

A Time-Varying Call Center Design via Lagrangian Mechanics *

Robert C. Hampshire
Heinz School of Public Policy and Management
Carnegie Mellon University, hamp@andrew.cmu.edu

Otis B. Jennings
Fuqua School of Business
Duke University, otisj@duke.edu

William A. Massey
Department of Operations Research and Financial Engineering
Princeton University, wmassey@princeton.edu

January 24, 2008

Abstract

We consider a multi-server delay queue with finite additional waiting spaces and time-varying arrival rates, where the customers waiting in the buffer may abandon. These are features that arise naturally from the study of service systems such as call centers. Moreover, we assume rewards for successful service completions and cost rates for service resources. Finally, we consider service level agreements that constrain the fractions of arriving customers that abandon as well as the ones that are blocked.

Applying the theory of Lagrangian mechanics to the fluid limit of a related Markovian service network model, we obtain near profit-optimal staffing and provisioning schedules. The nature of this solution consists of three modes of operation. A key step in deriving this solution is combining the modified offered load approximation for loss systems with our fluid model to estimate effectively both our service level agreement metrics and the profit for the original queueing model. Second-order profit improvements are achieved through a modified offered load version of the conventional square root rule.

Keywords: Asymptotic Analysis, Call Centers, Calculus of Variations, Delay Models, Dynamical Systems, Lagrangian, Multi-Server Queues, Optimal Control, Server Staffing.

*To appear in a special issue of *Probability in the Engineering and Informational Sciences* on the Analysis and Control of Queues in Manufacturing and Service Systems. First and third authors supported by NSF Grant DMI-0323668.

1 Introduction

According to Koole and Mandelbaum [23], 60 to 70 percent of the total costs for operating a call center involve wage and benefit expenses for personnel. It follows that determining the optimal amount of call center agents is of great interest to call center managers. Other features – such as facilities, equipment, maintenance, telephony access resources, and power consumption – contribute to the call center’s bottom line as well, and thus warrant thoughtful design and control. In this paper, we couple the decision of how to dynamically staff call center operators with the decision of how to control another potentially dynamic feature, buffer capacity, or more explicitly, the number of telephone lines dedicated to callers of a given type. Both of these features affect quality of service.

We consider a multi-server delay queue with finite additional waiting spaces and time-varying arrival rates, where the customers waiting in the buffer may abandon. We assume the call center manager receives payments for each successful service completion and incurs costs both for employing operators as well as for allocating telephone line resources. The performance of the call center is evaluated using two service level metrics: the fraction of callers who abandon and the fraction of callers who are blocked. Constraints, or *service level agreements* (SLA), for both of the metrics are imposed. Ultimately, we seek to develop a staffing and buffer capacity schedule that maximizes profits, while conforming to the SLA.

Determining a dynamic staffing schedule is a generally accepted undertaking and is extensively explored in the literature; see, e.g., Bhandari et. al [5], Jennings et al. [22], Green et al. [12] or Fieldman et al. [9]. Determining buffer capacity has been studied as well; see Harris et al. [20], Whitaker [30] as well as Wallace [29] and [27]. Each of these last four papers sets the buffer level to the minimum amount necessary to meet a given level of service. However, the papers are restricted to steady state analysis in an environment with stationary demand and non-adaptive, static buffer capacity. To the best of our knowledge, ours is the first paper to consider dynamic buffer capacity in a call center context. This paper is based on work found in Hampshire [17] and is an extension of work found in Hampshire and Massey [18].

As for why one might have time-varying buffer capacity, consider a call center serving multiple business lines, each with its own source of callers that effectively compete for service resources. One can envision a system so backlogged that buffer space is at or near capacity. Admitting a call of one type might inhibit one’s ability to accommodate others. Our approach to capturing the externality caused by allocating buffer capacity for a given caller type is to charge an internal cost; one can think of this as an opportunity cost.

Once we formulate our queueing model, we modify it slightly so that it conforms to a large class of queueing network models identified by Mandelbaum, Massey and Reiman [26] and referred to as *Markovian service networks* (MSN). This class of network models capture many important call center features such as time varying arrival rates, multi-server queues, service abandonments, as well as network routing due to service completions or service abandonments. Inspired by growing a business to match a corresponding growth in customer demand, we scale our MSN so that it converges to a deterministic “fluid” model that is a dynamical system. Its time evolution is governed by a set of non-linear differential equations.

Given the managerial economic structure of our queueing model, we can express the total

profit for the call center as an integral functional of the number of customers in the system over a fixed time interval. We call this our *profit functional*. Applying the cost structure to our fluid model yields a fluid approximation of the profit functional. Since the fluid model for our queueing system process is a *dynamical system*, we can use variational calculus methods from the theory of optimal control, see Gregory and Lin [13], to derive a staffing and provisioning schedule by analyzing the fluid approximation of the profit functional. From this variational analysis, we show that a penalty model can be used to analyze the typical performance metrics found in SLA for the design of call centers with only rewards.

Mathematically, our fluid model analysis augments the appropriate set of multipliers and state variables to construct a Lagrangian. We can then invoke the Euler-Lagrange equations to find the equations that determine a fluid optimal solution. For example, the multipliers for the Lagrangian simultaneously have the call center interpretation of an opportunity cost per admitted customer and the classical mechanical interpretation of generalized momentum. Additional references to variational calculus and its applications to mechanics can be found in books by Bryson and Ho [7] as well as Lanczos [24].

