Customer Acquisition, Retention, and Service Quality for a Call Center: Optimal Promotions, Priorities, and Staffing

Philipp Afèche • Mojtaba Araghi • Opher Baron
Rotman School of Management, University of Toronto

This Version: November 2012

We study the problem of maximizing profits for an inbound call center with abandonment by controlling customer acquisition, retention, and service quality via promotions, 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 characterize the optimal controls analytically based on a deterministic fluid approximation and show via simulation that these prescriptions yield near-optimal performance for the underlying stochastic model. This tractable modeling framework can be extended to further problems of joint customer relationship and call center management. Third, we derive three metrics which play a key role in call center 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. Fourth, we generate novel guidelines on managing a call center based on these metrics, the cost of promotions, and the capacity cost per call.

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 70-80% of a firm's interactions with its customers occur through call centers (Feinberg et al. 2002, Anton et al. 2004), and 92% of customers base their opinion of a company on their call center service experiences (Anton et al. 2004). More importantly, the call center service experience can have a dramatic impact on customer satisfaction and retention. Poor service is cited by 50% of customers as the 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 its 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."

This paper provides a starting point for building such understanding. To our knowledge, this is the first paper to consider the impact of queueing-related service quality on customer retention and long-term customer value. The standard approach in the call center literature has been to model a

firm's *customer base* as *independent of past interactions*. We model new and base (repeat) customers and study the problem of maximizing profits by controlling customer acquisition, retention, and service quality via promotions, priorities to new or base customers, and staffing.

This paper makes four contributions. First, we propose what seems to be the first call center model in which the customer base depends on past demand and queueing-related service quality. Second, we show via simulation that a deterministic fluid analysis yields near-optimal performance for the underlying stochastic model. This modeling framework can be extended to further problems of joint CRM and call center management. Third, we derive metrics which play a key role in call center decisions, and which link customer and financial parameters with operational service quality. Fourth, we generate novel results on how to manage a call center based on these metrics, the cost of promotions, and the capacity cost per call. We elaborate on these contributions in turn.

First, we develop a novel marketing-operations model of an inbound call center, which links elements of CRM with priority and staffing decisions. The model captures new customer arrivals in response to promotions, their conversion to base customers, the evolution and calls of the customer base, and call-related and call-independent profits and costs. Customers contribute to queueing and are impatient, which leads to abandonment and adversely affects customer retention. A notable feature of this model is that it can be tailored to a range of businesses that rely on a call center, such as credit card companies, phone service providers, or catalog marketing companies. In contrast to the marketing literature on CRM, the key novelty of our model is that customer flows and the customer base depend on the service quality, i.e., the probability of getting served, and in turn on customer acquisition, priorities, and capacity. In contrast to the call center literature, the key novelty of our model is that the customer base depends on past demand and service.

Second, we characterize the optimal controls analytically based on a deterministic fluid model approximation of the underlying stochastic queueing model, which is difficult to analyze directly. We validate these analytical prescriptions through a simulation study, which shows that they yield near-optimal performance for the stochastic system it approximates, with maximum profit losses below 1%. These results suggest that the main insights and guidelines based on the fluid model apply to the stochastic model as well. More generally, these results suggest that our modeling and analytical approach may prove quite effective in tackling further problems in this important area.

Third, we derive three metrics which are the basis for call center decisions, the expected *customer lifetime value* (CLV) of a base customer and the expected *one-time serving value* (OTV) of a new and base customer. A key feature of these metrics is that they depend not only on customer behavior and financial parameters, but also on operations through the service quality, which reflects the system load and the priority policy. In particular, unlike standard CLV metrics in the marketing literature, the CLV in our model reflects the impact of abandonment on retention. The OTV metrics also capture interaction effects between customers' service-related propensity to join and leave the customer base, and their call frequency while in the customer base.

Finally, we generate novel results on how to manage a call center. We show that it is optimal to prioritize the customers with the higher OTV. A notable feature of this policy is that it accounts for the financial impact of customers' future calls, in contrast to standard priority policies such as the

$c\mu$ rule. Next, for situations where the capacity is fixed, e.g., due to lags in hiring or training , we characterize the jointly optimal promotion and priority policy as a function of the promotion cost, the CLV and OTVs, and the capacity level. We further show how the jointly optimal promotion, priority and staffing policy depend on the CLV, the OTVs, and the capacity cost. Under the optimal policy, the most striking operating regime arises if new customers have the higher OTV, e.g., due to prohibitive switching costs for base customers, and capacity is relatively expensive. Under these conditions it is optimal to prioritize new customers and to *overload* the system. In this regime the primary goal of the call center is to serve and acquire new customers to grow the customer base, whereas *base customers* receive *deliberately poor service*. This result lends some theoretical support for the anecdotal evidence that locked-in customers of firms such us mobile phone service providers commonly experience long waiting times when contacting the call center.

The plan of this paper is as follows. In §2 we review the related literature. In §3 we specify the stochastic queueing model and the approximating deterministic fluid model, and we formulate the firm's profit maximization problem. In §4, we derive the CLV and OTV metrics and characterize the fluid model prescriptions on the optimal priority policy, promotion spending, and capacity level. In §5 we present simulation results that evaluate the performance of the fluid model prescriptions of §4 against simulation-based optimization results for the stochastic system described in §3. Our concluding remarks are in §6. All proofs are in the Online Supplement.

2 Literature Review

This paper is at the intersection of research streams on advertising, CRM, and call center management. We relate our work first to these literatures, and then to operations papers outside the call center context which also consider demand as a function of past service, as we do in this paper.

There is a vast literature on advertising. We refer to Feichtinger et al. (1994), Hanssens et al. (2001), and Bagwell (2007) for surveys. In contrast to our study, the overwhelming majority of these papers ignore the firms' supply constraints in fulfilling the demand generated by advertising. A number of papers consider advertising under supply constraints in different settings. Focusing on physical goods, Sethi and Zhang (1995) study joint advertising and production control, and Olsen and Parker (2008) study joint advertising and inventory control. Focusing on services, Horstmann and Moorthy (2003) study the relationship between advertising, capacity, and quality in a competitive market; in their model, unlike in ours, the quality attribute is independent of utilization.

CRM and models of CLV and related customer metrics are of growing importance in marketing. We refer to Rust and Chung (2006), Gupta and Lehmann (2008), and Reinartz and Venkatesan (2008) for surveys. Blattberg and Deighton (1996) develop a tool to optimize the (static) mix of acquisition and retention spending. Ho et al. (2006) derive static optimal spending policies in customer satisfaction in a model where customers' purchase rates, spending amounts and retention depend on their satisfaction from their last purchase. Several papers study the design of dynamic policies, focusing on marketing instruments such as direct mail (cf. Bitran and Mondschein 1996), pricing (Lewis 2005), cross-selling (Günes et al. 2010), and service effort (Aflaki and Popescu 2012).

In contrast to our setup, the CRM literature ignores supply constraints and the interaction between capacity, demand, and service quality. To our knowledge, Pfeifer and Ovchinnikov (2011) and Ovchinnikov et. al. (2012) are the only papers that consider a capacity constraint. They study its impact on the value of an incremental customer and on the optimal spending policy for acquisition and retention. In contrast to our model, theirs do not consider the effect of queueing and service quality on customer acquisition, the CLV, and retention.

The call center literature is extensive and growing. We refer to Gans et al. (2003), Akşin et al. (2007), and Green et al. (2007) for surveys. The bulk of these papers focus on operational controls, i.e., staffing and allocation policies to serve an *exogenous arrival process* of call center requests to the system, and they often consider *endogenous abandonment*. Some papers consider exogenous arrivals of *initial* requests but model policies to manage *endogenous retrials* before service due to congestion (e.g., Armony and Maglaras 2004), or after service due to poor service quality on earlier calls (e.g., de Véricourt and Zhou 2005). A growing number of papers study marketing and operational controls; they consider exogenous arrivals of *potential* requests but model some aspects of the *actual requests* as *endogenous*. This framework characterizes the study of cross-selling which increases service times to boost revenue. (Akşin and Harker 1999 started this stream; Akşin et al. 2007 review it; Gurvich et al. 2009 jointly consider staffing and cross-selling; Debo et al. 2008 study a service time-revenue tradeoff outside the call center setting.) This framework is also standard in the stream on pricing, scheduling, and delay information policies for queueing systems in general, rather than call centers in particular (cf. Hassin and Haviv 2003). Randhawa and Kumar (2008) model both initial requests and retrials as endogenous, based on the price and service quality.

In contrast to this paper, these call center and queueing research streams ignore the impact of service quality on *customer retention*, i.e., a firm's *customer base is independent of past interactions*. In a parallel effort, Farzan et al. (2012) do consider repeat purchases that depend on past service quality; however, in contrast to our model, they model service quality by a parameter that is independent of queueing. In the marketing literature, Sun and Li (2011) empirically estimate how the retention of customers depends on their allocation to onshore vs. offshore call centers, including on waiting and service time. In contrast to our paper, theirs does not model capacity constraints and the link to waiting time. However, their numerical results underscore the value of considering customer retention and CLV in call center policies. Specifically, they numerically solve a stochastic dynamic program that matches customers to service centers to maximize long term profit, and show via simulation that considering customer retention and CLV can significantly improve performance.

Schwartz (1966) seems to be the first to consider how past service levels affect demand, focusing on inventory availability. The operations literature has seen a growing interest over the last decade in studying how past service levels affect demand and how to manage operations in such settings. These papers consider operations outside the call center context. As such their models are fundamentally different from ours. Gans (2002) and Bitran et al. (2008) consider a general notion of service quality and do not model capacity constraints. Gans (2002) considers oligopoly suppliers that compete on static service quality levels and models customers who switch among them in Bayesian fashion based on their service history. Bitran et al. (2008) model a price- and

quality-setting monopoly and the evolution of its customer base depending on satisfaction levels and the number of past interactions. 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 competing firms with customers that 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 attract new/reacquire dissatisfied customers. The authors study the optimality of base-stock policies for the monopoly and the duopoly case. Adelman and Mersereau (2012) study how a supplier of a physical good should dynamically allocate its fixed capacity among a fixed portfolio of heterogeneous customers who never defect, but whose stochastic demands depend on goodwill derived from past fill rates.

3 Model and Problem Formulation

Consider a firm that serves two types of customers through its inbound call center. Base customers are part of the firm's customer base and repeatedly interact with the call center. New customers are first-time callers who may turn into base customers. Both customer types are impatient. Figure 1 depicts the customer flow through the system, showing the flow of new customers by dashed lines and the flow of base customers by solid lines. We describe the model of this system in two steps. In §3.1 we specify a conventional exact stochastic queueing model of the call center that captures variability in inter-arrival, service and abandonment times. In §3.2 we describe the approximating fluid model. Our analytical results in §4 are based on this fluid model.

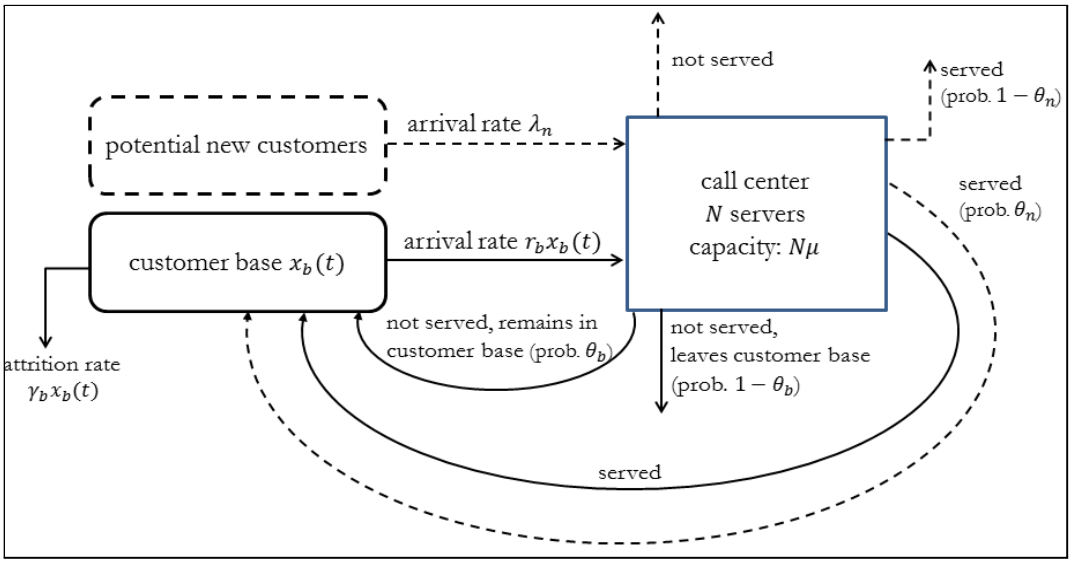

Figure 1: Flow of New and Base Customers Through the System.

3.1 The Stochastic Queueing Model

We model the call center as an N-server system. Service times are i.i.d. with mean $1/\mu$, so $N\mu$ is the system capacity. Calls arrive as detailed below. Customers wait in queue if the system is

busy upon arrival, but they are impatient. Abandonment times are independent and exponentially distributed with mean $1/\tau$, so $\tau > 0$ is the abandonment rate. Table 1 summarizes the notation.

We consider the system in steady-state under three stationary controls: the staffing policy sets the number of servers at a cost of C per server per unit time, the promotion policy controls the new customer call arrival rate, and the priority policy prioritizes new or base customer calls.

