new customers if \( \bar{V}_0 > C > V_1 \mu_1 \) and all customers if \( C \geq V_1 \mu_1 \). We contrast this policy with two alternative policies, marketing-driven and uncoordinated.
In the marketing-driven policy the marketing department controls both advertising and staffing, but assumes that all calls must be served. Let $\lambda_0^M$ and $N_0^M$ denote, respectively, the optimal new customer arrival rate and staffing under this policy. The new customer arrival rate $\lambda_0^M$ balances the marginal cost of acquiring a new customer with the profit when serving all her calls:

$$\lambda_0^M = \arg \left\{ \lambda_0 \geq 0 : \tilde{\pi}_1 (\tilde{V}_1 - C) = S' (\lambda_0) \right\} > 0 \text{ if } \tilde{V}_1 > C,$$

(43)

and $\lambda_0^M = 0$ otherwise. The staffing level satisfies $N_0^M = \lambda_0^M (s_0 + s_1)$, because all calls are served.

In the uncoordinated policy the marketing department also optimizes the advertising level assuming that all calls will be served, but the operations department optimizes the priority policy and staffing level (in line with Proposition 2), given the new customer arrival rate set by marketing. Let $\lambda_0^U$ and $N_0^U$ denote, respectively, the optimal new customer arrival rate and staffing level under this policy. Then $\lambda_0^U = \lambda_0^M$, because the new customer arrival rate in both policies is determined by (43). Because $V_0 \mu_0 \geq \tilde{V}_0 > V_1 \mu_1$, by Proposition 2 the optimal staffing corresponding to $\lambda_0^U$ satisfies $N_0^U = \lambda_0^U s_0$ if $\tilde{V}_0 > C > V_1 \mu_1$, and $N_0^U = \lambda_0^U (s_0 + s_1)$ if $V_1 \mu_1 \geq C$.

Both the marketing-driven and the uncoordinated policies reduce the advertising level for $C > V_1 \mu_1$, and may even cause the system to shut down (for $C > \tilde{V}_1$). This is the consequence of imposing a suboptimal service level that reduces the CLV of a new customer (since $\tilde{V}_0 > \tilde{V}_1$). The uncoordinated policy also yields a lower than optimal staffing level for $C > V_1 \mu_1$. However, in this cost range, the marketing-driven policy may yield lower or higher than optimal staffing, due to two countervailing effects: The new customer arrival rate is lower than optimal (i.e., $\lambda_0^M < \lambda_0^* M$) but all calls are served, rather than only those of new customers, as is optimal (i.e., $N_0^* M = \lambda_0^* M (s_0 + s_1)$ whereas $N_0^* = \lambda_0^* s_0$). The profit loss due to these suboptimal decisions can be significant, as evidenced by the extreme case, for $\tilde{V}_0 > C > \tilde{V}_1$, where they lead to suboptimal system shutdown.

This discussion underscores the importance of accounting for the optimal service level in evaluating customer lifetime value, particularly when basing substantial resource allocation decisions on this metric. Any policy that imposes an arbitrary uniform service level for all customer types, e.g., FIFO with an industry-standard abandonment probability, would be similarly suboptimal.

### 6.4 The Effects of Word of Mouth about Service Quality

We discuss how word of mouth (WOM) from base customers about their service quality affects our results. For brevity we focus on the jointly optimal advertising, staffing, and priority policies. Derivations and intermediate results are relegated to the Online Supplement.

We consider $\lambda_0$ to be the maximum new customer arrival rate that is generated by spending $S (\lambda_0)$ on advertising. To model negative WOM about service quality, we assume that the effective new customer arrival rate, $\lambda_0^e$, decreases in the base customer abandonment rate, $x_1 r_1 (1 - q_1)$.
\[
\lambda_0^e := \lambda_0 - \delta x_1 r_1 (1 - q_1) = \lambda_0 - \delta (x_1 r_1 - N_1 \mu_1).
\]  

The parameter \( \delta \geq 0 \) captures the WOM intensity. The second equality in (44) follows from (12).

We denote the effective new customer arrival rate by \( \lambda_0^e (N, \lambda_0) \), as it also depends on the capacity allocation \( N \) as discussed below. Let \( \Pi^w (N, N, \lambda_0) \) be the profit function with WOM.

WOM has two effects on problem (16)-(19). First, in the profit function the maximum new customer abandonment cost changes from \( \lambda_0 c_0 \), see (16), to \( \lambda_0 (N, \lambda_0) c_0 \). For \( m = 1 \) we get

\[
\Pi^w (N, N, \lambda_0) = N_0 V_0 \mu_0 + N_1 V_1 \mu_1 - \lambda_0^e (N, \lambda_0) c_0 - CN - S (\lambda_0).
\]

Second, the capacity allocation constraint for new customers changes from \( \nu \) to \( \lambda_0 (N, \lambda_0) s_0 \).

To consider how \( \lambda_0^e (N, \lambda_0) \) depends on the capacity allocation, let

\[
a := \frac{r_1 \theta_01}{\gamma_1 + r_1 (1 - \theta_1)}.
\]

Noting that the customer base, \( x_1 \), satisfies (13), we have from (44) and (47) that

\[
\lambda_0^e (N, \lambda_0) = \lambda_0 - \delta \left( N_0 \mu_0 a - N_1 \mu_1 \frac{\gamma_1}{r_1 \theta_01} a \right) = \lambda_0 - \delta a \left( \frac{N_0}{s_0} - \frac{N_1}{s_1} \right).
\]

Increasing \( N_0 \) increases the number of new customers served at a rate \( 1/s_0 \), which raises the base customer abandonment rate by \( a/s_0 \). However, increasing \( N_1 \) increases the number of base customers served at a rate \( 1/s_1 \), which reduces their abandonment rate by \( a/s_1 \). That is, \( a \) measures the increase (decrease) in base customer abandonments per new (base) customer served. If all base customers are served \( (N_1/s_1 = N_0/s_0) \), there is no negative WOM, and \( \lambda_0^e (N, \lambda_0) = \lambda_0 \) in (48).

In summary, with WOM the problem is to maximize the profit rate (45) subject to the capacity constraints (18)-(19), (46), and (48). Let \( \lambda_0^{w*} \) denote the optimal new customer arrival rate and \( N^{w*} \) the optimal staffing level with WOM. Proposition 5 summarizes the main effects of WOM.