The resulting fluid optimal staffing and provisioning policy reduces the queueing dynamics to three modes of operations that are each loss systems with the same time-varying arrival rate. Our fluid model is then refined by the modified offered load approximation, allowing it to effectively capture the average number of customers who abandon and the probability of blocking. The latter is a quantity that is not typically captured by a fluid model. The profit is increased by augmenting the fluid level staffing by an amount proportional to the square root of the system load. The modified fluid model is shown to be a good approximation of the mean behavior for the original finite buffer call center model.

Below, we list the main contributions of this work:

1. Making a connection between call center operations and classical mechanics that furthers the analysis of our staffing and provisioning problem.
2. Developing a fluid optimal schedule through Lagrangian methods that equate the SLA metrics to the application of penalties for lost service.
3. Identifying three primary modes of operation for our call center:
 - (a) Agent staffing with no buffer (agent mode),
 - (b) Buffer provisioning with no agents (music mode),
 - (c) No agents or buffer (busy signal mode).
4. Refining the fluid staffing schedule by employing offered load approximations such as square root staffing and the Erlang blocking formula to satisfy the SLA metrics.

Within the steadily growing call center literature, our paper fits into the single customer class/single operator type category. Other recent papers within this category include Armony et al. [1], Baron and Milner [2], Garnett et al. [11], and Zeltyn and Mandelbaum [31], [32]. Each of these papers studies systems with customer abandonment. A fluid model approach is used in [1] and the exponential abandonment assumption is relaxed in [32]. For more on contact centers see Brown et al. [6] and Gans et al. [10]

A recent call center paper with multiple customer classes is by Gurvich and Whitt [14]. In this paper, staffing with moderate cross training is set so that each customer achieves class-specific service levels through “fixed-queue-ratio” routing. Bassamboo et al. [3] combine multiple customer class with multiple server pools in a model with doubly stochastic arrival processes. In this paper, static staffing levels are set that balance personnel costs with abandonment penalties.

In Section 2, we present our call center model as a finite buffer, multiserver queue with abandonment. In Section 3, we present the basic results for Markovian service networks and create a Markovian service network model of our call center to reformulate the corresponding optimization problem. We do this by developing a fluid limit for the MSN model and present a fluid optimization problem. In Section 4, we give a Lagrangian formulation using penalty costs that leads to our optimal fluid model analysis and policy. Additionally, we present a refinement of the fluid model using modified offered load approximations. In Section 5, we numerically implement our provisioning and staffing algorithms based on the analysis in the previous section. We analyze an example with a non-linear cost structure and a time dependent customer arrival rate. After the series of approximations, we apply our optimal schedule to the original call center model. Finally, we summarize our results in Section 6.

2 Call Center Model

In this section we formulate a finite buffer call center model and present its corresponding agent and telephone line optimization problem. Our stochastic call center model is constructed by defining a set of parameters that are related to the dynamics of the call center. First, let $\lambda(t)$ equal the *arrival rate* of a nonhomogeneous Poisson process at time t . The constant μ is the common rate for the exponential distributions of the random *service times* for all the customers which we assume to be mutually independent. The constant β is the rate for the exponential distributions of the random *abandonment times* for all the customers which we assume to be mutually independent. The non-negative integer $L(t)$ equals the *number of call center agents* at time t . We assume that the number of available telephone lines exceeds the number of agents. Hence, we let the non-negative integer $K(t)$ equal the *number of additional lines* at time t .

Many probabilistic interpretations follow from these assumptions. The integral $\int_s^t \lambda(u) du$ equals the average number of arriving customers (incoming telephone calls) during the time interval $(s, t]$. Moreover, $1/\mu$ equals the average service time that customers spend listening (and talking) to an agent. Finally, $1/\beta$ equals the average time customers are willing to wait for an agent while they are listening to “music” before leaving the system.

We assume that both $K + L \equiv \{K(t) + L(t) \mid 0 \leq t \leq T\}$, the provisioning schedule for the call center telephone lines, and $L \equiv \{L(t) \mid 0 \leq t \leq T\}$, the staffing schedule for call center agents have a finite number of jumps in bounded intervals as a function of time. We also assume that the forecasted arrival rate $\lambda \equiv \{\lambda(t) \mid t \geq 0\}$ for customer demand is non-negative, real-valued and locally integrable. We then define $\{Q_{L/K}(t) \mid t \geq 0\}$ to be the number of customers in this $M_t/M_t/L_t/K_t$ queue with abandonment (see Figure 1).

Now we impose performance structure on our call center queueing model where we specify the following two *service level agreement* (SLA) targets:

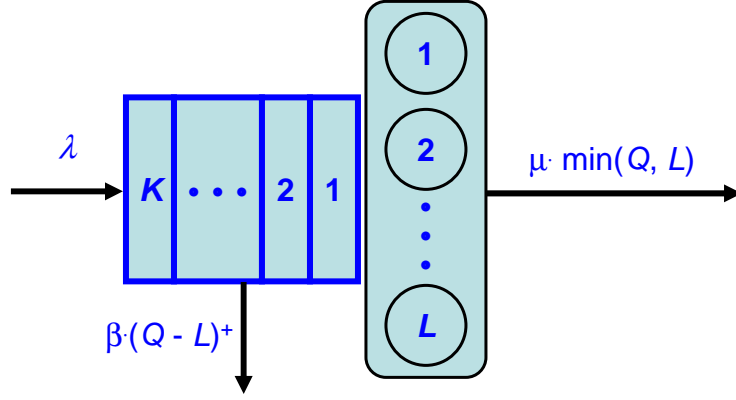


Figure 1: Queueing Model of Call Center

ϵ_a = maximum fraction of the mean number of customers who abandon,
 ϵ_b = maximum fraction of the mean number of customers who are blocked.