New customer calls arrive to the system following a stationary Poisson process with rate λ_n. The new customer call arrival rate depends on the firm's advertising spending (we use the terms advertising and promotion interchangeably). Let $S(\lambda_n)$ denote the advertising spending rate per unit time as a function of the new customer arrival rate it generates. We assume that the response

System Parameters	
N	Number of servers
μ	Service rate per server
λ_n	New customer call arrival rate
r_b	Call arrival rate per base customer
τ	Call abandonment rate
θ_n	P(new customer joins customer base after service)
θ_b	P(base customer remains in customer base after abandoning)
γ_b	Attrition rate per base customer (call-independent)
Economic Parameters	
p_n, p_b	Profit per served call of new, base customer
c_n, c_b	Cost per abandoned call of new, base customer
R	Profit rate per base customer (call-independent)
C	Cost rate per server
α, β	Parameters of advertising cost function $S(\lambda_n) = \alpha(\lambda_n)^\beta$
Steady-State Performance Measures	
x_b	Average number of base customers
q_n, q_b	Service probability of new, base customer calls
Π	Call center profit rate

Table 1: Summary of Notation.

of new customers to advertising spending follows the law of diminishing returns (Simon and Arndt 1980), so $S(\lambda_n)$ is strictly increasing and strictly convex in λ_n. For analytical convenience we assume that S is twice continuously differentiable and $S'(0) = 0$. Given the convexity of $S(\lambda_n)$ the assumption of a stationary advertising policy is quite plausible. We present our analytical results for a general increasing convex advertising cost function. We illustrate these results for the commonly assumed power model (cf. Hanssens et al. 2001), i.e.,

$$S(\lambda_n) = \alpha\lambda_n^\beta, \quad \alpha > 0, \beta > 1.$$

The constant α is a scale factor. The constant β is the inverse of the customers' response elasticity to the advertisement level, where $\beta > 1$ captures diminishing returns to advertising expenditures.

Let q_b and q_n denote the steady-state probability that base and new customer calls are served, respectively. We subsequently refer to each of these measures simply as a "service probability". These service probabilities depend on the system parameters and controls as discussed below.

On average the firm generates a profit of p_n per new customer call it serves and incurs a cost of $c_n \geq 0$ per new customer call it loses due to abandonment. Usually there is no direct penalty for not serving a new customer, so c_n can be interpreted as the loss of goodwill.

The following flows determine the evolution of the customer base. A new customer who receives service joins the customer base with probability $\theta_n > 0$, so new customers turn into base customers at an average rate of $\lambda_n q_n \theta_n$ per unit time. The times between successive calls of a base customer are independent and exponentially distributed with mean $1/r_b$, so $r_b \geq 0$ is the average call rate per base customer per unit time. We assume that $1/r_b >> 1/\mu, 1/\tau$, i.e. the mean time between the calls of a given customer is much larger than the mean service and abandonment times. A base customer who abandons the queue remains in the customer base with probability θ_b, but immediately terminates her relation with the company and leaves the customer base with probability $1 - \theta_b$. We do not explicitly model competition, but the parameter θ_b partially captures its effect. Other things equal, θ_b is lower in a competitive market than in a monopolistic one. The customer base is also subject to attrition due to service-independent reasons, such as customers moving away. The lifetimes of base customers in the absence of abandonment are independent and exponentially distributed with mean $1/\gamma_b$, so $\gamma_b > 0$ is the average call-independent attrition rate of a base customer.

Let x_b denote the long-run average number of base customers in steady-state; we also call x_b simply the average customer base. The system is stable since customer impatience ensures a stable queue at the call center, and the average customer base is finite since $\lambda_n < \infty$ and $\gamma_b > 0$.

We model two potential profit streams from base customers. First, the firm may generate a call-independent profit at an average rate of $R \geq 0$ per unit time per base customer, which captures monetary flows that are independent of call center interactions, such as monthly subscription and usage fees in the case of a mobile phone service provider. Second, base customers may also call with a purchase or a service request. In steady-state the average arrival rate of base customer calls is $x_b r_b$. On average the firm generates a profit of p_b per base customer call it serves and incurs a cost of $c_b \geq 0$ per base customer call it loses due to abandonment.

To summarize, we model the system as a two-station queueing network with two types of impatient customers and state-dependent routing. The call center itself is a $G/G/N + M$ system with service rate μ per server. Between successive call center visits, base customers enter an orbit that operates like a $G/M/\infty$ system with service rate $r_b + \gamma_b$. The new customer arrival process is Markovian and state-independent. The base customer arrival processes to the call center and to the orbit depend on the new customer arrival process, the capacity and the priority policy.

Let Π denote the firm's average profit rate in steady-state, which is given by

$$\Pi := \lambda_n \left(p_n q_n - c_n \left(1 - q_n\right)\right) + x_b \left(R + r_b \left[p_b q_b - c_b \left(1 - q_b\right)\right]\right) - S\left(\lambda_n\right) - CN, \tag{1}$$

where the first product is the profit rate from new customer calls, the second product is the profit rate from base customers (both call-independent and call-dependent), the third term is the

advertising cost rate, and the last term is the staffing cost rate. The firm aims to maximize its profit rate by choosing the number of servers N, the new customer arrival rate λ_n and the priority policy which affects the service probabilities q_b and q_n.

The profit rate (1) depends on three stationary performance measures, the average customer base x_b and the service probabilities q_b and q_n. The state-dependent nature of customer flows and feedbacks through the system make it difficult to analyze these measures for the stochastic model, even under the Markovian assumptions made above. We therefore approximate the stochastic model by a corresponding deterministic fluid model that we describe in §3.2.

Examples. By appropriately choosing key parameter values, the model can be tailored to a range of call centers. We discuss two characteristic cases which are summarized in Table 2.

Revenue Generation		Call-Independent		Call-Dependent
Type of Business		Credit Card	Phone Service	Catalog Marketing
Profit per served call	p_n, p_b	low or negative	low or negative	high
Profit rate per base customer (call-independent)	R	high	high	zero
P(new customer joins customer base after service)	θ_n	high	moderate	moderate
P(base customer remains in customer base after abandoning)	θ_b	moderate	very high	low
Attrition rate per base customer (call-independent)	γ_b	low	low	moderate

Table 2: Tailoring the Model to Call Center Characteristics: Examples.

Significant call-independent revenue generation. Consider a credit card company receiving calls from card holders and from potential new customers that are attracted by advertisements. The company offers a range of incentives and rewards to potential new customers to encourage them to apply for a credit card, so the profit of their first call p_n is usually near zero or even negative. The per-call profit of existing customers, p_b, is likely also negative, as card holders typically call with service requests, e.g., to redeem points or report lost cards, rather than to buy additional revenue-generating products, e.g., insurance. However, every month the credit card company receives on average a potentially significant call-independent profit R per card holder, which includes interest rate and subscription fee payments from card holders, and transaction fee payments from merchants where transactions took place. The probability θ_n that a potential new customer, who calls with a credit card application in response to a promotion, is approved and joins the customer base may be quite high. The probability θ_b that an existing customer who abandons the line remains in the customer base may be high or low, depending on her overall satisfaction, access to alternate credit lines, and the cost of switching credit card providers. Finally, most card holders keep their cards for a long period of time if they receive reasonable service, so that γ_b may be quite low.

The call center of a mobile phone service provider is similar to that of a credit card company in that, here too, much of the revenue is recurrent and independent of call center interactions. The main source of profit is the monthly subscription fee R, and additional communication charges.

Like a credit card company, such a business is also likely to experience low or even negative per-call profits p_n and p_b. However, unlike a credit card company, a mobile phone service provider may enjoy a significantly larger value of θ_b, because leaving the customer base is often subject to significant contract termination penalties and other switching costs.

Significant call-dependent revenue generation. In contrast to the above examples, the call center of a catalog marketing business may enjoy relatively significant per-call profits p_n and p_b, driven by merchandise sales, but generate little or no call-independent recurring revenues, i.e., R is small. Further, repeat customers may spend more per call than new customers, i.e., $p_b > p_n$. Since its customers are not subject to the same switching costs as those of credit card or phone service providers, a catalog marketing business likely faces a lower value of θ_b and a higher value of γ_b.

3.2 The Approximating Fluid Model

In this section we characterize the steady-state average customer base x_b and service probabilities q_b and q_n for the fluid model depending on the system load and the priority policy. One obvious caveat of the deterministic fluid model is that it does not account for queueing effects and customer impatience in evaluating the steady-state service probabilities. Specifically, in the stochastic system, customers may abandon even if there is enough capacity to serve them eventually. In contrast, in the fluid model all customers are served if there is enough capacity. However, the great advantage of the fluid model is its analytical tractability. It yields clear results on the optimal decisions, as shown in §4. Moreover, our simulation results in §5 show that the optimal decisions that the fluid model prescribes yield near-optimal performance for the stochastic system it approximates.

In steady-state, the size of the customer base must be constant in time:

$$x_b'(t) = \lambda_n q_n \theta_n - x_b(t)\left[\gamma_b + r_b(1-q_b)(1-\theta_b)\right] = 0, \tag{2}$$

where $\lambda_n q_n \theta_n$ is rate at which new customers join the customer base, and the second term in (2) is the customer base decay rate which is proportional to the size of the customer base. As discussed above, the departure rate of any base customer has two components, the service-independent attrition rate γ_b and the call-dependent term $r_b(1-q_b)(1-\theta_b)$, which is the product of a base customer's calling rate r_b, abandonment probability $1-q_b$, and probability of leaving the customer base after abandonment $1-\theta_b$. Solving (2) yields the steady-state average number of base customers

$$x_b := \frac{\lambda_n q_n \theta_n}{\gamma_b + r_b(1-q_b)(1-\theta_b)}. \tag{3}$$

The numerator in (3) is the inflow rate of new customers, the denominator the departure rate from the customer base, and $1/[\gamma_b + r_b(1-q_b)(1-\theta_b)]$ is the mean sojourn time in the customer base.

Let ρ be the system's load factor:

$$\rho := \frac{\lambda_n + x_b r_b}{N\mu}. \tag{4}$$

We call the system underloaded if $\rho \leq 1$, balanced if $\rho = 1$, and overloaded if $\rho > 1$. Similarly, let

$$\rho_n := \frac{\lambda_n}{N\mu}$$

be the system's new customer load factor. Using (3) and (4), we determine (x_b, q_b, q_n) and substitute into (1) to obtain the profit rate Π as a function of the system load and the priority policy.

Underloaded System ($\rho \leq 1$). If the system is underloaded, all customers are served in the fluid model regardless of the priority policy, so $q_b = q_n = 1$. It follows from (3) that

$$x_b = \frac{\lambda_n \theta_n}{\gamma_b}. \tag{5}$$

All new customers are served, a fraction θ_n turn into base customers, and they leave only for service-independent reasons with rate γ_b. By (4) and (5) the system is underloaded if and only if

$$\lambda_n \left(1 + \theta_n \frac{r_b}{\gamma_b}\right) \leq N\mu, \tag{6}$$

so $\lambda_n \left(1 + \theta_n r_b/\gamma_b\right)/N\mu$ is the maximum system load factor for given λ_n.

Substituting for x_b from (5) and $q_b = q_n = 1$ into (1) yields the steady-state profit rate:

$$\Pi = \lambda_n p_n + \frac{\lambda_n \theta_n}{\gamma_b}\left(R + r_b p_b\right) - S\left(\lambda_n\right) - CN. \tag{7}$$

Overloaded System ($\rho > 1$): Prioritize Base Customers. If base customers are prioritized in an overloaded system, new customers only get access to the residual capacity $\left(N\mu - x_b r_b\right)^+$, so

$$q_n = \frac{\left(N\mu - x_b r_b\right)^+}{\lambda_n} < 1,$$

where the inequality follows from (4) because the system is overloaded.

Remark 1*. Under any priority policy, the new customers' steady-state service probability $q_n > 0$.* To see why this must hold, note that if $q_n = 0$, no one joins the customer base and $x_b = 0$ by (3), so that all capacity is available for new customers. When base customers are prioritized, the customer base must therefore equilibrate at a level that leaves *some* (but insufficient) residual capacity for new customers, which also guarantees that all base customers are being served ($q_b = 1$):

$$0 < q_n = \frac{N\mu - x_b r_b}{\lambda_n} < 1. \tag{8}$$

Combined with (3), it follows from (8) that

$$x_b = \frac{N\mu\theta_n}{\gamma_b + \theta_n r_b} \text{ and } q_n = \frac{1}{\lambda_n}\frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}.$$

Substituting (x_b, q_b, q_n) into (1) yields for an overloaded system that prioritizes base customers:

$$\Pi = p_n \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b} - c_n\left(\lambda_n - \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}\right) + \frac{N\mu\theta_n}{\gamma_b + \theta_n r_b}\left(R + r_b p_b\right) - S\left(\lambda_n\right) - CN. \tag{9}$$

Overloaded System ($\rho > 1$): Prioritize New Customers. If new customers are prioritized, base customers are only served by the residual capacity $N\mu\left(1 - \rho_n\right)^+$, so

$$q_n = \min\left(\frac{1}{\rho_n}, 1\right) \text{ and } 0 \leq q_b = \frac{N\mu\left(1 - \rho_n\right)^+}{x_b r_b} < 1, \tag{10}$$

where the strict inequality for q_b follows from (4) because the system is overloaded. We consider in turn the two possible cases, $\rho_n < 1$ and $\rho_n \geq 1$.

Some residual capacity for base customers ($\rho_n < 1$). In this case $q_n = 1$ and $q_b > 0$ by (10). By (3) and (10) the average customer base in steady-state is

$$x_b = \frac{\theta_n \lambda_n + (1-\theta_b)(N\mu - \lambda_n)}{\gamma_b + r_b(1-\theta_b)},$$

where $x_b r_b q_b = N\mu - \lambda_n$ is the base customer throughput. Substituting (x_b, q_b, q_n) into (1) yields for the profit of an overloaded system that prioritizes new customers and serves some base customers:

$$\Pi = \lambda_n p_n + (p_b + c_b)(N\mu - \lambda_n) + \frac{\theta_n \lambda_n + (1-\theta_b)(N\mu - \lambda_n)}{\gamma_b + r_b(1-\theta_b)}(R - r_b c_b) - S(\lambda_n) - CN. \quad (11)$$

No residual capacity for base customers ($\rho_n \geq 1$). In this case $q_n \leq 1$ and $q_b = 0$ by (10). By (3) and (10) the average customer base in steady-state is

$$x_b = \frac{N\mu\theta_n}{\gamma_b + r_b(1-\theta_b)}.$$

Substituting (x_b, q_b, q_n) into (1) yields for this regime:

$$\Pi = N\mu p_n - (\lambda_n - N\mu)c_n + \frac{N\mu\theta_n}{\gamma_b + r_b(1-\theta_b)}(R - r_b c_b) - \alpha(\lambda_n)^\beta - CN. \quad (12)$$

4 Optimal Priority Policy, Promotion Level, and Staffing

In this section we solve a sequence of three increasingly general optimization problems for the fluid model. The solution of each problem serves as a building block for the next problem in the sequence. Before proceeding with the analysis, in §4.1 we derive three customer value metrics which play an important role in the structure of the optimal decisions. These metrics are novel in that they depend on the call center service quality. In §4.2 we consider the case in which the manager only controls the priority policy, whereas the new customer arrival rate and the call center capacity are fixed. In §4.3 we characterize the jointly optimal priority policy and promotion spending, taking the capacity as fixed. In §4.4 we solve the optimization problem over all three controls. In §4.5 we study the sensitivity of the optimal decisions to changes in the parameter values.