**Proposition 5** WOM affects the optimal policy and reduces the optimal profit if, and only if, \( \tilde{V}_0 > C > V_1 \mu_1 \). Let

\[
\mathcal{V}_1^w (\delta) := \frac{s_0 \delta a \tilde{V}_0 + s_1 V_1 \mu_1}{s_0 \delta a + s_1}.
\]

1. If \( \tilde{V}_0 > \mathcal{V}_1^w (\delta) > C > V_1 \mu_1 \), then with WOM it is optimal to serve all customers, \( N^{w*} = \lambda_0^{w*} s_1 \), but without WOM it is optimal to serve only new customers, \( N^* = \lambda_0^e s_0 \). WOM yields a lower advertising level, \( \lambda_0^{w*} < \lambda_0^e \), but a higher or lower staffing level than without WOM.

2. If \( \tilde{V}_0 > C > \mathcal{V}_1^w (\delta) > V_1 \mu_1 \), then with or without WOM, it is strictly optimal to serve only new customers, but WOM reduces advertising and staffing levels: \( \lambda_0^{w*} < \lambda_0^* \) and \( N^{w*} = \lambda_0^{w*} s_0/(1 + \delta a) < N^* = \lambda_0^* s_0 \). In the limit as \( \delta \to \infty \), WOM makes it unprofitable to operate.
Proposition 5 shows that WOM reduces, but does not eliminate, the capacity cost range in which it is optimal to serve only new customers and operate an overloaded system. For an overloaded system that prioritizes new customers, the metric $\overline{V}_1^w(\delta)$ captures the average value per unit of capacity from serving additional base customer calls (with weighted OTV $V_1\mu_1$) and all the new customer calls (with weighted OTV $\tilde{V}_0$) that are generated as a result. The ratio $\delta a/(1 + \delta a)$ measures the WOM-driven increase in the effective new customer arrival rate per increase in base customer throughput. Without WOM ($\delta = 0$) this ratio is clearly zero and $\overline{V}_1^w(0) = V_1\mu_1$, as in the basic model. With extreme WOM ($\delta \to \infty$) this ratio is one, that is, each additional base customer served generates an additional new customers. In this case, WOM imposes a system where all customers are served, so that $\overline{V}_1^w(\infty) = \tilde{V}_1$, as in the basic model for a balanced system. Consistent with Proposition 4, a balanced system is not profitable if $C > \tilde{V}_1$.

7 Fluid Model Validation: Simulation Results

In this section, we evaluate the performance of the fluid model results developed in §5 against simulation results for the stochastic system described in §3.1. We compare (1) the fluid model prescriptions with those from simulation-based optimization, and (2) the simulation-based profit under the fluid model prescriptions with the simulation-based optimal profit.

We report results for one representative case where an overloaded system may be optimal in §7.1, for one case where a balanced system is optimal in §7.2, and for further cases that document the robustness of these results in §7.3. For simplicity, we focus on homogeneous base customers.

Throughout, we assume exponentially distributed service times and Poisson arrivals of new and base customers. For the advertising cost function we use the commonly assumed power model $S(\lambda_0) = \alpha\lambda_0^\beta$, where $\alpha > 0$ is a scale factor and $\beta > 1$ is the inverse of the customers’ response elasticity to the advertising level (cf. Hansssens et al. 2001). We let $\alpha = 0.5$, $\beta = 1.5$.

We assume a 24x7 operation. One time unit equals one day. We initialize the size of the customer base with the value suggested by the fluid model, run the simulation for 5,100,000 new customer arrivals, and discard the first 100,000 arrivals in computing performance measures.

For the simulation-based optimization, we vary the new customer arrival rate $\lambda_0$ in the interval $[1000,250000]$, increasing the step size with $\lambda_0$: We set the step size to 25, 250, 500, 1000, and 2500, respectively, in the arrival rate intervals $[1000,10000]$, $[10000,20000]$, $[20000,40000]$, $[40000,100000]$, and $[100000,250000]$. For each value of $\lambda_0$, we vary the number of servers $N$ in unit increments.

The fluid model results are applicable only for medium to large call centers, with 100 servers or more. We limit the study to this scale range: We vary the server cost, $C$, in increments of 100, from 500 to the smaller of 5000 and the largest $C$ for which Proposition 3 prescribes $N^* \geq 100$. 

25
7.1 Representative Case 1: Higher OTV for New Customers ($\widetilde{V}_0 > V_1\mu_1$)

In this section, we report the results for a representative case where the optimal system may be overloaded: The weighted OTV of new customers exceeds that of base customers, i.e., $k^* = 0$. Table 2 summarizes the parameter values for this case. From (7) the base customers’ CLV as a function of their service probability ranges from $L(0) = $331.67 to $L(1) = $450.00. From (8), (10), and (27), the weighted OTVs of new and base customers are $\widetilde{V}_0 = $10,950 and $V_1\mu_1 = $2,367, respectively.

The highest capacity cost considered is $C = $3,400, for a total of 30 experiments. Both in the fluid model and in the simulation, prioritizing new customers is optimal. Figure 4 visualizes the fluid model performance. Figure 4(a) shows that the error in the new customer arrival rate prescribed by the fluid model is lower than 6%. The error in the number of servers prescribed by the fluid model is also lower than 6%, except for $C \in [2300, 2900]$ where the errors are significant – up to around 60%. They arise in the neighborhood of the server cost where the fluid-optimal operating regime switches from balanced to overloaded. Specifically, by Proposition 4, the fluid-optimal number of servers $N^*$ drops discontinuously in the server cost at $C = V_1\mu_1 = 2,367$ (from 325 to 130), whereas the simulation-optimal server number decreases gradually in this neighborhood. On average the profit loss is 0.89%. Figure 4(b) shows that the profit loss peaks in the neighborhood of the cost $C = V_1\mu_1$. However, at 5.9% this profit loss is significantly smaller than the error in the number of servers prescribed by the fluid model. Intuitively, whereas the fluid-optimal number of servers is discontinuous in the server cost at $C = V_1\mu_1$, the profit rate is continuous at this cost. Outside the cost range $[2300, 2900]$, the profit loss is smaller than 1.5%.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service rate of customers (per day)</td>
<td>$\mu_i$</td>
</tr>
<tr>
<td>Call abandonment rate of customers (per day)</td>
<td>$\tau_i$</td>
</tr>
<tr>
<td>Call arrival rate per base customer (per day)</td>
<td>$r_1$</td>
</tr>
<tr>
<td>Prob. that new customer joins the customer base after service</td>
<td>$\theta_{01}$</td>
</tr>
<tr>
<td>Prob. that base customer remains in the customer base after abandoning</td>
<td>$\theta_1$</td>
</tr>
<tr>
<td>Call-independent departure rate per base customer (per day)</td>
<td>$\gamma_1$</td>
</tr>
<tr>
<td>Call-independent profit rate per base customer (per day)</td>
<td>$R_1$</td>
</tr>
<tr>
<td>Profit per served call of new, base customers</td>
<td>$p_0, p_1$</td>
</tr>
<tr>
<td>Cost per abandoned call of new, base customers</td>
<td>$c_0, c_1$</td>
</tr>
<tr>
<td>Advertising cost function parameters (power model)</td>
<td>$\alpha, \beta$</td>
</tr>
</tbody>
</table>