Moreover, we add the following revenue and cost structure:

r = service completion revenue per customer,
 $c(L)$ = total staffing cost rate for L agents,
 $d(K + L)$ = total provisioning cost rate for $K + L$ telephone lines.

Our fundamental optimization problem can now be stated:

Optimization Problem 2.1 (Blocking Scheduling) *Find provisioning and staffing schedules K and L such that we*

$$\text{maximize } \int_0^T r\mu \cdot \mathbf{E} [\min(Q_{L/K}(t), L(t))] - c(L(t)) - d((K + L)(t)) dt \quad (2.1)$$

subject to the abandonment constraint of

$$\int_0^T \beta \cdot \mathbf{E} [(Q_{L/K}(t) - L(t))^+] dt \leq \mathcal{E}_a, \quad (2.2)$$

and the blocking constraint of

$$\int_0^T \lambda(t) \cdot \mathbf{P} \{Q_{L/K}(t) = K(t) + L(t)\} dt \leq \mathcal{E}_b, \quad (2.3)$$

where

$$\mathcal{E}_a \equiv \epsilon_a \cdot \left(Q_{L/K}(0) + \int_0^T \lambda(t) dt \right) \text{ and } \mathcal{E}_b \equiv \epsilon_b \cdot \left(Q_{L/K}(0) + \int_0^T \lambda(t) dt \right). \quad (2.4)$$

The integral of the first non-negative term in (2.1) is the expected revenue and we are maximizing the average profit subject to expected service level constraints.

We express our near optimal scheduling algorithm in terms of a solution to a set of *competing Lagrangian equations* which we now define. Given positive constants σ and τ , we say that the deterministic processes $p = \{p(t) \mid 0 \leq t \leq T\}$ and $q = \{q(t) \mid 0 \leq t \leq T\}$ solve the following set of differential equations:

1. If $\ell_1(p(t), q(t)) \geq \max(\ell_2(p(t), q(t)), \ell_3(p(t), q(t)))$ holds, then

$$\frac{d}{dt}p(t) = (p(t) - \tau)\gamma \quad \text{and} \quad \frac{d}{dt}q(t) = \lambda(t) - \gamma q(t). \quad (2.5)$$

2. If $\ell_2(p(t), q(t)) \geq \max(\ell_1(p(t), q(t)), \ell_3(p(t), q(t)))$ holds, then

$$\frac{d}{dt}p(t) = (p(t) - \sigma)\beta - d'(q(t)) \quad \text{and} \quad \frac{d}{dt}q(t) = \lambda(t) - \beta q(t). \quad (2.6)$$

3. If $\ell_3(p(t), q(t)) \geq \max(\ell_1(p(t), q(t)), \ell_2(p(t), q(t)))$ holds, then

$$\frac{d}{dt}p(t) = (p(t) + r)\mu - (c + d)'(q(t)) \quad \text{and} \quad \frac{d}{dt}q(t) = \lambda - \mu q(t). \quad (2.7)$$

Moreover, $\ell_1(p(t), q(t))$, $\ell_2(p(t), q(t))$ and $\ell_3(p(t), q(t))$ are defined to be

$$\ell_1(p(t), q(t)) \equiv (p(t) - \tau)\gamma q(t) - c(0) - d(0), \quad (2.8)$$

$$\ell_2(p(t), q(t)) \equiv (p(t) - \sigma)\beta q(t) - c(0) - d(q(t)) \quad (2.9)$$

$$\ell_3(p(t), q(t)) \equiv (p(t) + r)\mu q(t) - c(q(t)) - d(q(t)). \quad (2.10)$$

Finally, p and q are uniquely determined when we are given the value of $q(0)$ and set $p(T) = 0$.

Our algorithm is then to find positive constants σ and τ such that

$$\int_{\{\ell_2(p(t), q(t)) > \max(\ell_1(p(t), q(t)), \ell_3(p(t), q(t)))\}} \beta \cdot q(t) \cdot (1 - b(0, q(t))) dt = \mathcal{E}_a \quad (2.11)$$

and

$$\begin{aligned} & \int_{\{\ell_1(p(t), q(t)) > \max(\ell_2(p(t), q(t)), \ell_3(p(t), q(t)))\}} \lambda(t) dt + \int_{\{\ell_2(p(t), q(t)) > \max(\ell_1(p(t), q(t)), \ell_3(p(t), q(t)))\}} \lambda(t) \cdot b(0, q(t)) dt \\ & + \int_{\{\ell_3(p(t), q(t)) > \max(\ell_1(p(t), q(t)), \ell_2(p(t), q(t)))\}} \lambda \cdot b(\chi(t), q(t)) dt = \mathcal{E}_b \end{aligned} \quad (2.12)$$

where $b(\cdot, \cdot)$ is defined to be

$$b(a, z) \equiv \frac{z^{\lceil z+a\sqrt{z} \rceil}}{\lceil z+a\sqrt{z} \rceil!} \bigg/ \sum_{i=0}^{\lceil z+a\sqrt{z} \rceil} \frac{z^i}{i!}, \quad (2.13)$$