4.1 Customer Value Metrics and Service Quality

Let $L(q_b)$ denote the mean base *customer lifetime value (CLV)*, i.e., the total profit that she generates during her sojourn in the customer base, as a function of the service probability q_b:

$$L(q_b) := \frac{R + r_b(p_b q_b - c_b(1-q_b))}{\gamma_b + r_b(1-q_b)(1-\theta_b)}. \quad (13)$$

The CLV of a base customer is the product of her profit rate per unit time, the numerator in (13), by her average sojourn time in the customer base.

Let V_b denote the mean *one-time service value (OTV)* of a base customer, which measures the value of serving her current call but not any of her future calls:

$$V_b := p_b + c_b + (1 - \theta_b) L(0). \tag{14}$$

Serving a base customer's current call, but not any of her future calls, yields a profit $p_b + L(0)$, where p_b is the immediate profit, and $L(0)$ is the CLV given a zero service probability for this base customer. Not serving a base customer's call yields $-c_b + \theta_b L(0)$, where the first term captures the immediate cost and the second is her CLV given a zero service probability for this base customer. The difference between these two profits yields (14).

The CLV and base customer OTV satisfy the following intuitive relationship:

$$L(1) = L(0) + \frac{r_b}{\gamma_b} V_b, \tag{15}$$

where r_b/γ_b is the mean number of calls during a base customer's lifetime if all her calls are served.

Similarly, let V_n denote the mean OTV of a new customer, i.e., the value of serving a new customer's current call, but not any of her future calls:

$$V_n := p_n + c_n + \theta_n L(0). \tag{16}$$

Serving a new customer yields instant profit p_n, and with probability θ_n turns that customer into a base customer with lifetime value $L(0)$. Not serving a new customer results in a penalty c_n.

Remark 2. In cases where the firm controls the new customer arrival rate, as in §4.3-§4.4, the new customer OTV is $V_n - c_n$, because the firm does not attract new customers it does not intend to serve, and therefore does not incur the abandonment cost on such calls.

4.2 Optimal Priority Policy for Fixed Promotion Level and Staffing

Consider the case where the number of servers and the new customer arrival rate are fixed. This captures situations where staffing and/or advertising may not be at their optimal levels, e.g., due to hiring lead times, time lags between advertising and demand response, or poor coordination between marketing and operations. The remaining control is the priority policy to allocate capacity.

Proposition 1 *Fix the new customer arrival rate λ_n and the number of servers N. It is optimal to prioritize new customer calls if*

$$V_n = p_n + c_n + \theta_n L(0) \geq V_b = p_b + c_b + (1 - \theta_b) L(0),$$

and it is optimal to prioritize base customer calls otherwise.

A novel feature of the optimal priority policy specified in Proposition 1 is that it explicitly considers the financial impact of customers' future calls, in contrast to standard priority policies in the literature, such as the $c\mu$ rule. Figure 2 illustrates how the optimal priority policy depends on the system load and on the difference between the OTVs of new and base customers, $V_n - V_b$.

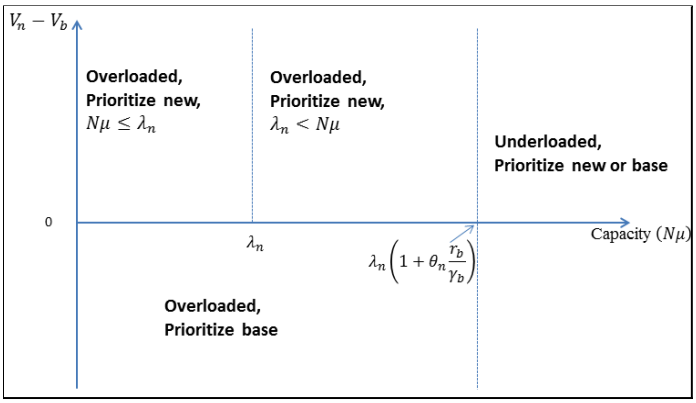

Figure 2: Optimal Priority Policy as Function of System Load and New vs. Base Customer OTVs.

In the region "Overloaded, Prioritize base", the OTV of a base customer exceeds that of a new customer, i.e., $V_b > V_n$. The system serves all base customers and a fraction of new customers (see Remark 1 in §3.2). The condition $V_b > V_n$ applies, for example, to a catalog marketing company (see Table 2) with base customers who generate a significantly larger profit per call vs. new customers (so $p_b > p_n$), and who are prone to leave the customer base if they are not served (so θ_b is low).

Conversely, if $V_n > V_b$, it is optimal to prioritize new customers, serving no base customers if $N\mu \leq \lambda_n$, and only a fraction of them if $\lambda_n < N\mu < \lambda_n (1 + \theta_n r_b/\gamma_b)$. The condition $V_n > V_b$ applies, for example, to a popular mobile phone service provider with potential new customers who are somewhat likely to join the customer base upon being served by the call center (so θ_n is moderate), and existing customers who do not easily leave the customer base (so θ_b is very high).

Remark 3. The profit depends on the priority policy only in conditions that result in throughput loss. In the deterministic fluid model, the system loses throughput if and only if it is overloaded; as Figure 2 shows the profit is independent of the priority policy in an underloaded system. An underloaded *stochastic* system, however, may experience throughput loss, due to queueing and abandonment. In stochastic systems with throughput loss, prioritizing customers with the higher OTV, in line with Proposition 1, improves profits even if $\rho < 1$, as our simulation results in §5 show.

4.3 Jointly Optimal Priority and Promotion Policy for Fixed Staffing Level

Consider situations where the staffing is fixed but the manager controls the priority policy and the arrival rate of new customers through the advertising budget. This setting is common because promotion levels are typically more adjustable in the short term compared to capacity levels, e.g., due to lead times in hiring and training, or because of inflexible outsourcing arrangements.

The optimal advertising policy balances the value and cost of attracting a new customer call. We say *net revenue* for the profit before advertising and staffing costs. From (1) the net revenue is

$$\lambda_n \left(p_n q_n - c_n \left(1 - q_n\right)\right) + x_b \left(R + r_b \left[p_b q_b - c_b \left(1 - q_b\right)\right]\right), \tag{17}$$

where the average number of base customers x_b and the service probabilities q_n and q_b depend on the system load and the priority policy as specified in §3.2. The value of an additional new

customer call is given by the marginal net revenue with respect to λ_n. We first discuss how the marginal net revenue depends on the system load and the priority policy (see Figure 3). We then specify the jointly optimal promotion and priority policy.

Underloaded System ($\rho \leq 1$). The profit rate (7) of an underloaded system is independent of the priority policy since all calls are served. The marginal net revenue satisfies

$$p_n + \theta_n L(1) = V_n - c_n + V_b \theta_n \frac{r_b}{\gamma_b}. \tag{18}$$

The LHS follows from (7) and (13); p_n is the profit of the new customer's first call, $\theta_n L(1)$ is the probability that she joins the customer base multiplied by her CLV if all her calls are served. The RHS follows from (15)-(16); $V_n - c_n$ is the new customer OTV which includes call-independent but excludes call-related future profits, $V_b \theta_n r_b / \gamma_b$ is the expected value of serving her future calls. To rule out the trivial case where it is unprofitable to attract new customers, we assume the following.

Assumption 1. $p_n + \theta_n L(1) > 0$.

Overloaded System ($\rho > 1$). Attracting new customers can only have positive value if more new customer calls can be served, which depends on the priority policy and the system load.

No additional new customer calls can be served if the system is overloaded and prioritizes base customers (see (9)), or if it prioritizes new customers and is overloaded with their calls (see (12)). In either case the marginal net revenue is $-c_n$, the abandonment penalty on the new customer call.

Additional new customer calls *can* be served in an overloaded system that prioritizes new customers, if their calls do not exhaust the capacity ($\rho_n < 1 < \rho$). In this regime the total throughput is constant, but the throughput of new customer calls and the size of the customer base increase in the new customer arrival rate. The marginal net revenue of a new customer satisfies

$$p_n + \theta_n L(0) - V_b = V_n - c_n - V_b. \tag{19}$$

The LHS follows from (11) and (13)-(14), the RHS from the definition of V_n in (16). By (19), attracting a new customer call to an overloaded system is profitable if $V_n - c_n > V_b$, i.e., her OTV exceeds the OTV of the base customer call it displaces. (By Remark 2, the new customer OTV excludes the abandonment cost if the firm controls the new customer arrival rate.)

By (18)-(19) the marginal value of a new customer call decreases in the system load, i.e., the net revenue function is concave in λ_n. Figure 3 summarizes this discussion.

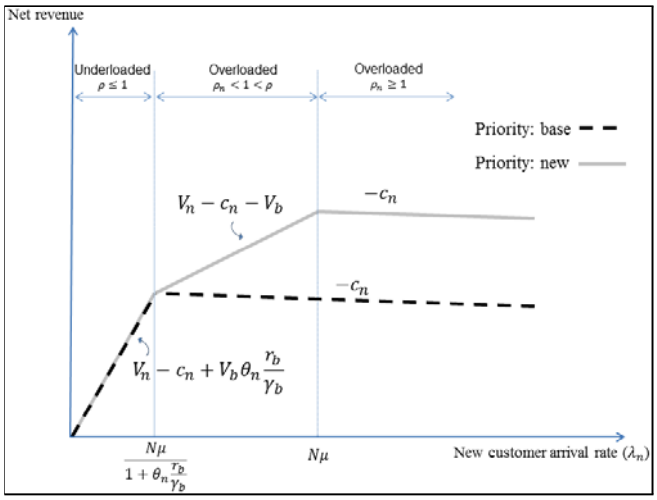

Figure 3: Net Revenue as Function of New Customer Arrival Rate and Priority Policy ($V_n - c_n > V_b$).

Jointly Optimal Promotion and Priority Policy. Let λ_n^* denote the optimal new customer arrival rate. Let $\overline{\lambda}_n$ be the new customer arrival rate at which the marginal net revenue in an underloaded system equals the marginal advertising cost. By (18) this arrival rate satisfies

$$V_n - c_n + V_b\theta_n \frac{r_b}{\gamma_b} = S'\left(\overline{\lambda}_n\right). \tag{20}$$

It is optimal to run an underloaded system if $\overline{\lambda}_n \leq N\mu/\left(1+\theta_n r_b/\gamma_b\right)$, in which case $\lambda_n^* = \overline{\lambda}_n$. Otherwise, as discussed above, it may be profitable to run an overloaded system that prioritizes new customers, but only if $V_n - c_n > V_b$. Let $\underline{\lambda}_n$ be the new customer arrival rate at which the marginal net revenue under this overloaded regime equals the marginal advertising cost. By (19)

$$V_n - c_n - V_b = S'\left(\underline{\lambda}_n\right), \tag{21}$$

where $\underline{\lambda}_n < \overline{\lambda}_n$ since the net revenue function is concave and the promotion cost is strictly convex.

In particular, if $S\left(\lambda_n\right) = \alpha\left(\lambda_n\right)^{\beta}$, then

$$\underline{\lambda}_n = \left(\frac{V_n - c_n - V_b}{\alpha\beta}\right)^{\frac{1}{\beta-1}} < \overline{\lambda}_n = \left(\frac{V_n - c_n + V_b\theta_n r_b/\gamma_b}{\alpha\beta}\right)^{\frac{1}{\beta-1}}. \tag{22}$$

Proposition 2 specifies the jointly optimal promotion and priority policy.

Proposition 2 *Fix the number of servers N. Under the optimal promotion policy, the optimal priority policy, new customer arrival rate, and system load depend as follows on the OTVs of new and base customer calls, and on the capacity:*

1. *If $V_n - c_n > V_b$ and $N\mu < \underline{\lambda}_n\left(1+\theta_n r_b/\gamma_b\right)$, prioritizing new customers strictly improves profits vs. prioritizing base customers, the system is overloaded, and $\lambda_n^* = \min\left\{\underline{\lambda}_n, N\mu\right\}$.*

   (a) If $N\mu \leq \underline{\lambda}_n$, the system serves all new but no base customers.

   (b) If $\underline{\lambda}_n < N\mu < \underline{\lambda}_n\left(1+\theta_n r_b/\gamma_b\right)$, the system serves all new but only some base customers.

2. Otherwise, profits are independent of the priority policy, the system is underloaded, and

$$\lambda_n^* = \min\left\{\overline{\lambda}_n, N\mu/\left(1+\theta_n r_b/\gamma_b\right)\right\}. \tag{23}$$

The jointly optimal promotion and priority policy gives rise to one of two operating regimes.