Table 2: Parameter Values for Representative Case 1: Higher OTV for New Customers ($\widetilde{V}_0 > V_1\mu_1$)

Figure 4: Fluid Model vs. Simulation: Percentage Errors in Optimal Number of Servers and New Customer Arrival Rate, and Percentage Profit Loss, as Functions of the Capacity Cost (Priority to New Customers).

7.2 Representative Case 2: Higher OTV for Base Customers \( (\tilde{V}_0 < V_1\mu_1) \)

In this section, we report the results for a representative case where the optimal system is balanced: The OTV of base customers is higher than the new customer OTV, i.e., \( k^* = 1 \). We set \( \theta_1 = 0.3 \). All other parameters remain the same as in Table 2. From (8), (10), and (27), \( \tilde{V}_0 = 4,317 \) and \( V_1\mu_1 = 6,789 \). For this case the highest capacity cost is \( C = 3900 \), for a total of 35 experiments.

Consistent with Proposition 1, prioritizing base customers is optimal in the simulation. The results are qualitatively similar to those for Case 1 (we omit the details due to space constraints): The fluid model prescriptions yield a near-optimal profit rate, and the percentage profit loss – on average 0.28%, is typically smaller than the error in the fluid-optimal number of servers and new customer arrival rate. However, in contrast to Case 1, here the error of the fluid model prescriptions and the profit loss do not peak in some intermediate range of \( C \), because the fluid model prescribes a balanced system for the entire cost range, so the fluid-optimal capacity is continuous in \( C \).

7.3 Robustness of the Fluid Model Prescriptions

To further test the robustness of the fluid model we simulate ten additional scenarios, to cover all the combinations of \( \theta_{01} \in \{0.3, 0.5\} \), \( \theta_1 \in \{0.3, 0.9\} \), and \( R_1 \in \{0.5, 1, 1.5\} \). All other parameters remain as in Table 2. For each scenario we consider on average approximately 31 capacity cost levels (ranging from three to 46 depending on the scenario), for a total of 313 experiments. The results for these scenarios are consistent with the ones reported in §§7.1-7.2. On average the profit loss is 1.7% in one scenario, and well below 1% in all others. We further find that the percentage profit loss under the fluid model prescriptions decreases in the call-independent base customer profit rate, \( R_1 \). Intuitively, the larger the call-independent component of the profit rate, the less significant is the profit effect of the fluid model error in estimating the abandonment rates.
8 Concluding Remarks

We develop a novel CRM-call center model to study the problem of maximizing profits by controlling customer acquisition, retention, and service quality via advertising, staffing, and priorities. The key distinction of our model is that it links the evolution and value of the customer base to past demand and queueing-related service quality, that is, to customers’ service probabilities. We characterize the optimal controls analytically based on a deterministic fluid approximation of the underlying stochastic model. For the case of homogenous base customers, we show via simulation that these prescriptions yield near-optimal performance. Our analysis yields several results and implications.

First, we provide new guidelines for call center management and CRM. These guidelines hinge on customer value metrics that depend on the service probabilities, and therefore reflect the effects of operations, unlike standard CLV metrics in the marketing literature. We show that the optimal service policy prioritizes base customers in decreasing order of their weighted OTVs (their $V_i\mu_i$-index) and gives new customers the priority level that maximizes their value per processing time (their $\overline{V}_k$-index). In contrast to standard priority policies, this policy accounts for the financial impact of customers’ future calls. We further show how the optimal staffing and advertising levels depend on these service-policy-dependent customer metrics, and the server and advertising costs. Specifically, we show that under practically plausible conditions, it is optimal to overload the system, whereby the firm serves only new and lucrative base customers, but not unprofitable base customers. This result highlights the key role of the priority policy for targeting retention spending (i.e., server costs) to profitable customers. By maximizing the expected life time value of a new customer, this targeted retention policy is also key for optimizing acquisition spending. Indeed, we show that marketing-focused policies that take CLV as independent of service levels and incorrectly assume all base customers must be served, yield lower than optimal advertising spending. Our results also suggest that the more advertising can be targeted to profitable customers, the lower the need for service denial to target retention spending, yielding a customer base with fewer, higher-value types, all (or most) receiving high service levels. These results highlight the importance of considering the effects of the service policy on CLV in balancing acquisition and retention spending, a principle that may also apply more broadly to CRM decisions in service settings beyond this model.

Second, from a modeling and methodological perspective, we conclude that our solution approach via the analysis of a deterministic fluid model is attractive, because the fluid model is both analytically tractable, yielding intuitively appealing prescriptions, and these prescriptions yield near-optimal profits for the approximated stochastic system. These results suggest that this approach may prove effective on further problems of joint CRM and call center management.

We close by outlining three research directions. First, in terms of customer modeling, it would be interesting to capture customer behavior that depends not only on the last interaction but on
longer service histories. Second, in terms of system modeling and solution method, a potentially fruitful avenue is to consider refinements to the fluid approximation we use in this paper, and to establish formal limit results. Third, in terms of estimation, many of our model parameters are reasonably well measurable based on data that call centers track. It would be interesting to estimate these parameters and also refine our model, based on such data. The results may be of value to quantify the impact of service quality attributes such as waiting time on CLV components.