$\chi(t)$ is the smallest positive number such that

$$\begin{aligned} & \frac{r\lambda(t)}{\sqrt{q(t)}} \cdot \frac{\phi(\chi(t))}{\Phi(\chi(t))} + (c+d) \left(q(t) + \chi(t) \sqrt{q(t)} \right) \\ &= \min_{a \geq 0} \left(\frac{r\lambda(t)}{\sqrt{q(t)}} \cdot \frac{\phi(a)}{\Phi(a)} + (c+d) \left(q(t) + a \sqrt{q(t)} \right) \right) \end{aligned} \quad (2.14)$$

with

$$\phi(x) \equiv \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \quad \text{and} \quad \Phi(x) \equiv \int_{-\infty}^x \phi(y) dy. \quad (2.15)$$

In Section 3, we show how the blocking features of our call center model can be approximated by impatient customers who experience a *fast abandonment*. This creates a new call center model that belongs to the family of Markovian service networks established in Mandelbaum, Massey and Reiman [26]. The appeal of these queueing networks is that they scale asymptotically to deterministic dynamical systems that we call *fluid models*. In Section 4, we show how a dynamic optimization of the fluid model leads to an equivalent formulation where the SLA constraints are replaced by penalties for each customer lost to regular or fast abandonment. Moreover, this fluid model analysis suggest optimal modes of operation for the queueing model that correspond to the dynamics of the Erlang loss model. In turn we apply various modified offered load approximations to transform the fluid model results into a useful approximation of the SLA blocking probabilities. The end result is a near optimal schedule for the original call center model that captures the SLA constraints.

3 Developing the Fluid Model

Now we modify our stochastic call center model to place it into a larger family of queueing models that have simple scaling properties. These asymptotic results form the basis of our optimal scheduling analysis. As illustrated in Figure 2, we use the concept of fast abandonment to approximate the call center model with blocking by a queueing model that has a strong law of large numbers limit theorem. Moreover, the limit of this scaled process is a deterministic dynamical system the we call our fluid model.

3.1 Fast Abandonment

In the current model, jobs are blocked when they arrive to a system that currently has $K(t) + L(t)$ jobs present. In our replacement model, such jobs are not blocked. Instead, we have an infinite buffer where those behind the $K(t)$ -th job in the buffer (which only happens when an additional $L(t)$ customers are in service) abandon at rate γ . See Figure 3 for a comparison of the two associated Markov chains. The top state-transition diagram is the one for the original call center model. The bottom diagram describes our replacement model.

Any jobs that arrives to the system with $K(t) + L(t)$ jobs already present and abandons before any other service or regular abandonment (at rate β) event occurs, has the same ultimate fate of being lost to the systems as a blocked customer for the blocked model. The larger the value of γ , the fewer of these jobs stay in the buffer long enough to reach the



Figure 2: Overview of Queueing Model Approximations

$K(t) + L(t)$ -th queueing location or lower. This is due to the fact that the excess number of customers beyond $K(t) + L(t)$ behaves stochastically like an infinite server queue, until a service or regular abandonment occurs, with arrival rate $\lambda(t)$ and service rate γ . Hence, we assume that γ is large relative to the other transition rates and refer to this phenomenon as “fast abandonment”, where γ equals the exponential fast abandonment rate. This modeling approach is effective, regardless of the size of the system.

The corresponding queueing system process $Q \equiv \{ Q(t) \mid t \geq 0 \}$ is defined by the following implicit equation:

$$\begin{aligned}
 Q(t) = & Q(0) + \Pi_1 \left(\int_0^t \lambda(s) ds \right) - \Pi_2 \left(\int_0^t \mu \cdot \min(Q(s), L(s)) ds \right) \\
 & - \Pi_3 \left(\int_0^t \beta \cdot ((Q(s) - L(s))^+ - (Q(s) - K(s) - L(s))^+) ds \right) \\
 & - \Pi_4 \left(\int_0^t \gamma \cdot (Q(s) - K(s) - L(s))^+ ds \right), \tag{3.1}
 \end{aligned}$$

where $\Pi_i \equiv \{ \Pi_i(t) \mid t \geq 0 \}$ for $i = 1, 2, 3, 4$ are a collection of independent and identically distributed, standard (rate 1) Poisson processes. The resulting queueing process is Markovian. It is the special case of a *Markovian service network* as defined in Mandelbaum, Massey and Reiman [26]. This paper also addresses the issues of existence and uniqueness of Markov processes defined by implicit equations such as (3.1).