By Part 1 of Proposition 2, overloading the system and prioritizing new customers is the *unique* optimal policy, if new customers have the higher OTV (i.e., $V_n - c_n > V_b$) and the marginal promotion cost to overload the system is sufficiently low (i.e., $N\mu < \underline{\lambda}_n\left(1+\theta_n r_b/\gamma_b\right)$). Under these conditions, the primary goal of the call center is to increase the number of base customers, not to serve them, resulting in *deliberately poor service to base customers*. It is optimal to attract and serve so many new customer calls that they displace some or all base customer calls. This overloading boosts the new customer throughput and the customer base without raising total throughput. The new customer load factor may be close to one and the system load factor well above one. The condition $V_n - c_n > V_b$ may hold in the example of a mobile phone service provider outlined in §3.1. Recall from (14) that $V_b = p_b + c_b + (1-\theta_b)L(0)$ and from (16) that $V_n - c_n = p_n + \theta_n L(0)$. If potential new customers are likely to join the customer base upon being served (so θ_n is significant), base customers do not easily leave the customer base, because of significant switching costs (so θ_b is high), and their per-call profit p_b and abandonment cost c_b are small in relation to their call-independent profit rate R, then $V_b \approx 0 < V_n - c_n$. Indeed, existing mobile phone service customers commonly experience long waiting times when contacting the call center with a service request.

By Part 2 of Proposition 2, serving all customers and prioritizing either new or base customers is optimal, if base customers have the higher OTV (i.e., $V_n - c_n \leq V_b$), or if new customers have the higher OTV and the marginal promotion cost to overload the system is high (i.e., $\underline{\lambda}_n\left(1+\theta_n r_b/\gamma_b\right) \leq N\mu$). Under these conditions, it is optimal to attract new customers only to replenish the customer base, while keeping the combined arrival rate of new and base customers at capacity if $N\mu/\left(1+\theta_n r_b/\gamma_b\right) \leq \overline{\lambda}_n$, and below capacity otherwise (see (23)). The condition $V_b > V_n - c_n$ may hold in the example of a catalog marketing company outlined in §3.1, where customers generate a significant profit per call, and base customers are prone to leave the customer base if not served (so θ_b is low). In this case, it is optimal to ensure that all customers are served.

Remark 4. Under both regimes of Proposition 2 it is optimal to serve all new customers and prioritize them, which is intuitive: Spending money to attract new customers is optimal only if they will be served, which is guaranteed by prioritizing them (Part 1) and/or by controlling promotions so the system experiences no throughput loss (Part 2). This logic largely holds for stochastic systems, with the following important qualification to Part 2 of Proposition 2: In stochastic systems it is preferable to prioritize customers with the higher OTV (in line with Proposition 1) regardless of load, since customers may abandon even if the system is underloaded (see Remark 3 in §4.2). In particular, *some abandonment of new customer calls may be optimal*. Unlike in the fluid model, in a stochastic system there is abandonment at load factors (including $\rho = 1$) where throughput is strictly lower than capacity. In this load range, it may be optimal to attract more new customers to boost the throughput at the expense of a higher abandonment rate. Therefore, *if base customers*

have the higher OTV, it may be the unique optimal policy to prioritize them and operate in this load range, even with $\rho > 1$. Under this policy some of the new customer calls, which the firm spends money to attract, are lost. The throughput loss under this policy differs fundamentally, both in rationale and in magnitude, from the one in the overloaded regime in Part 1 of Proposition 2. First, this throughput loss is not deliberate, but rather a side effect of increasing total throughput. Second, since base customers are prioritized, the new customer load factor is well below one, and the load factor ρ is below or above, but in any case close to one.

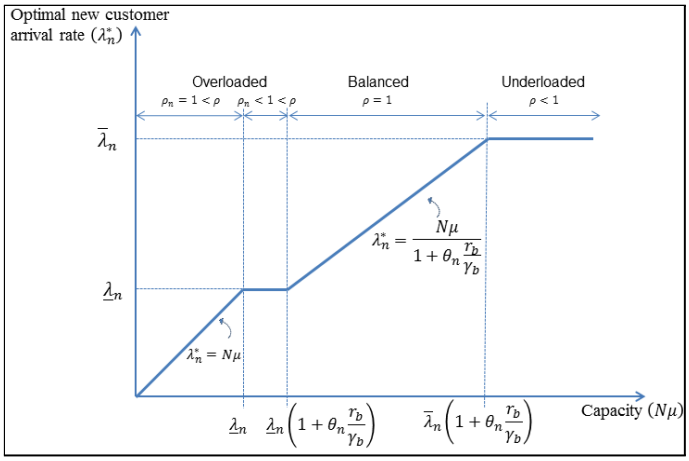

Figure 4: Optimal New Customer Arrival Rate as Function of Capacity ($V_n - c_n > V_b$).

Figure 4 illustrates Proposition 2 for the case $V_n - c_n > V_b$. For $N\mu < \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, the system is overloaded under the optimal new customer arrival rate. For $N\mu \leq \underline{\lambda}_n$, it is profitable to attract new customers to the point where their calls use all the available capacity, so that no base customer are served and $q_b = 0$. As the capacity increases from $N\mu = \underline{\lambda}_n$ to $N\mu = \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, the optimal new customer arrival rate remains fixed at $\lambda_n^* = \underline{\lambda}_n$, because the marginal advertising cost now exceeds the marginal net revenue of a new customer in an overloaded system; the extra capacity is better used to serve base customers who are already in the system. Their service probability increases from $q_b = 0$ at $N\mu = \underline{\lambda}_n$ to $q_b = 1$ at $N\mu = \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, which is the lowest capacity level such that the system is balanced under the optimal new customer arrival rate. At capacity levels close to the threshold $\underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, the marginal net revenue of attracting a new customer to a balanced system exceeds the marginal advertising cost. The optimal new customer arrival rate increases linearly with capacity, and the system remains in balance, up to $N\mu = \overline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, where $\lambda_n^* = \overline{\lambda}_n$. For larger capacity levels $N\mu > \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, the optimal new customer arrival rate is constant in the capacity, so that the system is underloaded.

4.4 Jointly Optimal Priority, Promotion, and Staffing Policy

Finally, we consider the jointly optimal priority policy, promotion spending, and staffing level.

We say *gross profit* for the sum of profit plus capacity cost (or equivalently, the difference of net revenue minus promotion cost). The optimal staffing policy balances the cost of adding an extra

server with the resulting marginal gross profit under the jointly optimal priority and promotion policy specified in Proposition 2. The following analysis focus on a *call* (rather than a server) as the unit of capacity. It compares the marginal gross profit and the capacity cost per call, i.e., C/μ.

Proposition 3 *Under the jointly optimal promotion and staffing policies, the optimal priority policy, new customer arrival rate, system load, and capacity depend as follows on the OTVs of new and base customer calls, and on the capacity cost:*

1. *If $V_n - c_n > V_b$, prioritizing new customers strictly improves profits vs. prioritizing base customers and the system is overloaded, if and only if $V_n - c_n > \frac{C}{\mu} \geq V_b$.*
   (a) *If $V_n - c_n > \frac{C}{\mu} > V_b$, the system serves only new customers, i.e., $\lambda_n^* = N^*\mu$, and*
   $$N^* = \arg\left\{ N \geq 0 : V_n - c_n - S'(N\mu) = \frac{C}{\mu} \right\}. \tag{24}$$
   (b) *If $V_b = \frac{C}{\mu}$, the system serves all new and a fraction of base customer calls, $\lambda_n^* = \underline{\lambda}_n$, and*
   $$N^*\mu \in \left[\underline{\lambda}_n, \underline{\lambda}_n\left(1 + \theta_n r_b/\gamma_b\right)\right]. \tag{25}$$
   (c) *If $\frac{C}{\mu} < V_b$, profits are independent of the priority policy, the system serves all customers, $\lambda_n^* = N^*\mu/\left(1 + \theta_n r_b/\gamma_b\right)$, and*
   $$N^* = \arg\left\{ N \geq 0 : \frac{V_n - c_n + V_b\theta_n r_b/\gamma_b - S'\left(\frac{N\mu}{1+\theta_n r_b/\gamma_b}\right)}{1 + \theta_n r_b/\gamma_b} = \frac{C}{\mu} \right\}. \tag{26}$$
   (d) *Otherwise, it is not profitable to operate: $N^* = \lambda_n^* = 0$.*

2. *If $V_n - c_n \leq V_b$, profits are independent of the priority policy and the system is not overloaded.*
   (a) *If $\left(V_n - c_n + V_b\theta_n r_b/\gamma_b\right)/\left(1 + \theta_n r_b/\gamma_b\right) > \frac{C}{\mu}$, the system serves all customers, $\lambda_n^* = N^*\mu/\left(1 + \theta_n r_b/\gamma_b\right)$, and N^* is given by (26).*
   (b) *Otherwise, it is not profitable to operate: $N^* = \lambda_n^* = 0$.*

By Proposition 3, an overloaded system is optimal if and only if the capacity cost per call is smaller than the OTV of a new customer but (weakly) exceeds that of a base customer.

In Part 1.(a) of Proposition 3, serving base customers is not optimal since the capacity cost per call exceeds their OTV, i.e., $V_b < C/\mu$. This scenario implies the regime in Part 1.(a) of Proposition 2, i.e., $\lambda^* = N\mu \leq \underline{\lambda}_n$; it is optimal to expand capacity only so long as the corresponding optimal new customer arrival rate is at capacity. The marginal gross profit of a call therefore equals the new customer OTV minus the marginal advertising cost at capacity, i.e., $V_n - c_n - S'(N\mu)$. By (24), at the optimal capacity this marginal gross profit equals the cost per call.

In particular, if $S(\lambda_n) = \alpha(\lambda_n)^\beta$ then (24) yields

$$N^*\mu = \left(\frac{V_n - c_n - C/\mu}{\alpha\beta}\right)^{\frac{1}{\beta-1}}.$$

In Part 1.(b) of Proposition 3, the base customer OTV equals the cost of a call, i.e., $C/\mu = V_b$, which implies the capacity scenario in Part 1.(b) of Proposition 2, i.e., $N\mu \in [\underline{\lambda}_n, \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$. In this capacity range, the optimal new customer arrival rate remains fixed at $\lambda_n^* = \underline{\lambda}_n$, because the marginal net revenue of a new customer call is lower than the marginal advertising cost required to attract it. Additional capacity is used to increase the service probability of base customers, so the marginal gross profit of a call equals the base customer OTV. Since this OTV equals the capacity cost per call, any capacity in the range $[\underline{\lambda}_n, \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$ is optimal and yields same profit.

In Parts 1.(c) and 2.(a) of Proposition 3, the capacity cost is sufficiently low so that it is optimal to serve all calls, which implies the capacity scenario in Part 2 of Proposition 2. In this case $\lambda_n^* = \min\left\{\overline{\lambda}_n, N\mu/(1 + \theta_n r_b/\gamma_b)\right\}$ by (23), where $\overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$ is the optimal total arrival rate in an underloaded system and $\overline{\lambda}_n$ is given by (20). Since the optimal capacity cannot exceed this rate, consider $\lambda_n^* = N\mu/(1 + \theta_n r_b/\gamma_b) < \overline{\lambda}_n$. In this capacity range, it is optimal to match an increase in capacity by an equal increase in the *total* call arrival rate; new customer calls increase at a rate of $1/(1 + \theta_n r_b/\gamma_b)$ per extra unit of capacity, and base customer calls increase to absorb the remaining capacity share. Therefore, the marginal gross profit of capacity equals

$$\frac{V_n - c_n + V_b \theta_n r_b/\gamma_b - S'\left(\frac{N\mu}{1+\theta_n r_b/\gamma_b}\right)}{1 + \theta_n r_b/\gamma_b},$$

where the numerator is the marginal net revenue minus marginal advertising cost of a new customer call, and the denominator reflects the fact that only a fraction of additional capacity translates into a higher new customer arrival rate. By (26), this marginal gross profit equals the cost of a call at the optimal capacity $N^*\mu$. In particular, if $S(\lambda_n) = \alpha (\lambda_n)^\beta$ then

$$N^*\mu = \left(\frac{V_n - c_n + V_b \theta_n r_b/\gamma_b - \frac{C}{\mu}\left(1 + \theta_n \frac{r_b}{\gamma_b}\right)}{\alpha\beta}\right)^{\frac{1}{\beta-1}} \left(1 + \theta_n \frac{r_b}{\gamma_b}\right).$$

Remark 5. For a stochastic system, Parts 1.(c) and 2.(a) of Proposition 3 require the following qualification (see Remark 4 in §4.3). It is preferable to prioritize customers with the higher OTV regardless of load. In particular, if base customers have the higher OTV (Part 2.(a) of Proposition 3), it may be the unique optimal policy to prioritize them. Under this policy, some of the new customer calls, which the firm spends money to attract, are lost.

Corollary 1 highlights the fact that the ratio of optimal profit to advertising expenditure under the power model depends *only* on the response elasticity to advertising.

Corollary 1 *If the advertising cost follows the power model, i.e., $S(\lambda_n) = \alpha (\lambda_n)^\beta$, then under the optimal priority, promotion and staffing policy, the ratio of profit to advertising expenditure is $\beta - 1$.*

4.5 Sensitivity of Optimal Decisions to Model Parameters

The results in §4.2-4.4 show that the optimal policy critically depends on the difference between the OTVs of new and base customers, i.e., $V_n - V_b$ if the new customer arrival rate is fixed and

$V_n - c_n - V_b$ if the firm controls the new customer arrival rate (see Remark 2 in §4.1). Under the optimal promotion policy, if $V_n - c_n \leq V_b$, the optimal system serves all calls, whereas if $V_n - c_n > V_b$, the optimal system serves all calls of new customers but possibly none or only some of base customers. Corollary 2 specifies the sensitivity of $V_n - V_b$ to customer-related parameters.