References


Reinartz, W., J.S. Thomas, V. Kumar. 2005. Balancing acquisition and retention resources to maximize customer profitability. *J. Marketing* 69(1) 63–79.


Online Supplement: Proofs

Derivation of Problem Formulation (16)-(19). This problem obtains as follows from (4)-(6).

First consider the profit function (16). Substitute \( N_0q_0 = N_0\mu_0 \) from (11), \( x_ir_iq_i = N_i\mu_i \) for \( i \geq 1 \) from (12), and \( x_i \) from (13) into the profit function (4) to get

\[
\Pi = N_0\mu_0 (p_0 + c_0) + \sum_{i=1}^{m} \left( N_i\mu_i (p_i + c_i) + \frac{N_0\mu_0\theta_{0i} + N_i\mu_i (1 - \theta_i)}{\gamma_i + r_i (1 - \theta_i)} (R_i - r_ic_i) \right) - \lambda_0c_0 - CN - S(\lambda_0).
\]

Noting from (7)-(9) that

\[
V_0 = p_0 + c_0 + \sum_{i=1}^{m} \frac{R_i - r_ic_i}{\gamma_i + r_i (1 - \theta_i)},
\]

\[
V_i = p_i + c_i + (1 - \theta_i) \frac{R_i - r_ic_i}{\gamma_i + r_i (1 - \theta_i)}, \quad i = 1, 2, ..., m,
\]

and substituting into (50) yields (16).

The capacity allocation constraints (17)-(18) follow from the service probability constraints (5): \( q_0 \leq 1 \) and (11) and (14) yield (17); and \( q_i \leq 1 \) for \( i \geq 1 \), and (12), (13), and (15) yield (18).

Finally, the capacity constraint (19) is immediate from (6) and (11)-(12).

**Proof of Lemma 1.** Clearly (25) implies (24). The properties (25) and (26) follow from two facts. (i) From (21), \( \bar{V}_i \) is a convex combination of \( \bar{V}_{i-1} \) and \( V_i\mu_i \):

\[
\bar{V}_i = \frac{s_{i-1}\bar{V}_{i-1} + s_iV_i\mu_i}{s_{i-1} + s_i}, \quad i = 1, 2, ..., m,
\]

so \( \bar{V}_i \) is smaller than \( \bar{V}_{i-1} \), equal to \( \bar{V}_{i-1} \), or larger than \( \bar{V}_{i-1} \), for every \( i \) such that \( \bar{V}_{i-1} > V_i\mu_i \), \( \bar{V}_{i-1} = V_i\mu_i \), or \( \bar{V}_{i-1} < V_i\mu_i \), respectively. (ii) By assumption \( V_i\mu_i > V_{i+1}\mu_{i+1} \) for \( i \geq 1 \).

**Proof of Lemma 2.** The proof of (30)-(31) is identical to that of (25)-(26). By the definition of \( k^* \) in (28), the inequality \( k \leq k^* \) is equivalent to \( \bar{V}_{k-1} \leq \bar{V}_k \) if \( k \geq 1 \). Using (27) we have

\[
\bar{V}_{k-1} \leq \bar{V}_k \iff \bar{V}_{k-1} - \bar{V}_k \leq c_0 \left( \frac{1}{s_{k-1}} - \frac{1}{s_k} \right),
\]

which is satisfied because \( \bar{V}_{k-1} - \bar{V}_k \leq 0 \) by (25), \( c_0 \geq 0 \), and \( s_k > s_{k-1} > 0 \).

**Proof of Proposition 1.** Part 1. Prioritizing base customers in decreasing order of their \( V_i\mu_i \)-index is equivalent to the capacity allocation constraints

\[
N_i < N_0\mu_0s_i \implies N_{i+1} = 0, \quad i = 1, ..., m - 1.
\]

That is, customers of type \( i + 1 \) are only served if all of type \( i \) are served (constraint (18) is binding for \( i \)). This is optimal because \( V_i\mu_i > V_{i+1}\mu_{i+1} \) for \( i \geq 1 \) and by inspection of (16) and (18)-(19).
We next establish the optimal priority of new customers. Let $\Pi^l(N, \lambda_0)$ denote the maximum profit under the policy that prioritizes base customers of type $i \in \{1, \ldots, l\}$ over new customers, new customers over base customers of type $i > l$, and base customers in decreasing order of their $V_i\mu_t$-index. This policy is equivalent to (52) and the following capacity allocation constraints. All base customer calls of type $i \in \{1, \ldots, l\}$ are served, i.e., (18) is binding:

$$N_i = N_0\mu_0s_i, \quad i = 1, \ldots, l,$$

and if some new customers are not served (so (17) is slack) then no calls of type $i > l$ are served:

$$N_0 < \lambda_0s_0 \implies N_i = 0, \quad i > l. \quad (54)$$

To complete the proof of Part 1 we show that $\Pi^k(N, \lambda_0) \geq \Pi^l(N, \lambda_0)$ for $l \neq k$. The profit $\Pi^l(N, \lambda_0)$ is the maximum value of the objective function in (16) under a capacity allocation vector $\mathbf{N}$ that satisfies (17)-(19) and (52)-(54). Using (20), (21), and (53) we get

$$\sum_{i=0}^{l} N_i V_i \mu_i = N_0 \mu_0 s_l V_l \text{ and } \sum_{i=0}^{l} N_i = N_0 \mu_0 s_l,$$

and substituting into (16) and (19), respectively, it is straightforward to verify that

$$\Pi^l(N, \lambda_0) = \min (\lambda_0 s_l, N) V_l + \sum_{i=l+1}^{m} \min (\lambda_0 s_i, (N - \lambda_0 s_{i-1})^+) V_i \mu_i - \lambda_0 c_0 - CN - S(\lambda_0). \quad (55)$$

It remains to show $\Pi^k(N, \lambda_0) \geq \Pi^l(N, \lambda_0)$ for $l \neq k$. We have

$$\Pi^l(N, \lambda_0) - \Pi^{l-1}(N, \lambda_0) = \min (\lambda_0 s_l, N) V_l - \min (\lambda_0 s_{l-1}, N) V_{l-1} - \min (\lambda_0 s_l, (N - \lambda_0 s_{l-1})^+) V_l \mu_l.$$