To size γ , let the time for one of the L busy servers to be free is $\min(Y_1, \dots, Y_L)$, where the Y_i 's are a collection of i.i.d. random variables having the same distribution as Y and L is some generic positive integer. The time for one of the K occupied to be free due to abandonment is $\min(Z_1, \dots, Z_K)$, where the Z_i 's are a collection of i.i.d. random variables having the same distribution as Z , they are all independent of the Y_i 's, and K is some other generic integer. Therefore, we want to select γ such that

$$\mathbb{P}(X > \min(Y_1, \dots, Y_L, Z_1, \dots, Z_K)) \leq \epsilon, \tag{3.2}$$

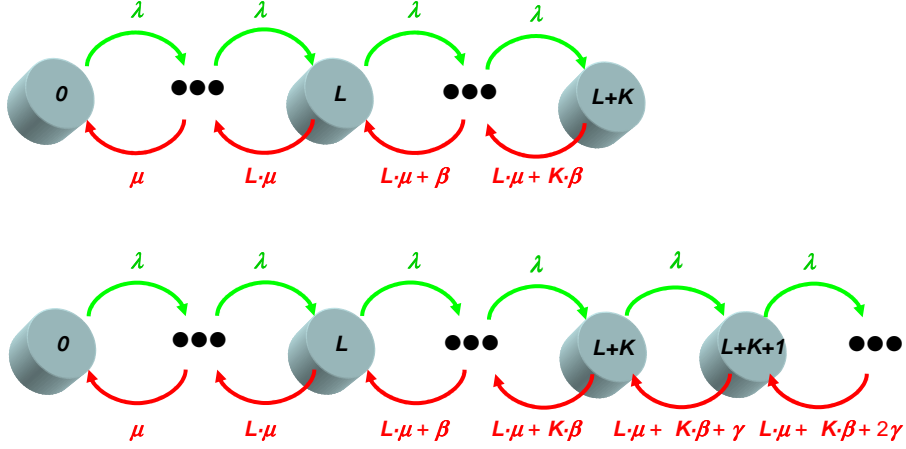


Figure 3: Markov Chain Comparison for the Fast Abandonment Assumption

where ϵ is a small positive number. Now $\min(Y_1, \dots, Y_L, Z_1, \dots, Z_K)$ is exponentially distributed with rate $L\mu + K\beta$, so we have

$$P(X > \min(Y_1, \dots, Y_L, Z_1, \dots, Z_K)) = \frac{L\mu + K\beta}{L\mu + K\beta + \gamma}. \quad (3.3)$$

This ratio is less than ϵ whenever $\gamma > (L\mu + K\beta) \cdot (1/\epsilon - 1)$. A simple rule of thumb is then to set $\gamma = (L\mu + K\beta)/\epsilon$. If the probability of blocking is already low for a specific system, then setting $\epsilon = 0.1$ is more than sufficient. Our rule of thumb then becomes $\gamma = 10.0 \cdot (L\mu + K\beta)$.

3.2 Uniform Acceleration

By appealing to the strong law of large numbers, we construct a deterministic approximation for the sample path (and mean) behavior of any Markovian service network. Suppose that we construct the *uniformly accelerated* version (see Mandelbaum, Massey and Reiman [26]) of the model (3.1) with scale factor $\eta > 0$, such that

$$\begin{aligned} Q^\eta(t) = & Q^\eta(0) + \Pi_1 \left(\int_0^t \eta \lambda(s) ds \right) - \Pi_2 \left(\int_0^t \mu \cdot \min(Q(s), \eta \cdot L(s)) ds \right) \\ & - \Pi_3 \left(\int_0^t \beta \cdot ((Q^\eta(s) - \eta \cdot L(s))^+ - (Q^\eta(s) - \eta \cdot K(s) - \eta \cdot L(s))^+) ds \right) \\ & - \Pi_4 \left(\int_0^t \gamma \cdot (Q^\eta(s) - \eta \cdot K(s) - \eta \cdot L(s))^+ ds \right), \end{aligned} \quad (3.4)$$

and consider the pointwise limit of $Q^\eta \equiv \{Q^\eta(t) \mid t \geq 0\}$ as $\eta \rightarrow \infty$.

In the context of call centers, we motivate this asymptotic scaling by considering the expansion of a business in response to growing customer demand. The “size” of this call center business is given by the number of call center agents $L(t)$ and telephone lines $K(t) + L(t)$. Similarly the “size” of the aggregate customer demand is given by the arrival rate $\lambda(t)$. The service rate and abandonment rates μ and β correspond to personal decisions made by

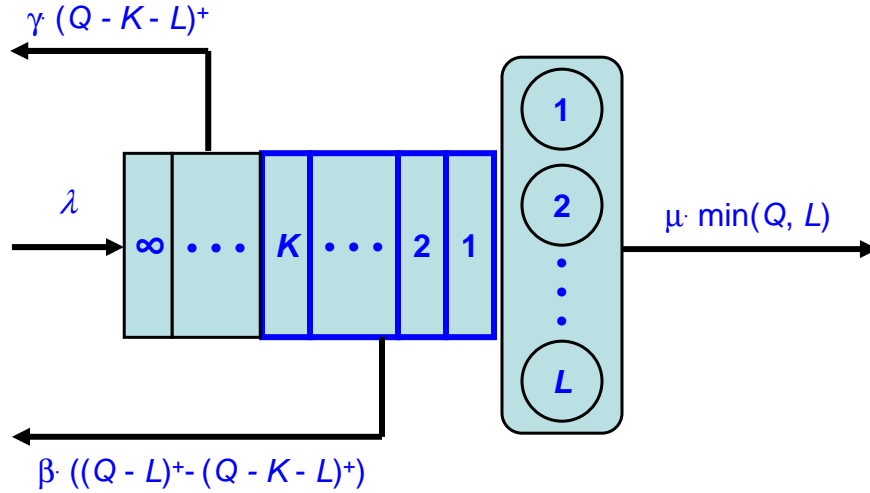


Figure 4: Markovian Service Network Model of Call Center with Fast Abandonment

individual customers and agents that are independent of the total size of customer demand or the total size of the call center. This follows from the fact that the typical customer is unaware of both of these dimensions for the call center. Based on these assumptions, it is reasonable to scale the parameters $\lambda(t)$, $K(t)$ and $L(t)$ upwards by some positive factor η , but *not* to scale μ , β or γ .