Corollary 2 *The call center's preference to operate an overloaded system in which calls of new customers displace those of base customers increases in the difference $V_n - V_b$.*

1. *$V_n - V_b$ increases in the per-call profit and abandonment cost of a new customer, p_n and c_n, respectively, and decreases in the corresponding base customer metrics, p_b and c_b.*
2. *If $R > r_b c_b$ ($R < r_b c_b$), then $V_n - V_b$ increases (decreases) in the probability of joining the customer base after service, θ_n, and in the probability of staying in the customer base after abandoning, θ_b. If $R = r_b c_b$, then $V_n - V_b$ is constant in θ_n and θ_b.*
3. *If $\theta_n > 1 - \theta_b$ ($\theta_n < 1 - \theta_b$), then $V_n - V_b$ increases (decreases) in the base customers' call-independent profit rate, R, and decreases (increases) in their call arrival rate, r_b.*
   If $\theta_n = 1 - \theta_b$, then $V_n - V_b$ is constant in R and r_b.
4. *If $(\theta_n + \theta_b - 1)(R - r_b c_b) > (<)\, 0$ then $V_n - V_b$ decreases (increases) in the call-independent attrition rate, γ_b. If $\theta_n = 1 - \theta_b$ and $R = r_b c_b$, then $V_n - V_b$ is constant in γ_b.*

The unambiguous effects in Part 1 of Corollary 2 are intuitively clear: The system is more eager to serve new customers the higher their per-call profit and their abandonment cost. By Part 2 of Corollary 2 the sensitivity of $V_n - V_b$ to θ_n and θ_b depends on the profitability of a base customer who is *not* served, since $R > r_b c_b$ implies $L(0) > 0$ by (13). If such a customer is profitable, then the higher the propensity of new customers to turn into base customers (i.e., high θ_n) and to remain loyal (i.e., high θ_b), the larger the new customer OTV relative to the base customer OTV. High values of θ_n and θ_b may characterize firms that enjoy high brand loyalty, face weak competition, and/or saddle their customers with high process-related or monetary switching costs. Conversely, if $R < r_b c_b$, e.g., in the absence of call-independent revenues, the new customer OTV decreases in θ_n and θ_b relative to the OTV of base customers, because the firm loses money on base customers that it does not serve. By Part 3 of Corollary 2, the sensitivity of $V_n - V_b$ to R and r_b depends on θ_n and $1 - \theta_b$. The OTVs of *both* customer types increase in base customers' call-independent profit rate R, and decrease in their call arrival rate r_b. If $\theta_n > 1 - \theta_b$, the new customer OTV is more sensitive to each of these changes than the base customer OTV, because new customers are more likely to join than base customers are to leave the company. In such settings, the OTV difference $V_n - V_b$ increases in the call-independent profit rate R (or a decrease in r_b), which promotes serving more new customers to grow the customer base. Conversely, if it is harder to gain and easier to lose base customers, i.e., $\theta_n < 1 - \theta_b$, an increase in R reduces V_n relative to V_b, making it more attractive to serve base customers. Finally, by Part 4 of Corollary 2, $V_n - V_b$ decreases in γ_b, if new customers are both more (less) likely to convert to base customers than to leave the customer base and are (not) profitable as base customers that are not served.

5 Fluid Model Validation: Simulation Results

In this section, we study the accuracy of the fluid model approximation by comparing its performance with simulation results for the stochastic system described in §3.1. We assume exponentially distributed service times. §5.1 specifies the parameter values for this simulation study. We report our results in three steps. In §5.2 we report, for a wide range of *fixed* load factors, the accuracy of the fluid model in approximating key steady-state performance measures. We then report the performance of the fluid model in approximating the *optimal* decisions and profit, in §5.3 for fixed capacity, and in §5.4 under the jointly optimal priority policy, promotion, and staffing level.

5.1 Parameter Values

Table 3 summarizes the parameter values for the simulation study. These values may be representative for a call center of a business with significant call-independent revenue, such as the example of a mobile phone service provider outlined in §3.1.

Parameter		Value
Service rate per server (per day)	μ	100
Call abandonment rate (per day)	τ	100
Call arrival rate per base customer (per day)	r_b	0.01
P(new customer joins customer base after service)	θ_n	0.3
P(base customer remains in customer base after abandoning)	θ_b	0.9
Attrition rate per base customer (per day)	γ_b	0.002
Profit rate per base customer (per day)	R	1.0
Profit per served call	p_n, p_b	$10, -10$
Cost per abandoned call	c_n, c_b	$0.25, 0.5$
Advertising cost function parameters (power model)	α, β	$0.5, 1.5$

Table 3: Parameter Values for Simulation.

One time unit equals one day. We assume a 24x7 operation. The mean service and abandonment times are 14.4 minutes each (since $\mu = \tau = 100$ calls per day). A base customer calls on average once every one hundred days (since $r_b = 0.01$ per day). Her mean lifetime in the absence of abandonment is 500 days (since $\gamma_b = 0.002$ per day). Increasing the new customer call arrival rate by one call increases the total arrival rate at most by $1 + \theta_n r_b/\gamma_b = 2.5$ calls.

The average profit per served call of a base customer is negative; since existing customers already have a contract with the company, they typically call with service requests, not to buy new products/services. Base customers generate profit from subscription fees and communication charges, at an average rate of $R = \$1$ per day per customer, i.e., \$30 per month per customer. Based on (13) the CLV varies from $L(0) = \$331.67$ to $L(1) = \$450.0$ depending on the base customer service probability. From (16) and (14) the OTVs of new and base customers are $V_n = \$109.75$ and $V_b = \$23.67$, respectively. Since $V_n - c_n > V_b$, the fluid model results of §4 suggest that it is optimal to design and operate an overloaded system for some capacity (cost) levels.

We assume an advertising cost function that follows the power model and let $\alpha = 0.5$, $\beta = 1.5$.

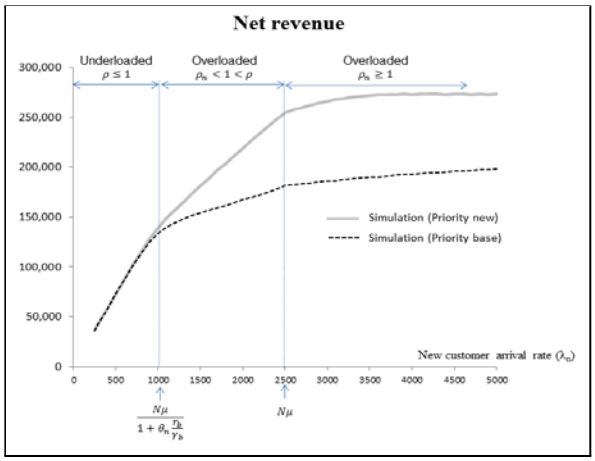

Figure 5: Simulated System: Net Revenue as Function of New Customer Arrival Rate and Priority Policy.

5.2 Accuracy of Steady-State Performance Measures for a Fixed Load

In this section we report the accuracy of the fluid model in approximating four key steady-state performance measures: the average size of the customer base x_b; the abandonment probabilities of new and base customers, $1 - q_n$ and $1 - q_b$, respectively; and the net revenue rate (17). For fixed new customer arrival rate and capacity, the net revenue rate is uniquely determined by (x_b, q_n, q_b).

We simulated the system for capacity levels between $N\mu = 2,500$ and $N\mu = 110,000$, increasing $N\mu$ in increments of $2,500$. For each capacity level we varied the new customer arrival rate λ_n, such that the maximum load factor $\rho = \lambda_n \left(1 + \theta_n r_b/\gamma_b\right)/N\mu$ ranges from $\rho = 0.2$ to $\rho = 5$, in increments of 0.1. In every case, we set the initial size of the customer base at the equilibrium size suggested by the fluid model, ran the simulation for $1,100,000$ new customer arrivals, and discarded results from the first $100,000$ arrivals in computing performance measures. (Starting from an empty system, a much longer warm-up period is required to reach steady-state).

First, consider the profitability of prioritizing new vs. base customers in the stochastic system. Figure 5 shows the net revenues under these two priority policies, as a function of the new customer arrival rate, for a simulated system with capacity $N\mu = 2,500$ calls per day (these graphs are representative of those for different capacity levels). Since $1+\theta_n r_b/\gamma_b = 2.5$, the system is underloaded if $\lambda_n \leq 2,500/\left(1 + \theta_n r_b/\gamma_b\right) = 1,000$. Consistent with the fluid model, the net revenue functions are virtually identical as long as the system is underloaded, and prioritizing new customers is strictly more profitable as the system gets overloaded, because the OTV of new customers exceeds that of base customers. These simulation results indicate that, as predicted by our fluid model results (see Proposition 1), it is optimal to prioritize new customers in the stochastic system. Therefore, the following simulation results focus on the new customer priority policy.

Figure 6 compares the performance of a simulated system vs. the fluid model, for capacity $N\mu = 2,500$. These graphs look similar for different capacity levels, except that the approximation errors decrease dramatically in the number of servers, as expected for a fluid model (see Table 4).

When the system is underloaded (i.e., $\rho \leq 1$, for $\lambda_n \leq 1,000$) the fluid model is very accurate, as long as all customers get served. As the new customer arrival rate increases to the high end

of the underloaded regime, the accuracy of the fluid model decreases. Because it ignores queueing effects, it underestimates the abandonment probabilities of both customer types, and this estimation error is larger for the low-priority base customer calls; see Figure 6(a)-(b). Underestimating these abandonment probabilities leads to overestimating the average size of the customer base and the net revenue, although these errors are still quite small for $\lambda_n \leq 1,000$; see Figure 6(c)-(d). For a balanced system (i.e., $\lambda_n = 1,000$) the fluid model overestimates the net revenue by only 4.24%. The accuracy of the fluid model further decreases as the new customer arrival rate increases, so that the system is overloaded with residual capacity for base customers (i.e., $\rho_n < 1 < \rho$, for $1,000 < \lambda_n < 2,500$). In this regime, the fluid model underestimates the abandonment probability of high-priority new customer calls and overestimates that of low-priority base customer calls; see Figure 6(a)-(b). The fluid model ignores that high-priority customers may abandon the queue and therefore underestimates the residual capacity available for base customers. Underestimating the abandonment of new customers results in overestimating the size of the customer base and the net revenue; see Figure 6(c)-(d). These approximation errors increase in the new customer arrival rate, and they are maximized at the point where the new customer load factor $\rho_n \approx 1$ (i.e., $\lambda_n \approx N\mu = 2,500$); at this point the fluid model overestimates the net revenue by 7.65%.

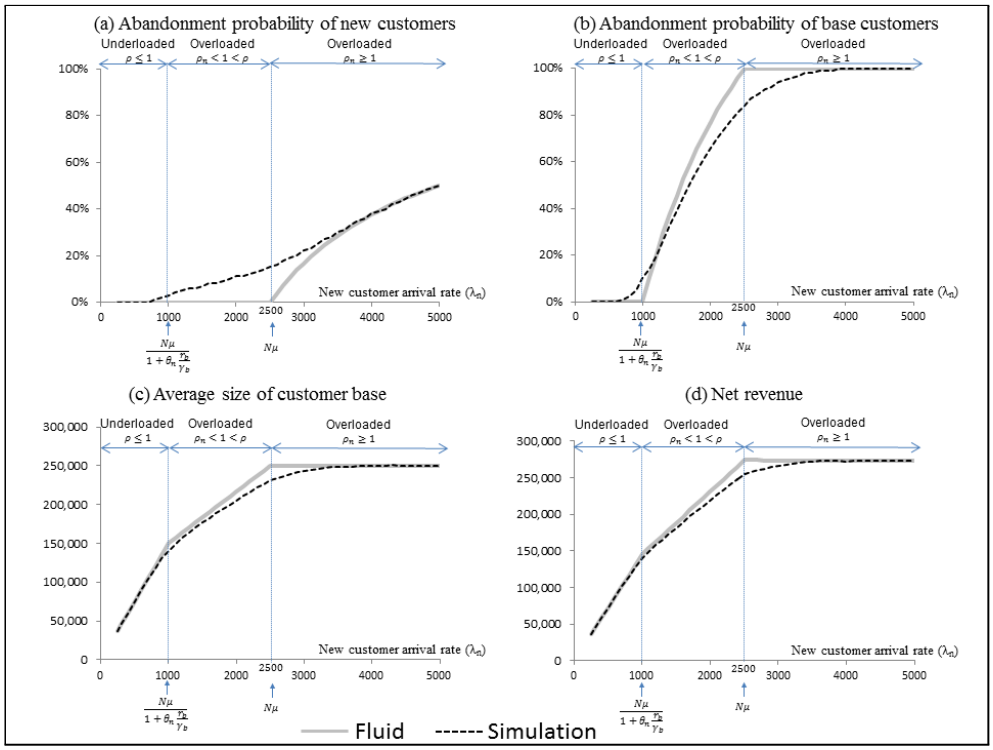

Figure 6: Fluid Model vs. Simulation: Steady-State Performance as Function of New Customer Arrival Rate (Priority to New Customers).

As the system becomes overloaded with new customer calls the approximation errors diminish and vanish eventually. Unlike in the fluid model, in the stochastic system the new customer throughput increases in $\rho_n > 1$, up to the point where it approaches capacity, at $\lambda_n \approx 3,700$.

$N\mu$	2,500	5,000	10,000	20,000	30,000	50,000	70,000	90,000	110,000
Error	7.65%	4.06%	2.25%	1.31%	0.96%	0.72%	0.62%	0.53%	0.47%

Table 4: Error in Approximating the Net Revenue for $\rho_n = 1$, as Function of Capacity.

In summary, the net revenue under fluid assumptions matches that in the stochastic system for load factors below one ($\rho < 1$) and new customer load factors well above one ($\rho_n > 1$). In these ranges the total and new customer throughput in the fluid model are the same as in the stochastic system. In between these ranges, the fluid model overestimates the net revenue. As noted above, the system exhibits similar behavior at every capacity level, i.e., the percentage error in the net revenue estimate is largest at $\rho_n \approx 1$, but this error decreases significantly with capacity (Table 4).

5.3 Accuracy of Joint Priority and Promotion Prescriptions for Fixed Staffing

In this section we report the performance of the fluid model in approximating the *optimal* new customer arrival rate and gross profit, for fixed capacity. As noted in §5.1 the new customer OTV exceeds the base customer OTV, i.e., $V_n - c_n > V_b$. By Part 1 of Proposition 2, in the fluid model the system is overloaded under the jointly optimal priority and promotion policy for relatively low capacity levels, i.e., $N\mu < \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$, and only new customers are served if $N\mu < \underline{\lambda}_n$. By Part 2 of Proposition 2, the optimal system is balanced if $N\mu \in [\underline{\lambda}_n (1 + \theta_n r_b/\gamma_b), \overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$ and underloaded if $N\mu > \overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$. Using (22) we obtain the following values for the thresholds: $\underline{\lambda}_n = 13,098$, $\underline{\lambda}_n (1 + \theta_n r_b/\gamma_b) = 32,744$, and $\overline{\lambda}_n (1 + \theta_n r_b/\gamma_b) = 93,444$ (where $\overline{\lambda}_n = 37,378$).