Using (51) to substitute for $V_l$, and noting that

$$\min (\lambda_0 s_l, (N - \lambda_0 s_{l-1})^+) = \min (\lambda_0 s_l, N) - \min (\lambda_0 s_{l-1}, N),$$

we have

$$\tilde{s}_l \left( \Pi^l(N, \lambda_0) - \Pi^{l-1}(N, \lambda_0) \right) = (V_l \mu_l - V_{l-1}) (\tilde{s}_l \min (\lambda_0 s_{l-1}, N) - \tilde{s}_{l-1} \min (\lambda_0 s_l, N)).$$

Note that $V_l \mu_l \geq V_{l-1} \iff V_l \geq V_{l-1}$ by (51) and $\tilde{s}_l \min (\lambda_0 s_{l-1}, N) - \tilde{s}_{l-1} \min (\lambda_0 s_l, N) \geq 0$, so

$$\Pi^l(N, \lambda_0) \geq \Pi^{l-1}(N, \lambda_0) \iff V_l \geq V_{l-1}.$$

It follows from (25) in Lemma 1 that $\Pi^k(N, \lambda_0) \geq \Pi^l(N, \lambda_0)$ for $l \neq k$.

Part 2. The capacity allocation specified in (32) is implied by (52)-(54) for $l = k$.

Part 3. By Part 1 and (55), $\Pi^*(N, \lambda_0)$ satisfies (33). Lemma 1 implies that $\Pi^*(N, \lambda_0)$ is concave in $N$. ■

**Proof of Proposition 2.** The condition for profitable operation, and Parts 1-2 follow from the profit function (33) and from (25)-(26) in Lemma (1): (i) if $C \geq V_k$ then $\Pi^*(N, \lambda_0) \leq 0$ for
\( N \geq 0, \) and (ii) if \( C < V_k \) then \( \Pi^*(N, \lambda_0) \) is strictly increasing for \( N \in [0, \lambda_0 s_k] \), nondecreasing for \( N \in [\lambda_0 s_{i-1}, \lambda_0 s_i] \) and \( i = k+1, \ldots, m \) if and only if \( V_i \mu_i \geq C \), and strictly decreasing for \( N > \lambda_0 s_m \). This establishes (35). It follows from the optimal priority rule and the capacity allocation (32) of Proposition 1, that it is optimal to serve the customer types as claimed. Furthermore, because \( V_i \mu_i > V_{i+1} \mu_{i+1} \) for \( i \geq 1 \), the system is overloaded, i.e., \( N^* < \lambda_0 s_m \) if and only if \( C > V_m \mu_m \).

Part 3. The profit rate (36) follows from (33), by using (35) to substitute for \( N^* \) and by noting that \( s_k \bar{V}_k - c_0 = \bar{s}_k \bar{V}_k \) from (27).

**Proof of Proposition 3.** We first prove the condition for profitable operation, that is, \( \max_{\lambda_0 \geq 0} \Pi^* (\lambda_0) > 0 \) if and only if \( C < \bar{V}_k^* \), where \( \Pi^* (\lambda_0) \) is given by (36) for \( C < \bar{V}_k \). Note that this condition for profitable operation is more restrictive than \( C < \bar{V}_k \) and less restrictive than \( C < \bar{V}_k \), because \( \bar{V}_k^* \leq \bar{V}_k \) from (24), (27) and (29), and \( \bar{V}_k \leq \bar{V}_k \) from Lemma 2.

Because \( \Pi''^* (\lambda_0) < 0 < S'' (\lambda_0) \) and \( \Pi^* (0) = 0 \), the maximizer \( \lambda_0^* \) is unique, with \( \lambda_0^* = 0 \) if \( \Pi''^* (0) \leq 0 \), and \( \lambda_0^* = \arg \{ \lambda_0 \geq 0 : \Pi''^* (\lambda_0) = 0 \} > 0 \) if \( \Pi''^* (0) > 0 \). Because \( S'' (0) = 0 \), (36) implies

\[
\Pi''^* (0) = \bar{s}_k (\bar{V}_k - C) + \sum_{i=k+1}^{m} s_i (V_i \mu_i - C)^+.
\]

(56)

Note that \( \Pi''^* (0) \) strictly decreases in \( C \). Therefore, \( \Pi''^* (0) > 0 \Leftrightarrow C < \bar{V}_k^* \) holds if, and only if, \( \Pi''^* (0) = 0 \) for \( C = \bar{V}_k^* \), which we show next. For \( C = \bar{V}_k^* \) we have from (56) that

\[
\Pi''^* (0) = \bar{s}_k (\bar{V}_k - C) + \sum_{i=k+1}^{k^*} s_i (V_i \mu_i - C) = \bar{s}_k^* (\bar{V}_k^* - C) = 0,
\]

where the first equality follows because \( k \leq k^* \) (Lemma 2) and \( V_j \mu_j \geq \bar{V}_k^* > V_i \mu_i \) for \( 1 \leq j \leq k^* < l \) ((30)-(31) in Lemma 2 and \( V_i \mu_i \) decreases in \( i \), and the second equality from (21) and (27).

Parts 1-2 follow from Parts 1-2 of Proposition 2 because \( \bar{V}_k^* \leq \bar{V}_k \) and \( \bar{V}_k^* \leq V_j \mu_j \) for \( 1 \leq j \leq k^* \).

Part 3 follows by noting that for \( C < \bar{V}_k^* \), the analysis in Part 1 implies

\[
\Pi''^* (0) = \bar{s}_k^* (\bar{V}_k^* - C) + \sum_{i=k^*+1}^{m} s_i (V_i \mu_i - C)^+ > 0,
\]

and noting that \( \Pi^* (\lambda_0) = \lambda_0 \Pi''^* (0) - S (\lambda_0) \) in light of (36) yields (38) and (39).

**Proof of Proposition 4.** Apply Proposition 3 with \( m = 1 \), for Part 1 using \( k^* = 0 \) and \( C > V_1 \mu_1 \), and for Part 2 using \( k^* = 0 \) or \( k^* = 1 \) and \( C \leq V_1 \mu_1 \).