From the general theory for Markovian service networks (see [26]), it follows that whenever we have $\lim_{\eta \rightarrow \infty} Q^\eta(0)/\eta = Q(0)$, there is a deterministic process $q \equiv \{q(t) \mid t \geq 0\}$ such that

$$\lim_{\eta \rightarrow \infty} \sup_{0 \leq t \leq T} \left| \frac{1}{\eta} Q^\eta(t) - q(t) \right| = 0 \quad \text{a.s.} \quad (3.5)$$

Moreover, this process q is a *dynamical system* that is governed by the differential equation

$$\begin{aligned} \frac{d}{dt} q(t) &= \lambda(t) - \beta \cdot ((q(t) - L(t))^+ - (q(t) - K(t) - L(t))^+) \\ &\quad - \gamma \cdot (q(t) - K(t) - L(t))^+ - \mu \cdot \min(q(t), L(t)). \end{aligned} \quad (3.6)$$

We refer to q as the *fluid model* for the call center. We now study q to construct an optimal staffing and provisioning schedule. Henceforth we refer to this procedure simply as *fluid scheduling*.

4 Fluid Scheduling with Offered Load Refinements

In this section we construct our near-optimal scheduling algorithm for the blocking models as follows:

1. Lagrangian formulation of fluid scheduling with SLA targets.
2. Equivalence to Lagrangian formulation of fluid scheduling with penalties.
3. Concavity cost assumptions to solve fluid scheduling problem via competing Lagrangians.
4. Using modified offered load estimates to refine the fluid schedule analysis and estimate the blocking scheduling problem.

In Section 4.1, we state the analogous scheduling problem for the fluid model with SLA targets. We then use variational calculus to transform this problem into the analysis of an associated Lagrangian. In Section 4.2, we show that the Lagrangian analysis of this fluid scheduling problem is equivalent to one with penalties followed by a calibration of them to the SLA targets. The penalties capture the cost of customer departures without receiving service due to the blocking or abandonment. In Section 4.3, we simplify the analysis of this fluid scheduling problem with penalties by assuming our cost functions c and d are increasing and concave. This allows us to solve this problem in terms of what we call competing Lagrangians. In turn this analysis suggests that the queueing system makes transitions between discrete operational modes. For our call center blocking model, these modes corresponds to the dynamics of an Erlang loss model. Finally, in Section 4.4, we combine this insight with modified offered load approximation methods for the Erlang loss model. This allows us to refine our fluid approximations and formulate a near-optimal schedule that allows the blocking model to attain critical SLA targets such as blocking probabilities. These are quantities that are typically beyond the reach of fluid models. The end result of our analysis is a new technique that we call the *fluid modified offered load (FMOL)* approximation.

4.1 Lagrangian Formulation with SLA Targets

Now we state a fluid version of the scheduling problem where the analogues to (2.2) and (2.3) are replaced by integral equality constraints. For this to be possible we must have $\epsilon_a + \epsilon_b < 1$. For this section and the next two, we simplify our notation by suppressing the time dependence, i.e. use q instead of $q(t)$. When taking time derivatives, we denote that as \dot{q} instead of $\frac{d}{dt}q(t)$. The exceptions are parameters β, γ, μ and r which are always constants.

Optimization Problem 4.1 (Fluid Scheduling with SLA Targets) *Find K and L so that we*

$$\text{maximize } \int_0^T r\mu \cdot (q \wedge L) - c(L) - d(K + L) dt, \quad (4.1)$$

with the control constraint:

$$\dot{q} = \lambda - \beta \cdot ((q - L)^+ - (q - K - L)^+) - \gamma \cdot (q - K - L)^+ - \mu \cdot (q \wedge L) \quad (4.2)$$

and integral equality constraints

$$\int_0^T \beta \cdot ((q - L)^+ - (q - K - L)^+) = \mathcal{E}_a \quad \text{and} \quad \int_0^T \gamma \cdot (q - K - L)^+ = \mathcal{E}_b. \quad (4.3)$$

Now, our integral constraints become *isoperimetric* (see Bryson and Ho [7]). Including them as equality constraints involves adding two new state variables x and y and two new Lagrange multipliers σ and τ . This problem is equivalent to finding K and L so that we

$$\text{maximize } \int_0^T \widehat{\mathcal{L}}(K, L, p, q, \dot{q}, \sigma, \tau, x, \dot{x}, y, \dot{y}) dt, \quad (4.4)$$

where

$$\begin{aligned} \widehat{\mathcal{L}}(K, L, p, q, \dot{q}, \sigma, \tau, x, \dot{x}, y, \dot{y}) &= r\mu \cdot (q \wedge L) - c(L) - d(K + L) \\ &\quad + p \cdot \left\{ \dot{q} - \lambda + \beta \cdot ((q - L)^+ - (q - K - L)^+) \right. \\ &\quad \left. + \gamma \cdot (q - K - L)^+ + \mu \cdot (q \wedge L) \right\} \\ &\quad + \sigma \cdot \left(\dot{x} - \beta \cdot ((q - L)^+ - (q - K - L)^+) \right) \\ &\quad + \tau \cdot \left(\dot{y} - \gamma \cdot (q - K - L)^+ \right). \end{aligned} \quad (4.5)$$

with $x(0) = y(0) = 0$, $x(T) = \mathcal{E}_a$, and $y(T) = \mathcal{E}_b$.