Figure 7(a) shows that the optimal new customer arrival rate prescribed by the fluid model closely approximates the optimal new customer arrival rate for the simulated system, except for capacity levels in two intervals, containing the thresholds $\underline{\lambda}_n$ and $\overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$, respectively. These are the largest capacity levels at which the optimal new customer arrival rate and the optimal total arrival rate, respectively, equal capacity under the fluid model solution. In these ranges, approximately for $N\mu$ in the intervals $[10000, 30000]$ and $[90000, 110000]$, the optimal new customer arrival rate prescribed by the fluid model is larger than optimal in the stochastic system.

Figure 7(b) shows that the maximum percentage errors in these intervals are approximately 6.5% at $N\mu = 15,000$, and 0.8% at $N\mu = 102,500$, respectively. More importantly, these errors in approximating the optimal new customer arrival rate of the stochastic system translate into a significantly smaller error in the corresponding optimal gross profit. Specifically, the highest loss in gross profit when operating the stochastic system with the optimal new customer arrival rate prescribed by the fluid model is 0.6%, at $N\mu = 15,000$.

5.4 Accuracy of Joint Priority, Promotion, and Staffing Prescriptions

Finally, in this section, we report the performance of the fluid model in approximating the *optimal* capacity and profit, as a function of the capacity cost. We vary the capacity cost per call C/μ from 0 to 50, in unit increments. As noted in §5.1 we have $V_n - c_n > V_b$ and $V_b = 23.67$. By Proposition

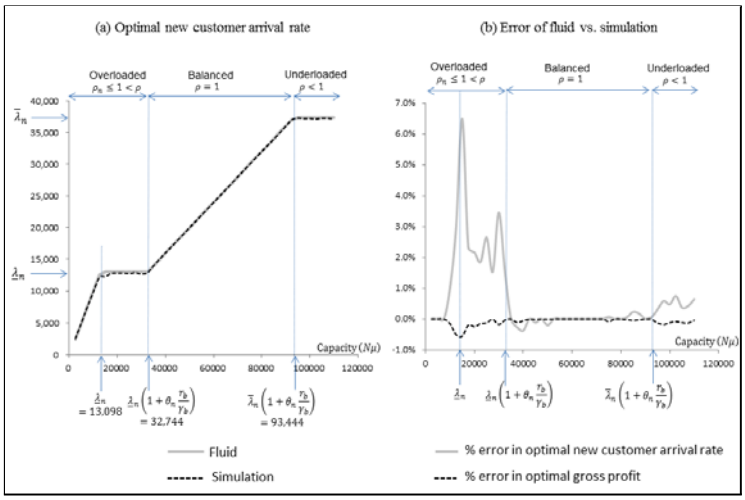

Figure 7: Fluid Model vs. Simulation: Optimal New Customer Arrival Rate, and Percentage Errors in Optimal New Customer Arrival Rate and Corresponding Gross Profit (= Net Revenue - Advertising Cost), as Functions of Capacity (Priority to New Customers).

3, V_b is an important threshold in the fluid model solution: it prescribes an overloaded system in which only new customers are served for $C/\mu > V_b$, and a balanced system in which all customers are served for $C/\mu < V_b$; for $C/\mu = V_b$, any capacity in the interval $[\underline{\lambda}_n, \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$ is optimal.

Figure 8(a) shows that the optimal capacity prescribed by the fluid model closely approximates the optimal capacity for the simulated system, except for $C/\mu \in [23, 29]$, which contains $V_b = 23.67$. Figure 8(b) shows that while the maximum percentage error in this interval is approximately 60.6%, at $C/\mu = 24$, the error outside this interval is much smaller, at less than 5%. The substantial fluid approximation error in the optimal capacity level for $C/\mu \in [23, 29]$ is consistent with the correspondingly large errors in approximating the optimal new customer arrival rate: for $C/\mu = V_b$, the fluid model prescribes a capacity level in $[\underline{\lambda}_n, \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$, and as shown in Figure 7(b), at these capacity levels, the error of the fluid model in approximating the optimal new customer arrival rate is largest. More importantly, however, these errors in approximating the optimal capacity level of the stochastic system translate into a significantly smaller error in the corresponding optimal profit: Figure 8(b) shows that for $C/\mu \in [23, 29]$ the maximum loss in profit is approximately 5.7%, at $C/\mu = 24$. Outside this interval the error in approximating the optimal profit is less than 1%.

6 Concluding Remarks

This paper proposes and analyzes a novel call center model that considers the impact of past demand and service quality on customer retention. We study the problem of maximizing profits by controlling customer acquisition, retention, and service quality via promotions, priorities, and staffing. The key feature of our model is that the customer base depends on the abandonment rates of new and base customers, reflecting their priority and the system load. We specify a stochastic queueing model, characterize the optimal controls analytically based on a deterministic fluid model,

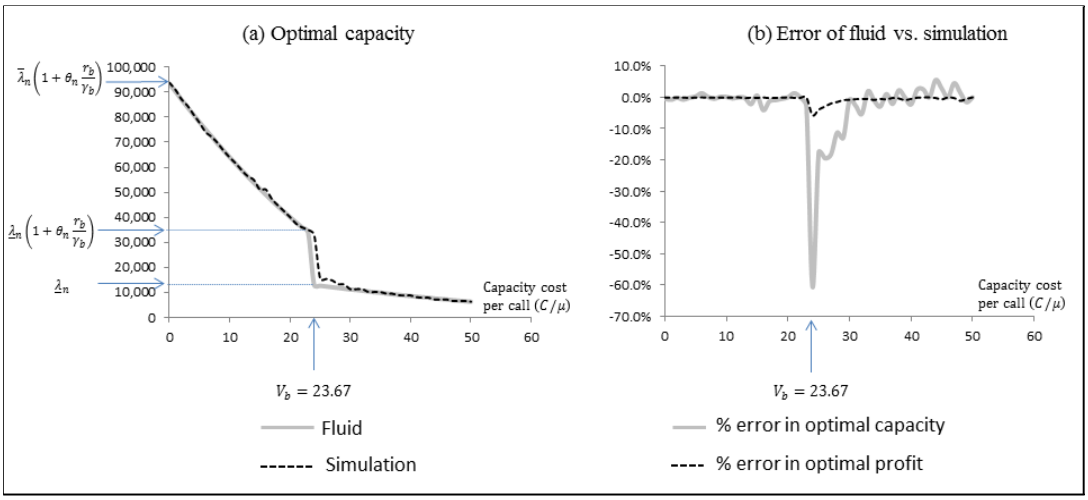

Figure 8: Fluid Model vs. Simulation: Optimal Capacity, and Percentage Errors in Optimal Capacity and Corresponding Profit, as Functions of Capacity Cost Per Call (Priority to New Customers).

and show via simulation that these prescriptions yield near-optimal performance for the underlying stochastic model. The following findings and implications emerge from our analysis.

First, we provide novel insights and guidelines on call center management. We derive three metrics which form the basis for call center decisions, the CLV of a base customer and the OTVs of new and base customers. These metrics, unlike standard ones in marketing, depend on operations through the service quality, i.e., the probability of getting served. We show that it is optimal to prioritize the customers with the higher OTV. In contrast to standard priority policies such as the $c\mu$ rule, this policy accounts for the financial impact of customers' future calls. We further show how the jointly optimal promotion, priority and staffing policy depends on the CLV, the OTVs, and the capacity cost. The results on the optimal promotion and staffing levels underscore the importance of considering the interaction between customer metrics and operations; i.e., the contribution of an additional new customer to net revenues depends on the system load and the priority policy. These results also provide insights on the optimal service quality. Specifically, offering deliberately poor service to base customers is optimal only if new customers have the higher OTV, e.g., due to prohibitive switching costs for base customers, and the capacity cost per call exceeds the OTV of a base customer. Under these conditions it is optimal to prioritize new customers and to *overload* the system, possibly significantly. In this regime the call center's primary goal is to serve and acquire new customers to grow the customer base, not to serve base customers. This result lends some theoretical support for the anecdotal evidence that locked-in customers of firms such us mobile phone service providers commonly experience long waiting times when contacting the call center.

Second, from a modeling and methodological perspective, we conclude that the benefits of our solution approach via the analysis of the deterministic fluid model outweigh its disadvantages. The fluid model is particularly appealing because it is both analytically tractable, as is obvious from §4, and its prescriptions yield near-optimal (gross) profit performance for the approximated stochastic system, as shown in §5. Furthermore, the fluid model analysis provides valuable intuition on the

problem structure. The obvious caveat of the fluid model is that it does not account for queueing and abandonment in underloaded systems. This leads to two types of errors, which can be handled, however. (*i*) As discussed in Remarks 3-5, the fluid model incorrectly assumes that the profit is independent of the priority policy in an underloaded system. Since customers may abandon in an underloaded stochastic system, it is preferable in stochastic systems to prioritize customers with the higher OTV (in line with Proposition 1) regardless of load. (*ii*) Even under the "correct" priority policy, as shown in §5.3-§5.4 the fluid model is less accurate in approximating the optimal decision for the stochastic system, than in approximating the optimal (gross) profit. A natural solution to this problem is to use the fluid model to identify the solution candidate as a starting point for determining the optimal decision via simulation-based optimization. In summary, the tractability and accuracy of the fluid model demonstrated in this paper suggest that a similar approach may prove effective on further problems of joint CRM and call center management.

We close by outlining three future research directions. First, in terms of customer modeling, we model homogenous base customers that defect based only on their last call center interaction. It is important to consider heterogenous customers that differ based on service-independent attributes and/or their service histories. Second, in terms of system modeling and solution methodology, one 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 data, as discussed in §3.1, our model can be tailored to a range of call center characteristics. Many of our model inputs are reasonably well measurable based on data that call centers track. It would be quite interesting to estimate our model parameters and also to refine our model, based on such data. The results could be of value to measure and compare CLV and OTV metrics within and across call centers, and more importantly, to study the impact of service quality attributes such as waiting time on these metrics.

References

Adelman, D., A.J. Mersereau. 2012. Dynamic capacity allocation to customers who remember past service. Forthcoming in *Management Science*.

Aflaki, S., I. Popescu. 2012. Managing retention in service relationships. Working paper, INSEAD.

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

Akşin, O.Z., M. Armony, V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Prod. & Oper. Management* **16**(6) 665-688.

Anton, J., T. Setting, C. Gunderson. 2004. Offshore company call centers: A concern to U.S. consumers. Technical Report, Purdue University Center for Customer-Driven Quality.

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

Bagwell, K. 2007. The economic analysis of advertising. M. Armstrong, R. Porter (eds.) *Handbook of Industrial Organization*, Vol. 3, Chapter 28. North-Holland, Amsterdam.

Bitran, G., S. Mondschein. 1996. Mailing decisions in the catalog sales industry. *Management Sci.* **42**(9) 1364–1381.

Bitran, R. G., P. Rocha e Oliveira, A. Schilkrut. 2008. Managing customer relationships through price and service quality. Working paper, IESE, Spain.

Blattberg, R. C., J. Deighton. 1996. Manage marketing by the customer equity test. *Harvard Bus. Rev.* **74**(4) 136–145.

Debo, L.G., B. Toktay, L.K. Wassenhove. 2008. Queueing for expert services. *Management Sci.* **54**(8) 1497–1512.

de Véricourt, F., Y.-P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Farzan. A., H. Wang, Y.-P. Zhou. 2012. Setting quality and speed in service industry with repeat customers. Working paper, University of Washington.

Feichtinger, G., R. F. Hartl, S. P. Sethi. 1994. Dynamic optimal control models in advertising: recent developments. *Management Sci.* **40**(2) 195-226.

Feinberg, R.A., L. Hokama, R. Kadam, I.S. Kim. 2002. Operational determinants of caller satisfaction in the banking/financial services call center. *Int. J. Bank Marketing* **20**(4) 174-180.

Gans, N. 2002. Customer loyalty and supplier quality competition. *Management Sci.* **48**(2) 207-221.

Gans, N., Koole, G., Mandelbaum, A. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management* **5**(2) 79–141.

Gaur, V., Y. Park. 2007. Asymmetric consumer learning and inventory competition. *Management Sci.* **53**(2) 227-240.

Genesys Telecommunications Labs. 2007. *Genesys Global Consumer Survey*. Daily City, CA, USA.

Green, L.V., P.J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Prod. & Oper. Management* **16**(1) 13–39.

Günes, E. D., O. Z. Akşin, E. L. Örmeci, S. H. Özden. 2010. Modelling customer reactions to sales attempts: if cross-selling backfires. *J. Service Res.* **13**(2) 168-183.

Gupta, S., D.R. Lehmann. 2008. Models of customer value. Wierenga B. (ed.) *Handbook of Marketing Decision Models*. Springer Science.

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

Hall, J., E. Porteus. 2000. Customer service competition in capacitated systems. *Manufacturing Service Oper. Management* **2**(2) 144-165.

Hanssens, D.M., L.J. Parsons, R.L. Schultz. 2001. *Market response models: Econometric and time series analysis*. Springer.

Hassin R., M. Haviv. 2003. *To Queue or not to queue: Equilibrium behavior in queueing systems*. Kluwer, Boston.

Ho, T.H., Y.H. Park, Y.P. Zhou. 2006. Incorporating satisfaction into customer value analysis: Optimal investment in lifetime value. *Marketing Sci.* **25**(3) 260-277.

Horstmann, I. J., S. Moorthy. 2003. Advertising spending and quality for services: The role of capacity. *Quant. Marketing Econ.* **1**(3) 337–365.

Lewis, M. A. 2005. Dynamic programming approach to customer relationship pricing. *Management Sci.* **51**(6) 986–994.

Lilien, G.L., P. Kotler, K.S. Moorthy. 1992. *Marketing Models*. Prentice Hall.