**Proof of Proposition 5.** For convenience we restate the problem with WOM as follows:

\[
\max_{N \geq 0, N \geq 0, \lambda_0 \geq 0} \Pi^w (N, N, \lambda_0) = N_0 V_0^w \mu_0 + N_1 V_1^w \mu_1 - \lambda_0 c_0 - CN - S (\lambda_0)
\]

s.t.

\[
N_0 \leq s_0 \left( \frac{\lambda_0}{1 + \delta a} + N_1 \frac{1}{\delta s_1} \right),
\]

\[
N_1 \leq N_0 \mu_0 s_1,
\]

\[
N_0 + N_1 \leq N.
\]
where the customer metrics
\[ V_0^w : = V_0 + \delta c_0 a, \]
\[ V_1^w : = V_1 - \delta c_0 \frac{\gamma_1}{r_1 \theta_0} a, \]
adjust those in the basic model by accounting for the indirect effects of the capacity allocation \( \mathbf{N} \) on the abandonment cost through the effective new customer arrival rate. The profit rate (57) follows from (45), (48), and (61)-(62). The constraint (58) follows from (46) and (48).

We proceed in four steps.
1. Lemma 3 characterizes the optimal capacity allocation \( \mathbf{N}^w^* \) for fixed \( N \) and \( \lambda_0 \).
2. Lemma 4 characterizes the jointly optimal \( \mathbf{N}^w^* \) and \( N^w^* \) for fixed \( \lambda_0 \).
3. Lemma 5 characterizes the jointly optimal \( \mathbf{N}^w^*, N^w^* \) and \( \lambda_0^w^* \).
4. We use Proposition 4 and Lemma 5 to prove the claims of Proposition 5.

We use the following relationships among the customer value metrics.

\[ \nabla_1 = \frac{s_0 V_0 \mu_0 + s_1 V_1 \mu_1}{s_0 + s_1} = \frac{s_0 V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 + s_1}, \]
\[ \nabla_1^w (\delta) = \frac{s_0 \frac{\delta a}{1+\delta a} V_0^w + s_1 V_1^w \mu_1}{s_0 + s_1} = \frac{s_0 \frac{\delta a}{1+\delta a} V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 + s_1}. \]

The first equalities in (63) and (64) hold by (21) and (49), respectively, the second equalities because \( s_0 V_0^w \mu_0 = s_0 V_0 \mu_0 + \delta c_0 a \) by (61), \( s_1 V_1^w \mu_1 = s_1 V_1 \mu_1 - \delta c_0 a \) by (62), and \( \bar{V}_0 = (V_0 - c_0)\mu_0 \) by (27).

\textbf{Lemma 3} Fix the new customer arrival rate \( \lambda_0 \) and the number of servers \( \mathbf{N} \).
If \( V_0^w \mu_0 > V_1^w \mu_1 \) then it is optimal to prioritize new customers, the optimal capacity allocation is

\[ N_0^w^* = \min \left( N, \frac{\lambda_0 s_0}{1+\delta a} + (N - \frac{\lambda_0 s_0}{1+\delta a}) \frac{s_0 \frac{\delta a}{1+\delta a}}{s_0 \frac{\delta a}{1+\delta a} + s_1}, \lambda_0 s_0 \right), \]
\[ N_1^w^* = \min \left( (N - \frac{\lambda_0 s_0}{1+\delta a}) + \frac{s_1}{s_0 \frac{\delta a}{1+\delta a} + s_1}, \lambda_0 s_0 \right), \]

and the optimal profit rate is

\[ \Pi^w^*(N, \lambda_0) = -\lambda_0 c_0 - S(\lambda_0) - CN + \begin{cases} \frac{\lambda_0 s_0}{1+\delta a} V_0^w \mu_0 + (N - \frac{\lambda_0 s_0}{1+\delta a}) \nabla_1^w (\delta), & \lambda_0 s_0 \leq \frac{1}{1+\delta a}, \\ \lambda_0 s_0 \nabla_1, & \frac{1}{1+\delta a} < N < \lambda_0 s_1, \\ \lambda_0 s_1 \nabla_1, & \lambda_0 s_1 \leq N. \end{cases} \]

If \( V_0^w \mu_0 \leq V_1^w \mu_1 \) then it is optimal to prioritize base customers, the optimal capacity allocation is

\[ N_i^w^* = \min \left( \lambda_0 s_1, N \right) \frac{s_i}{s_1}, \quad i = 0, 1, \]

and the optimal profit rate is

\[ \Pi^w^*(N, \lambda_0) = \min \left( \lambda_0 s_1, N \right) \nabla_1 - \lambda_0 c_0 - CN - S(\lambda_0). \]
Proof. If \( V_0^w \mu_0 > V_1^w \mu_1 \), it is clear from the profit function (57), and the convexity of the feasible region (58)-(60), that the solution maximizes \( N_0 \). Therefore it is optimal to strictly prioritize new customers. We establish \( N^w \) in (65)-(66) for each of the possible capacity intervals:

If \( N \leq \lambda_0 s_0 / (1 + \delta a) \), then (60) binds, so \( N^w = N \) and \( N^w_1 = 0 \), which agrees with (65)-(66).

If \( \lambda_0 s_0 / (1 + \delta a) < N < \lambda_0 \bar{s}_1 \), then (58) and (60) are binding, which yields

\[
N^w_0 = \frac{\lambda_0 s_0}{1 + \delta a} + (N - \frac{\lambda_0 s_0}{1 + \delta a}) \frac{s_0 \delta a}{s_0 \delta a + s_1},
\]

(70)

\[
N^w_1 = \frac{s_1}{s_0 \delta a + s_1},
\]

(71)

If \( \lambda_0 \bar{s}_1 \leq N \), then (58)-(59) bind, so \( N^w_0 = \lambda_0 s_0 \) and \( N^w_1 = \lambda_1 s_1 \). In this case there is enough capacity to serve all calls, and the effective new customer arrival rate equals \( \lambda_0 \).

That the solutions for \( N > \lambda_0 s_0 / (1 + \delta a) \) agree with (65)-(66) follows because \( N^w_0 \) in (70) satisfies \( N^w_0 < \lambda_0 s_0 \) and \( N^w_1 \) in (71) satisfies \( N^w_1 < \lambda_0 \bar{s}_1 \) if, and only if, \( N < \lambda_0 \bar{s}_1 \).