What follows from the Euler-Lagrange equations for x and y (see Lanczos [24]) is that σ and τ are constants, since

$$\frac{d}{dt} \frac{\partial \widehat{\mathcal{L}}}{\partial \dot{x}} = \frac{\partial \widehat{\mathcal{L}}}{\partial x} = 0 \quad \Rightarrow \quad \dot{\sigma} = 0 \quad \text{and} \quad \frac{d}{dt} \frac{\partial \widehat{\mathcal{L}}}{\partial \dot{y}} = \frac{\partial \widehat{\mathcal{L}}}{\partial y} = 0 \quad \Rightarrow \quad \dot{\tau} = 0. \quad (4.6)$$

In classical mechanics, x and y are considered *position variables* with σ and τ as their corresponding *momentum variables*. These conditions imply that we have a *conservation of momentum* principle. In the next section we show how the SLA targets can be replaced by an appropriate selection of values for the constants σ and τ .

4.2 Equivalent Formulation with Penalties

Since σ and τ can be viewed as constants, then any solution to Optimization Problem 4.1 can also be used to find schedules K and L such that we

$$\text{maximize } \int_0^T \mathcal{L}(K, L, p, q, \dot{q}) dt, \quad (4.7)$$

where

$$\begin{aligned} \mathcal{L}(K, L, p, q, \dot{q}) &\equiv r\mu \cdot (q \wedge L) - \sigma \beta \cdot ((q - L)^+ - (q - K - L)^+) \\ &\quad - \tau \gamma \cdot (q - K - L)^+ - c(L) - d(K + L) \\ &\quad + p \cdot \left\{ \dot{q} - \lambda + \beta \cdot ((q - L)^+ - (q - K - L)^+) \right. \\ &\quad \left. + \gamma \cdot (q - K - L)^+ + \mu \cdot (q \wedge L) \right\}, \end{aligned} \quad (4.8)$$

provided that the isoperimetric constraints of (4.3) are satisfied. This is a Lagrangian formulation of the following equivalent optimization program:

Optimization Problem 4.2 (Fluid Scheduling with Penalties) Find K and L so that we

$$\begin{aligned} \text{maximize } \int_0^T & r\mu \cdot (q \wedge L) - \sigma\beta \cdot ((q - L)^+ - (q - K - L)^+) \\ & - \tau\gamma \cdot (q - K - L)^+ - c(L) - d(K + L) dt, \end{aligned} \quad (4.9)$$

with the control constraint:

$$\dot{q} = \lambda - \beta \cdot ((q - L)^+ - (q - K - L)^+) - \gamma \cdot (q - K - L)^+ - \mu \cdot (q \wedge L). \quad (4.10)$$

We then solve for the fluid schedule by finding positive constants σ and τ such that the isoperimetric constraints of (4.3) are satisfied. Thus we reinterpret our two constants as follows:

σ = music abandonment penalty per customer (those above $L(t)$ but below $K(t) + L(t)$),

τ = busy signal abandonment penalty per customer (those above $K(t) + L(t)$).

The search for the appropriate σ and τ is referred to in this paper as the *calibration* of the penalties to the SLA targets.

In the language of classical mechanics (see Lanczos [24]), our profit integral is called the *action*. Our Lagrangian has the units of *energy* for a physical system and plays the role of a *profit rate* for our call center model. The fluid approximation q for the total number of customers in the call center plays the classical mechanical role of the *position* variable. The multiplier p is the *generalized momentum* variable. For our call center fluid model it has the economic interpretation of the *opportunity cost per customer*. It measures of the impact of an additional customer joining the system on the total profit. Customers may influence the total profit in three ways. First, the call center may incur a penalty when a customer abandons due to having no agents or telephone lines available. Second, the call center incurs a penalty when a customer abandons due to having no agents available. Finally, the call center receives revenue only when a customer departs due to a service completion.

Approaching the end of the time interval T , an additional arriving customer has little effect on the total profit. The late arriving customer neither has time to abandon nor time to depart due to a service completion. Thus the Lagrange multiplier process $\{p(t) \mid 0 \leq t \leq T\}$ has the terminal condition of $p(T) = 0$. We assume that no penalty is assessed for customers remaining in the queue after time T . Customers remaining in the system after this time become the initial load for the next planning interval.

Having established that the fluid scheduling problem with SLA targets is equivalent to a calibrated fluid scheduling problem with penalties, it remains to solve either problem. In the next section, we formulate a solution in terms of competing Lagrangians for the fluid problem with penalties when we assume that our cost functions c and d are increasing and concave. We use the word “increasing” instead of the more commonly used phrase “non-decreasing” to include functions with zero derivatives.