Liu, L., W. Shang, S. Wu. 2007. Dynamic competitive newsvendors with service-sensitive demands. *Manufacturing Service Oper. Management* **9**(1) 84–93.

Olsen, T.L., R.P. Parker. 2008. Inventory management under market size dynamics. *Management Science* **54**(10) 1805-1821.

Ovchinnikov, A., B. Boulu, P.E. Pfeifer. 2012. Revenue management with lifetime value considerations. Working paper, Darden School of Business, University of Virginia.

Pfeifer, P.E., A. Ovchinnikov. 2011. A note on willingness to spend and customer lifetime value for firms with limited capacity. *J. Interactive Marketing* **25**(3) 178-189.

Randhawa, R. S., S. Kumar. 2008. Usage restriction and subscription services: operational benefits with rational customers. *Manufacturing and Service Oper. Management* **10**(3) 429-447.

Reinartz, W., R. Venkatesan. 2008. Decision models for customer relationship management (CRM). Wierenga B. (ed.) *Handbook of Marketing Decision Models*. Springer Science.

Rust, R.T., T.S. Chung. 2006. Marketing models of service and relationships. *Marketing Sci.* **25**(6) 560–580.

Schwartz, B. L. 1966. A new approach to stockout penalties. *Management Sci.* **12**(12) B538–B544.

Sethi, S.P., Q. Zhang. 1995. Multilevel hierarchical decision making in stochastic marketing-production systems. *SIAM J. Control. and Optimization* **33**(2) 528-553.

Simon, J.L., J. Arndt. 1980. The shape of the advertising response function. *J. Advertising Res.* **20**(4) 11–28.

Sun, B., S. Li. 2011. Learning and acting on customer information: A simulation-based demonstration on service allocations with offshore centers. *J. Marketing Research* **48**(1) 72-86.

Online Supplement: Proofs

As a basis for the proofs, Lemma 1 provides the profit rates from §3.2 in more convenient forms.

Lemma 1 *The profit rate as a function of the load factor and the priority policy is given by:*

1. *For an underloaded system ($\rho \leq 1$):*

$$\Pi = \lambda_n \left(V_n - c_n + V_b \theta_n \frac{r_b}{\gamma_b} \right) - S(\lambda_n) - CN. \tag{27}$$

2. *For an overloaded system ($\rho > 1$) that prioritizes base customers:*

$$\Pi = \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b} \left(V_n + V_b \theta_n \frac{r_b}{\gamma_b} \right) - \lambda_n c_n - S(\lambda_n) - CN. \tag{28}$$

3. *For an overloaded system with $\rho_n < 1 < \rho$ that prioritizes new customers :*

$$\Pi = \lambda_n (V_n - c_n) + (N\mu - \lambda_n) V_b - S(\lambda_n) - CN. \tag{29}$$

4. *For an overloaded system with $\rho_n \geq 1$ that prioritizes new customers:*

$$\Pi = N\mu V_n - \lambda_n c_n - S(\lambda_n) - CN. \tag{30}$$

Proof. *Part 1.* In this case the profit rate is given by (7). Factoring out λ_n yields:

$$\Pi = \lambda_n \left(p_n + \frac{\theta_n}{\gamma_b} (R + r_b p_b) \right) - S(\lambda_n) - CN. \tag{31}$$

We observe that

$$\frac{\theta_n}{\gamma_b} (R + r_b p_b) = \theta_n L(1) = \theta_n L(0) + V_b \theta_n \frac{r_b}{\gamma_b}. \tag{32}$$

The first equality follows from (13), the second from (15). Substituting (32) into (31) gives

$$\Pi = \lambda_n \left(p_n + \theta_n L(0) + V_b \theta_n \frac{r_b}{\gamma_b} \right) - S(\lambda_n) - CN.$$

Equation (27) now follows using (16).

Part 2. In this case the profit rate is given by (9). Factoring out $N\mu\gamma_b / (\gamma_b + \theta_n r_b)$ yields:

$$\Pi = \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b} \left(p_n + c_n + \frac{\theta_n}{\gamma_b} (R + r_b p_b) \right) - \lambda_n c_n - S(\lambda_n) - CN.$$

Equation (28) now follows from (32) and (16), as above.

Part 3. In this case the profit rate is given by (11). From (13) we have

$$L(0) = \frac{R - r_b c_b}{\gamma_b + r_b (1 - \theta_b)}. \tag{33}$$

Substituting (33) into (11) gives:

$$\begin{aligned}\Pi &= \lambda_n p_n + (p_b + c_b)(N\mu - \lambda_n) + (\theta_n \lambda_n + (1-\theta_b)(N\mu - \lambda_n))L(0) - S(\lambda_n) - CN \\ &= \lambda_n (p_n + \theta_n L(0)) + (N\mu - \lambda_n)(p_b + c_b + (1-\theta_b)L(0)) - S(\lambda_n) - CN \end{aligned} \tag{34}$$

Using the definitions of V_b and V_n from (14) and (16) respectively, the profit rate in (34) can be written as (29).

Part 4. In this case the profit rate is given by (12). Again, by substituting (33) we have:

$$\begin{aligned}\Pi &= N\mu p_n - (\lambda_n - N\mu)c_n + N\mu\theta_n L(0) - S(\lambda_n) - CN \\ &= N\mu(p_n + c_n + \theta_n L(0)) - \lambda_n c_n - S(\lambda_n) - CN.\end{aligned}$$

Equation (30) now follows using (16).■

Proof of Proposition 1. For an underloaded system the profit is independent of the priority policy by (27). For an overloaded system we consider the two possible cases: $\rho_n < 1 < \rho$ and $\rho_n \geq 1$.

Overloaded system with $\rho_n < 1 < \rho$. It is optimal to prioritize new customers if the profit rate (29) exceeds the profit rate in (28) under base customer prioritization, i.e., if

$$0 \leq \lambda_n V_n + (N\mu - \lambda_n)V_b - \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}\left(V_n + V_b\theta_n\frac{r_b}{\gamma_b}\right) = \left(\lambda_n - \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}\right)(V_n - V_b). \tag{35}$$

By (6), in an overloaded system $\lambda_n > N\mu/(1 + \theta_n r_b/\gamma_b)$, which means that (35) holds if and only if $V_n - V_b \geq 0$, establishing Proposition 1 for this load range.

Overloaded system with $\rho_n \geq 1$. By (28) and (30), it is optimal to prioritize new customers if

$$0 \leq N\mu V_n - \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}\left(V_n + V_b\theta_n\frac{r_b}{\gamma_b}\right) = \left(N\mu - \frac{N\mu\gamma_b}{\gamma_b + \theta_n r_b}\right)(V_n - V_b) = \left(\frac{N\mu\theta_n r_b}{\gamma_b + \theta_n r_b}\right)(V_n - V_b),$$

which again holds if and only if $V_n - V_b \geq 0$, completing the proof.■

Proof of Proposition 2. The proof proceeds as follows. First, we show that it is optimal to prioritize new customers under the optimal advertising policy. Next, we characterize the optimal new customer arrival rate for a system that prioritizes new customers, as a function of the load factor. Finally, we characterize the optimal load factor and the optimal new customer arrival rate for each part of the Proposition.

Prioritizing new customers is optimal. The profit rate of an overloaded system that prioritizes base customer calls is given by (28). It decreases in λ_n since $S' > 0$ and

$$\frac{d\Pi}{d\lambda_n} = -c_n - S'(\lambda_n) < 0.$$

It follows that it is not profitable to overload a system that prioritizes base customers. That it is optimal to prioritize new customers follows because by (27), the profit rate of an underloaded system is independent of the priority policy.

Next, we characterize the optimal new customer arrival rate for a system that prioritizes new customers, for each of the following possible load ranges: $\rho \leq 1$, $\rho_n < 1 < \rho$, and $\rho_n \geq 1$.

Underloaded system ($\rho \leq 1$). In this load range the profit rate is given by (27), and its first order condition (FOC) with respect to λ_n is

$$\frac{d\Pi}{d\lambda_n} = V_n - c_n + V_b \theta_n \frac{r_b}{\gamma_b} - S'(\lambda_n) = 0. \tag{36}$$

The FOC (36) has an unique solution by Assumption 1 and since $S'(0) = 0 < S''$. Let $\overline{\lambda}_n$ be this solution, and note that this definition is identical to (20). Furthermore

$$\overline{\lambda}_n := S'^{-1}(V_n - c_n + V_b \theta_n r_b/\gamma_b) > 0. \tag{37}$$

For the system to be underloaded at this arrival rate, $\overline{\lambda}_n$ must satisfy (6). Therefore if

$$\overline{\lambda}_n (1 + \theta_n r_b/\gamma_b) \leq N\mu, \tag{38}$$

then $\overline{\lambda}_n$ is the optimal new customer arrival rate. Otherwise the profit rate increases in λ_n when the system is underloaded.

Overloaded system with $\rho_n < 1 < \rho$. The profit rate is given by (29), and the FOC is

$$\frac{d\Pi}{d\lambda_n} = V_n - c_n - V_b - S'(\lambda_n) = 0. \tag{39}$$

If $V_n - c_n - V_b \leq 0$, the profit rate decreases in λ_n in this load range. If $V_n - c_n - V_b > 0$, let $\underline{\lambda}_n$ be the unique solution of (39), and note that this definition is identical to (21). Furthermore

$$\underline{\lambda}_n := S'^{-1}(V_n - c_n - V_b) > 0. \tag{40}$$

The condition $\rho_n < 1 < \rho$ is equivalent to $\lambda_n < N\mu < \lambda_n (1 + \theta_n r_b/\gamma_b)$. Therefore, $\underline{\lambda}_n$ is the optimal new customer arrival rate if

$$\frac{N\mu}{(1 + \theta_n r_b/\gamma_b)} < \underline{\lambda}_n < N\mu. \tag{41}$$

Otherwise, if $\underline{\lambda}_n \leq N\mu/(1 + \theta_n r_b/\gamma_b)$ then the profit rate decreases in λ_n while $\rho_n < 1 < \rho$, and if $N\mu \leq \underline{\lambda}_n$, then the profit peaks at a new customer arrival rate for which $\rho_n \geq 1$.

Overloaded system with $\rho_n \geq 1$. In this load range the profit rate is given by (30), and we have

$$\frac{d\Pi}{d\lambda_n} = -c_n - S'(\lambda_n) < 0. \tag{42}$$

Thus, when $\rho_n \geq 1$, the optimal new customer arrival rate equals $N\mu$, for which $\rho_n = 1$.

Now that we have established the optimal decision for each load range, we prove each part of the Proposition.

Part 1. If $V_n - c_n > V_b$ and $N\mu < \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$: We know

$$\underline{\lambda}_n = S'^{-1}(V_n - c_n - V_b) < S'^{-1}(V_n - c_n + V_b \theta_n r_b/\gamma_b) = \overline{\lambda}_n,$$

where the inequality follows because $V_n - c_n - V_b < V_n - c_n + V_b\theta_n r_b/\gamma_b$ and S'^{-1} is strictly increasing (since $S'' > 0$). Therefore:

$$N\mu < \underline{\lambda}_n \left(1 + \theta_n r_b/\gamma_b\right) < \overline{\lambda}_n \left(1 + \theta_n r_b/\gamma_b\right),$$

which means that while the system is underloaded, (38) does not hold and the profit rate increases in λ_n, so the optimal system is overloaded and prioritizing new customers stricly improves profits. For an overloaded system, we have two cases:

Part 1.(a). If $N\mu \leq \underline{\lambda}_n$, then (41) does not hold and the profit rate increases in λ_n as long as $\rho_n < 1$. As discussed, the profit rate decreases in λ_n when $\rho_n \geq 1$. Thus, $\lambda_n^* = N\mu$, for which $\rho_n = 1$, and the system serves new customers only.

Part 1.(b). If $\underline{\lambda}_n < N\mu < \underline{\lambda}_n \left(1 + \theta_n r_b/\gamma_b\right)$, then (41) holds, $\lambda_n^* = \underline{\lambda}_n$, and the system serves all new but only some base customers.

Part 2. If $V_n - c_n \leq V_b$, then by the discussions following (39) and (42), the profit rate decreases in λ_n for overloaded systems. Therefore, the optimal system is underloaded, profits are independent of the priority policy, and

$$\lambda_n^* = \min\left\{\overline{\lambda}_n, N\mu/\left(1 + \theta_n r_b/\gamma_b\right)\right\}.$$

If $\overline{\lambda}_n \leq N\mu/\left(1 + \theta_n r_b/\gamma_b\right)$, then (38) holds, and $\overline{\lambda}_n$ is the optimal new customer arrival rate. Otherwise the profit is maximized at $\lambda_n^* = N\mu/\left(1 + \theta_n r_b/\gamma_b\right)$, for which $\rho = 1$.

The proof of the case $V_n - c_n > V_b$ and $N\mu \geq \underline{\lambda}_n \left(1 + \theta_n r_b/\gamma_b\right)$ is similar and therefore omitted.■

Proof of Proposition 3. For convenience, let $\Pi^*(N)$ denote the optimal profit rate as a function of capacity, under the jointly optimal priority and promotion policy specified in Proposition 2. Write $\Pi^{*\prime}(N)$ for its first derivative.

Part 1. Note that $V_n - c_n > V_b$ implies $\underline{\lambda}_n > 0$ by (40). First, we characterize $\Pi^*(N)$ and $\Pi^{*\prime}(N)$ as a function of capacity, for four capacity intervals that partition the capacity range. Then, we use these properties to prove the claims.

Interval 1: $N\mu \leq \underline{\lambda}_n$. By Part 1.(a) of Proposition 2 the optimal system is overloaded, prioritizing new customers strictly improves profits, and $\lambda_n^* = N\mu$, so that $\rho_n = 1$. Substituting λ_n^* in the profit rate (30) yields

$$\Pi^*(N) = N\mu\left(V_n - c_n\right) - S\left(N\mu\right) - CN \tag{43}$$

and

$$\Pi^{*\prime}(N) = \mu\left(V_n - c_n - S'\left(N\mu\right) - \frac{C}{\mu}\right). \tag{44}$$

It follows from (40) and (44) that

$$\Pi^{*\prime}(N) = \mu\left(V_b - \frac{C}{\mu}\right) \text{ for } N\mu = \underline{\lambda}_n. \tag{45}$$

Interval 2: $\underline{\lambda}_n < N\mu < \underline{\lambda}_n \left(1 + \theta_n r_b/\gamma_b\right)$. By Part 1.(b) of Proposition 2 the optimal system is overloaded, prioritizing new customers strictly improves profits, and $\lambda_n^* = \underline{\lambda}_n$, so that $\rho_n < 1 < \rho$.