The profit (67) follows by substituting \( N^w \) in (57), using (64) and (70)-(71) for \( \lambda_0 s_0 / (1 + \delta a) < N < \lambda_0 \bar{s}_1 \), and (63) and \( N^w_1 = \lambda_0 \bar{s}_1 \) for \( \lambda_0 \bar{s}_1 \leq N \).

Next consider \( V_0^w \mu_0 \leq V_1^w \mu_1 \). It is clear from the profit function (57) and because the feasible region (58)-(60) is convex, that the solution maximizes \( N_1 \), so it is optimal to strictly prioritize base customers. That is, (59) is binding, so \( N_1 = N_0 s_1 / s_0 \) and (58) simplifies to \( N_0 \leq s_0 \lambda_0 \), which yields \( N^w \) in (68). The profit (67) follows by substituting \( N^w \) in (57), using (63).

 Lemma 4 Fix the new customer arrival rate \( \lambda_0 \). Under the optimal priority and staffing policies it is profitable to operate if, and only if, \( C < \max(V_0^w \mu_0, V_1^w) \). In this case:

1. If \( V_0^w \mu_0 > C > V_1^w (\delta) \), it is optimal to prioritize and serve only new customers, the system is overloaded, \( N^w = N^w_0 = \lambda_0 s_0 / (1 + \delta a) \), and the optimal profit is

\[
\Pi^w (\lambda_0) = \frac{\lambda_0}{1 + \delta a} (s_0(V_0 - C) - c_0) - S(\lambda_0).
\]

(72)

2. If \( V_0^w \mu_0 > V_1^w (\delta) \geq C \), or if \( V_1^w (\delta) \geq V_0^w \mu_0 \) and \( V_1 > C \), it is optimal to serve all customers, the system is balanced \( N^w = \lambda_0 \bar{s}_1 \), and the optimal profit is

\[
\Pi^w (\lambda_0) = \lambda_0 (\bar{s}_1(V_1 - C) - c_0) - S(\lambda_0).
\]

(73)

Proof. The results follow from Lemma 3 and the two possible rankings of \( V_0^w \mu_0 \) and \( V_1^w \mu_1 \): If \( V_0^w \mu_0 > V_1^w \mu_1 \) then

\[
V_0^w \mu_0 > V_1 = \frac{s_0 V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 + s_1} > \frac{s_0 \delta a}{s_0 \delta a + s_1} \frac{V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 \delta a + s_1} = V_1^w (\delta) > V_1^w \mu_1.
\]

(74)

The first equality holds by (63), the second by (64). The inequalities follow from \( V_0^w \mu_0 > V_1^w \mu_1 \). If \( V_0^w \mu_0 \leq V_1^w \mu_1 \) then it follows similarly that

\[
V_0^w \mu_0 \leq V_1 = \frac{s_0 V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 + s_1} \leq \frac{s_0 \delta a}{s_0 \delta a + s_1} \frac{V_0^w \mu_0 + s_1 V_1^w \mu_1}{s_0 \delta a + s_1} = V_1^w (\delta) \leq V_1^w \mu_1.
\]

(75)
The condition for profitable operation, \( \max_{N \geq 0} \Pi^{w^*}(N, \lambda_0) > 0 \) if and only if \( C < \max(V_0^w \mu_0, \nabla_1) \), follows from (74)-(75) and the profit functions (67) for \( V_0^w \mu_0 > V_1^w \mu_1 \) and (69) for \( V_0^w \mu_0 \leq V_1^w \mu_1 \).

Part 1. If \( V_0^w \mu_0 > C > \nabla_1^w (\delta) \) then \( V_0^w \mu_0 > V_1^w \mu_1 \), so by Lemma 3 it is optimal to prioritize new customers and \( \Pi^{w^*}(N, \lambda_0) \) satisfies (67), which is maximized at \( N^{w^*} = N_0^{w^*} = \lambda_0 \sigma_0 / (1 + \delta) \). Substituting into (67) yields (72), because \( s_0 V_0^w \mu_0 = s_0 \nabla_0 + \delta \sigma_0 \) from (21), (61) and \( s_0 \mu_0 = 1 \).

Part 2. If \( C < \max(V_0^w \mu_0, \nabla_1) \) but the conditions of Part 1 do not hold, then (74)-(75) imply that either \( V_0^w \mu_0 > \nabla_1^w (\delta) \geq C \), or \( \nabla_1^w (\delta) \geq V_0^w \mu_0 \) and \( \nabla_1 > C \), and furthermore, \( \Pi^{w^*}(N, \lambda_0) \) in (67) or (69) is maximized at \( N^{w^*} = \lambda_0 \overline{s}_1 \): Serving all customers is optimal under either priority policy. Substituting into (67) or (69) yields (73). □

**Lemma 5** Under the optimal advertising, priority and staffing policies it is profitable to operate if, and only if, \( C < \max(\overline{V}_0, \overline{V}_1) \). In this case:

1. If \( \overline{V}_0 > C > \nabla_1^w (\delta) \), it is optimal to prioritize and serve only new customers, the system is overloaded, \( N^{w^*} = N_0^{w^*} = \lambda_0^{w^*} s_0 / (1 + \delta) \), and the optimal new customer arrival rate satisfies

\[
\lambda_0^{w^*} = \arg \left\{ \lambda_0 \geq 0 : \frac{1}{1 + \delta} s_0 (\overline{V}_0 - C) = S' (\lambda_0) \right\} > 0.
\]

2. If \( \overline{V}_0 > \nabla_1^w (\delta) \geq C \), or if \( \overline{V}_0 \leq \nabla_1^w (\delta) \) and \( C < \overline{V}_1 \), it is optimal to serve all customers, the system is balanced, \( N^{w^*} = \lambda_0^{w^*} \overline{s}_1 \), and the optimal new customer arrival rate satisfies

\[
\lambda_0^{w^*} = \arg \left\{ \lambda_0 \geq 0 : \overline{s}_1 (\overline{V}_1 - C) = S' (\lambda_0) \right\} > 0.
\]

**Proof.** By Lemma 4, for fixed \( \lambda_0 \) it is profitable to operate if and only if \( C < \max(V_0^w \mu_0, \nabla_1) \), in which case the profit function \( \Pi^{w^*}(\lambda_0) \) satisfies (72) or (73). From (27) it follows that \( s_0 (\overline{V}_0 - C) - c_0 = s_0 (\overline{V}_0 - C) \) in (72) and \( \overline{s}_1 (\overline{V}_1 - C) - c_0 = \overline{s}_1 (\overline{V}_1 - C) \) in (73). Therefore, if \( C \geq \max(\overline{V}_0, \overline{V}_1) \) then \( \max_{\lambda_0 \geq 0} \Pi^{w^*}(\lambda_0) = 0 \), so it is not profitable to operate.