Substituting λ_n^* in (29) yields:

$$\Pi^*(N) = \underline{\lambda}_n (V_n - c_n) + (N\mu - \underline{\lambda}_n) V_b - S(\underline{\lambda}_n) - CN. \tag{46}$$

Note that $\underline{\lambda}_n$, which is defined in (40), is independent of N, so

$$\Pi^{*\prime}(N) = \mu \left(V_b - \frac{C}{\mu} \right). \tag{47}$$

Interval 3: $\underline{\lambda}_n (1 + \theta_n r_b/\gamma_b) \leq N\mu \leq \overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$. Note that $\overline{\lambda}_n > 0$ by (37). By Part 2 of Proposition 2 the optimal system is underloaded and $\lambda_n^* = N\mu/(1 + \theta_n r_b/\gamma_b)$, so that $\rho = 1$. Substituting λ_n^* in (27) yields:

$$\Pi^*(N) = \frac{N\mu}{1 + \theta_n r_b/\gamma_b} \left(V_n - c_n + V_b \theta_n \frac{r_b}{\gamma_b} \right) - S\left(\frac{N\mu}{1 + \theta_n r_b/\gamma_b} \right) - CN \tag{48}$$

and

$$\Pi^{*\prime}(N) = \frac{\mu}{1 + \theta_n r_b/\gamma_b} \left[V_n - c_n + \left(V_b - \frac{C}{\mu} \right) \theta_n \frac{r_b}{\gamma_b} - S^{\prime}\left(\frac{N\mu}{1 + \theta_n r_b/\gamma_b} \right) - \frac{C}{\mu} \right]. \tag{49}$$

It follows from (37) and (49) that

$$\Pi^{*\prime}(N) = -C, \text{ for } N\mu = \overline{\lambda}_n (1 + \theta_n r_b/\gamma_b). \tag{50}$$

Interval 4: $\overline{\lambda}_n (1 + \theta_n r_b/\gamma_b) < N\mu$. By Part 2 of Proposition 2 the optimal system is underloaded and $\lambda_n^* = \overline{\lambda}_n$ (which is defined in (37)), so that $\rho < 1$. Substituting λ_n^* in the profit rate (27) yields:

$$\Pi^*(N) = \overline{\lambda}_n \left(V_n - c_n + V_b \theta_n \frac{r_b}{\gamma_b} \right) - S(\overline{\lambda}_n) - CN \tag{51}$$

and $\Pi^{*\prime}(N) = -C$, because $\overline{\lambda}_n$ is independent of N.

Observe from (43), (46), (48), and (51), that $\Pi^*(N)$ is (i) continuous in N, (ii) strictly concave in N on Interval 1 and Interval 3 (because $S'' > 0$), (iii) linear on Interval 2 and Interval 4, and (iv) strictly decreasing on Interval 4. Therefore, any N^* that satisfies the first order necessary conditions $\Pi^{*\prime}(N^*) = 0$ maximizes Π^*, and by (50)-(51) any such N^* is in the Intervals 1-3.

Part 1.(a) In this case the maximizer N^* is unique and in Interval 1, since Π^* is strictly concave on Interval 1, and because $\Pi^{*\prime}(0) > 0$ and $\Pi^{*\prime}(N) < 0$ for $N\mu = \underline{\lambda}_n$ (the first inequality follows from (44) since $V_n - c_n > \frac{C}{\mu}$ and $S'(0) = 0$, the second inequality holds by (45) since $\frac{C}{\mu} > V_b$). The maximizer N^* in (24) is the unique solution of $\Pi^{*\prime}(N) = 0$ with $\Pi^{*\prime}(N)$ given in (44).

Part 1.(b) In this case any capacity in Interval 2 maximizes Π^*, because Π^* is strictly concave on Intervals 1 and 3, and $\Pi^{*\prime}(N) = 0$ on Interval 2, i.e., for $N\mu \in [\underline{\lambda}_n, \underline{\lambda}_n (1 + \theta_n r_b/\gamma_b)]$, where $\Pi^{*\prime}(N) = 0$ holds by (45) and (47) since $\frac{C}{\mu} = V_b$.

Part 1.(c) In this case the maximizer N^* is unique and in Interval 3, because $\Pi^{*\prime}(N) > 0$ on Interval 2 (by (47) since $\frac{C}{\mu} < V_b$), Π^* is strictly concave on Interval 3, and $\Pi^{*\prime}(N) = -C$, for $N\mu = \overline{\lambda}_n (1 + \theta_n r_b/\gamma_b)$ by (50). The maximizer N^* in (26) is the unique solution of $\Pi^{*\prime}(N) = 0$ with $\Pi^{*\prime}(N)$ given in (49).

Part 1.(d) In this case $V_n - c_n \leq \frac{C}{\mu}$ so that $\Pi^{*\prime}(0) \leq 0$ by (44), i.e., it is unprofitable to operate.

Part 2. First, $V_n - c_n \leq V_b$ implies by Part 2 of Proposition 2 that the optimal system is underloaded, the profit function $\Pi^*(N)$ in (48) holds for $N\mu \leq \overline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, and the one in (51) for $N\mu < \overline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$. By the discussion following (51), $\Pi^*(N)$ is strictly concave in $N\mu \leq \overline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$ and strictly decreasing in $N\mu \geq \overline{\lambda}_n(1 + \theta_n r_b/\gamma_b)$, so $\Pi^*(N)$ has an unique maximizer N^* which is positive if and only if

$$\Pi^{*\prime}(0) > 0 \Leftrightarrow (V_n - c_n + V_b\theta_n r_b/\gamma_b)/(1 + \theta_n r_b/\gamma_b) > \frac{C}{\mu}, \tag{52}$$

where the equivalence holds by (49). Part 2.(a) applies if (52) holds. In this case N^* in (26) is the unique solution of $\Pi^{*\prime}(N) = 0$ with $\Pi^{*\prime}(N)$ given in (49). Part 2.(b) applies if (52) is violated.■

Proof of Corollary 1. Using Proposition 3, we prove Corollary 1 for the case $V_n - c_n - V_b > 0$. The proof for the case $V_n - c_n - V_b \leq 0$ is similar and is omitted. We prove the claim for the three scenarios that correspond to Parts 1.(a)-(c) of Proposition 3 (under the conditions of Part 1.(d) it is not profitable to operate). In this proof only, we write $\Pi(N, \lambda_n)$ to express the dependence on the capacity and the new customer arrival rate. Note that $S(\lambda_n) = \alpha(\lambda_n)^\beta$ and $S'(\lambda_n) = \beta S(\lambda_n)/\lambda_n$.

1.(a) If $V_n - c_n > C/\mu > V_b$, then by Part 1.(a) of Proposition 3, $\lambda_n^* = N^*\mu$, and N^* solves $\Pi^{*\prime}(N) = 0$ with $\Pi^{*\prime}(N)$ given in (44), so that

$$N^* = \frac{1}{\mu} S'^{-1}\left(V_n - c_n - \frac{C}{\mu}\right). \tag{53}$$

Since $\lambda_n^* = N^*\mu$ we have $\rho_n = 1$. Substituting for λ_n^* and N^* in the profit rate (30) yields:

$$\Pi(N^*, \lambda_n^*) = N^*\mu\left(V_n - c_n - \frac{C}{\mu}\right) - S(N^*\mu). \tag{54}$$

Substituting for N^* from (53) into (54) yields

$$\Pi(N^*, \lambda_n^*) = N^*\mu\left(S'(N^*\mu)\right) - S(N^*\mu) = S(N^*\mu)(\beta - 1), \tag{55}$$

where the second equality holds since $S'(\lambda_n) = \beta S(\lambda_n)/\lambda_n$.

1.(b) If $V_b = C/\mu$, then by Part 1.(b) of Proposition 3, $N^*\mu \in [\underline{\lambda}_n, \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)]$ and $\lambda_n^* = \underline{\lambda}_n$. First, note that in the case $N^*\mu = \lambda_n^* = \underline{\lambda}_n$ the ratio of optimal profit to promotion spending must be the same as in Part (a), because the load factor is the same so that the same profit rate function (54) holds, and $S'(\underline{\lambda}_n) = V_n - c_n - V_b = S'(N^*\mu)$, where the first equality holds by (40) and the second since $N^*\mu = \underline{\lambda}_n$. Next, for $N^*\mu \in [\underline{\lambda}_n, \underline{\lambda}_n(1 + \theta_n r_b/\gamma_b)]$ the optimal profit and the optimal new customer arrival rate are constant, so the ratio of optimal profit to promotion spending is constant for any capacity in this range.

1.(c) If $C/\mu < V_b$, then by Part 1.(c) of Proposition 3, $\lambda_n^* = N^*\mu/(1 + \theta_n r_b/\gamma_b)$, and N^* solves $\Pi^{*\prime}(N) = 0$ with $\Pi^{*\prime}(N)$ given in (49), so that

$$N^* = \frac{1}{\mu} S'^{-1}\left(V_n - c_n + V_b\theta_n\frac{r_b}{\gamma_b} - \frac{C}{\mu}\left(1 + \theta_n\frac{r_b}{\gamma_b}\right)\right)\left(1 + \theta_n\frac{r_b}{\gamma_b}\right). \tag{56}$$

Since $\lambda_n^* = N^*\mu/\left(1+\theta_n r_b/\gamma_b\right)$ we have $\rho = 1$. Substituting for λ_n^* and N^* in yields:

$$\Pi\left(N^*, \lambda_n^*\right) = \frac{N^*\mu}{1+\theta_n r_b/\gamma_b}\left(V_n - c_n + V_b\theta_n\frac{r_b}{\gamma_b} - \frac{C}{\mu}\left(1+\theta_n\frac{r_b}{\gamma_b}\right)\right) - S\left(\frac{N^*\mu}{1+\theta_n r_b/\gamma_b}\right). \tag{57}$$

Substituting for N^* from (56) into (57) yields

$$\Pi\left(N^*, \lambda_n^*\right) = \frac{N^*\mu}{1+\theta_n r_b/\gamma_b}S'\left(\frac{N^*\mu}{1+\theta_n r_b/\gamma_b}\right) - S\left(\frac{N^*\mu}{1+\theta_n r_b/\gamma_b}\right) = S\left(\lambda_n^*\right)\left(\beta - 1\right),$$

where the second equality holds since $S'\left(\lambda_n\right) = \beta S\left(\lambda_n\right)/\lambda_n$ and $\lambda_n^* = N^*\mu/\left(1+\theta_n r_b/\gamma_b\right)$.■

Proof of Corollary 2. By the definitions of V_b and V_n in (14) and (16), respectively, we have

$$V_n - V_b = \left(p_n + c_n\right) - \left(p_b + c_b\right) + \left(\theta_n + \theta_b - 1\right)\frac{R - r_b c_b}{\gamma_b + r_b\left(1-\theta_b\right)}. \tag{58}$$

Part 1. It is clear by inspection of (58) that $V_n - V_b$ increases in p_n and c_n and decreases in p_b. Write $\partial f/\partial x$ for the partial derivative of a function f with respect to x. We have

$$\frac{\partial\left(V_n - V_b\right)}{\partial c_b} = -1 - \frac{r_b\left(\theta_n + \theta_b - 1\right)}{\gamma_b + r_b\left(1-\theta_b\right)} = -\frac{\gamma_b + r_b\theta_n}{\gamma_b + r_b\left(1-\theta_b\right)} < 0.$$

Part 2. It is clear by inspection of (58) that $V_n - V_b$ increases (decreases) in θ_n if $R > r_b c_b$ ($R < r_b c_b$), and that $V_n - V_b$ is constant in θ_n and θ_b if $R = r_b c_b$. That $V_n - V_b$ also increases (decreases) in θ_b if $R > r_b c_b$ ($R < r_b c_b$) follows because from (58) we have

$$\frac{\partial\left(V_n - V_b\right)}{\partial\theta_b} = \frac{R - r_b c_b}{\gamma_b + r_b\left(1-\theta_b\right)}\left(1 + \frac{r_b\left(\theta_n + \theta_b - 1\right)}{\gamma_b + r_b\left(1-\theta_b\right)}\right)$$

and the term in brackets is strictly positive.

Part 3. It is clear by (58) that $V_n - V_b$ increases (decreases) in R if $\theta_n + \theta_b - 1 > 0$ ($\theta_n + \theta_b - 1 < 0$), and that $V_n - V_b$ is constant in R and r_b if $\theta_n + \theta_b - 1 = 0$. That $V_n - V_b$ decreases (increases) in r_b if $\theta_n + \theta_b - 1 > 0$ ($\theta_n + \theta_b - 1 < 0$) follows because from (58) we have

$$\begin{aligned}\frac{\partial\left(V_n - V_b\right)}{\partial r_b} &= \left(\theta_n + \theta_b - 1\right)\left(-\frac{c_b}{\gamma_b + r_b\left(1-\theta_b\right)} - \frac{\left(R - r_b c_b\right)\left(1-\theta_b\right)}{\left(\gamma_b + r_b\left(1-\theta_b\right)\right)^2}\right) \\ &= -\left(\theta_n + \theta_b - 1\right)\frac{c_b\gamma_b + R\left(1-\theta_b\right)}{\left(\gamma_b + r_b\left(1-\theta_b\right)\right)^2}.\end{aligned}$$

Part 4. It is clear by (58) that $V_n - V_b$ decreases (increases) in γ_b if $\left(\theta_n + \theta_b - 1\right)\left(R - r_b c_b\right) > (<)\, 0$ and is constant in γ_b otherwise.■