Parts 1-2 cover the two possible cases for \( C < \max(\overline{V}_0, \overline{V}_1) \). Note that \( \overline{V}_1 = (s_0 \overline{V}_0 + s_1 \overline{V}_1 \mu_1) / (s_0 + s_1) \) by (21) and (27), and \( \nabla_1^w (\delta) \) by (64) are convex combinations of \( \overline{V}_0 \) and \( V_1 \mu_1 \), and \( \nabla_1 (\delta) \) puts a lower weight on \( \overline{V}_0 \). Therefore one of the following rankings holds:

\[
\overline{V}_0 > \overline{V}_1 > \nabla_1^w (\delta) > V_1 \mu_1, \quad (78)
\]
\[
\overline{V}_0 \leq \overline{V}_1 \leq \nabla_1^w (\delta) \leq V_1 \mu_1. \quad (79)
\]

Part 1. If \( \overline{V}_0 > C > \nabla_1^w (\delta) \), then because \( V_0^w \mu_0 \geq \overline{V}_0 \) by (27) and (61), the conditions of Part 1 of Lemma 4 hold. Therefore, it is optimal to prioritize and serve only new customers, and

\[
\Pi^{w^*}(\lambda_0) = \frac{\lambda_0}{1 + \delta} s_0 (\overline{V}_0 - C) - S (\lambda_0)
\]

from (72) and (27). Because \( \Pi^{w^*}(\lambda_0) < 0 < S''(\lambda_0) \) and \( \Pi^*(0) = 0 \), the maximizer \( \lambda_0^{w^*} \) is uniquely defined as \( \lambda_0^{w^*} = 0 \) if \( \Pi^{w^*}(0) \leq 0 \) and \( \lambda_0^{w^*} = \arg \{ \lambda_0 \geq 0 : \Pi^{w^*}(\lambda_0) = 0 \} > 0 \) if \( \Pi^{w^*}(0) > 0 \). Noting that

\[
\Pi^{w^*}(\lambda_0) = \frac{1}{1 + \delta} s_0 (\overline{V}_0 - C) - S' (\lambda_0),
\]
and $S'(0) = 0$, it follows that $\Pi^{w*}(0) > 0$ and $\lambda_0^{w*}$ satisfies (76).

Part 2. If $C < \max(\bar{V}_0, \bar{V}_1)$ but the conditions of Part 1 do not hold, then (78)-(79) imply that either $\bar{V}_0 > \bar{V}_1^w(\delta) \geq C$, or $\bar{V}_0 \leq \bar{V}_1^w(\delta)$ and $C < \bar{V}_1$. Furthermore, the conditions of Part 2 of Lemma 4 hold: If $\bar{V}_0 > \bar{V}_1^w(\delta) \geq C$ then $V_0^w \mu_0 > \bar{V}_1^w(\delta) \geq C$. If $\bar{V}_0 \leq \bar{V}_1^w(\delta)$ and $C < \bar{V}_1$ then $C < \min(\bar{V}_1, \bar{V}_1^w(\delta))$ because $\bar{V}_1 \leq \bar{V}_1$ by (27) and $\bar{V}_1 \leq \bar{V}_1^w(\delta)$ by (79), so that either $V_0^w \mu_0 > \bar{V}_1^w(\delta) \geq C$, or $\bar{V}_1^w(\delta) \geq V_0^w \mu_0$ and $\bar{V}_1 > C$ holds. By Part 2 of Lemma 4, it is optimal to serve all customers, and we have from (73) and (27) that

$$\Pi^{w*}(\lambda_0) = \lambda_0 \bar{V}_1(1 - C) - S(\lambda_0).$$

Because $S'(0) = 0$, we have $\Pi^{w*}(0) > 0$. By the same argument as in Part 1, $\lambda_0^{w*}$ satisfies (77).□

It remains to prove the claims of Proposition 5, which follow from Proposition 4 and Lemma 5. If $\bar{V}_0 > C > V_1 \mu_1$ does not hold, then (78)-(79) imply that Part 2 of Proposition 4 and Part 2 of Lemma 5 apply, so the solution is the same with or without WOM. Suppose that $\bar{V}_0 > C > V_1 \mu_1$.

Part 1. If $\bar{V}_0 > \bar{V}_1^w(\delta) > C > V_1 \mu_1$, then Part 1 of Proposition 4 and Part 2 of Lemma 5 apply, so $\lambda_0^s$ satisfies (40) and $\lambda_0^{w*}$ satisfies (77). We have $\lambda_0^{w*} < \lambda_0^s$ and $\Pi^{w*}(\lambda_0^{w*}) < \Pi^s(\lambda_0^s)$, because $\Pi^{w*}(\lambda_0)$ and $\Pi^s(\lambda_0)$ are strictly concave and $C > V_1 \mu_1$ implies the inequality in

$$\Pi^{w*}(\lambda_0) = s_0(\bar{V}_0 - C) + s_1(V_1 \mu_1 - C) - S'(\lambda_0) < s_0(\bar{V}_0 - C) - S'(\lambda_0) = \Pi^s(\lambda_0).$$

Part 2. If $\bar{V}_0 > C > \bar{V}_1^w(\delta) > V_1 \mu_1$, then Part 1 of Proposition 4 and Part 1 of Lemma 5 apply, so $\lambda_0^s$ satisfies (40) and $\lambda_0^{w*}$ satisfies (76). We have $\lambda_0^{w*} < \lambda_0^s$ and $\Pi^{w*}(\lambda_0^{w*}) < \Pi^s(\lambda_0^s)$, because $\Pi^{w*}(\lambda_0)$ and $\Pi^s(\lambda_0)$ are strictly concave and

$$\Pi^{w*}(\lambda_0) = \frac{1}{1 + \delta a} s_0(\bar{V}_0 - C) - S'(\lambda_0) < s_0(\bar{V}_0 - C) - S'(\lambda_0) = \Pi^s(\lambda_0).$$