Mechanics	Call Center Operations
position	number of customers in service (q)
velocity	customer flow rate (\dot{q})
Lagrangian	value rate ($\mathcal{L}(q, \dot{q})$)
momentum	opportunity cost per customer (p)
Hamiltonian	opportunity cost rate ($\mathcal{H}(p, q)$)
action	Bellman value function ($\mathcal{V}(q)$)

Table 1: Classical Mechanical Terms and their Call Center Counterparts

4.3 Competing Lagrangians

To simplify our analysis, we assume that c and d are both non-negative, increasing, *concave* functions. This follows from assuming economies of scale for the costs of staff and telephone lines. A case can be made for non-negative, increasing, *convex* functions in the context of a limited supply of agents with specialized skill sets (see Borst, Mandelbaum and Reiman [4]).

With the concave assumption, we show that Optimization Problem 4.2 reduces to the analysis of three “competing” Lagrangians \mathcal{L}_1 , \mathcal{L}_2 and \mathcal{L}_3 , where

$$\mathcal{L}_1(p, q, \dot{q}) \equiv \mathcal{L}(0, 0, p, q, \dot{q}) = p \cdot (\dot{q} - \lambda) + (p - \tau) \gamma q - c(0) - d(0) \quad (4.11)$$

$$\mathcal{L}_2(p, q, \dot{q}) \equiv \mathcal{L}(q, 0, p, q, \dot{q}) = p \cdot (\dot{q} - \lambda) + (p - \sigma) \beta q - c(0) - d(q) \quad (4.12)$$

$$\mathcal{L}_3(p, q, \dot{q}) \equiv \mathcal{L}(0, q, p, q, \dot{q}) = p \cdot (\dot{q} - \lambda) + (p + r) \mu q - c(q) - d(q). \quad (4.13)$$

This result follows from our fundamental lemma.

Lemma 4.3 *If c and d are increasing, concave functions, then we have*

$$\max_{K, L \geq 0} \mathcal{L}(K, L, p, q, \dot{q}) = \mathcal{L}(0, 0, p, q, \dot{q}) \vee \mathcal{L}(q, 0, p, q, \dot{q}) \vee \mathcal{L}(0, q, p, q, \dot{q}) \quad (4.14)$$

which yields

$$\max_{K, L \geq 0} \int_0^T \mathcal{L}(K, L, p, q, \dot{q}) dt = \int_0^T \max_{K, L \geq 0} \mathcal{L}(K, L, p, q, \dot{q}) dt. \quad (4.15)$$

Proof: We assume that $\lim_{L \rightarrow \infty} c(L)$ is finite, and denoted by $c(\infty)$. We make similar assumptions for d and let $d(\infty)$ be that limit. When these limits are infinite, the resulting proof is simpler.

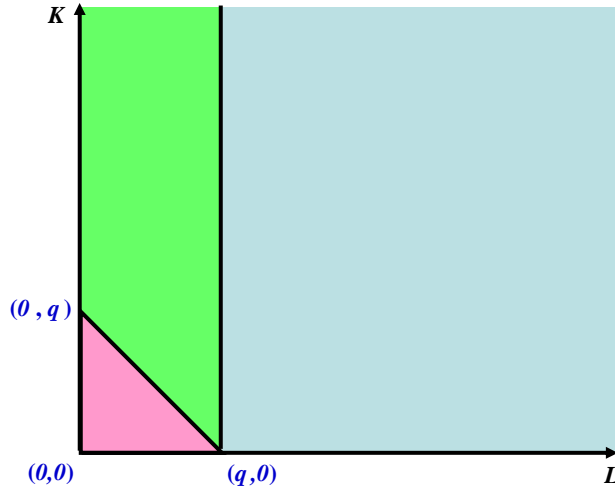


Figure 5: Proof of Fundamental Lemma for Competing Lagrangians.

First observe that given the definition of the Lagrangian by (4.8) we have

$$\lim_{L \rightarrow \infty} \mathcal{L}(K, L, p, q, \dot{q}) = r\mu \cdot q + p \cdot \left\{ \dot{q} - \lambda + \mu \cdot q \right\} - c(\infty) - d(\infty) \leq \mathcal{L}_3(p, q, \dot{q}). \quad (4.16)$$

Similarly, we have

$$\lim_{K \rightarrow \infty} \mathcal{L}(K, L, p, q, \dot{q}) = r\mu \cdot (q \wedge L) - \sigma\beta \cdot (q - L)^+ \quad (4.17)$$

$$+ p \cdot \left\{ \dot{q} - \lambda + \mu \cdot (q \wedge L) + \beta \cdot (q - L)^+ \right\} - c(L) - d(\infty) \leq \left(-\sigma\beta \cdot q + p \left\{ \dot{q} - \lambda + \beta \cdot q \right\} - c(0) - d(\infty) \right) \quad (4.18)$$

$$\vee \left(r\mu \cdot q + p \left\{ \dot{q} - \lambda + \mu \cdot q \right\} - c(q) - d(\infty) \right) \vee \left(r\mu \cdot q + p \left\{ \dot{q} - \lambda + \mu \cdot q \right\} - c(\infty) - d(\infty) \right) \leq \mathcal{L}_2(p, q, \dot{q}) \vee \mathcal{L}_3(p, q, \dot{q}). \quad (4.19)$$

The critical step in this proof is from (4.17) to (4.18). This follows from the fact that every summand on the righthand side of (4.17), except for c , is a piecewise linear function of L . More specifically, they are linear over the intervals $(0, q]$ and $[q, \infty)$. Since $-c$ is everywhere convex, then the limit of the \mathcal{L} as $K \rightarrow \infty$ is a convex function of L over each of these intervals. Now we apply the maximum modulus property of convex functions. The inequality (4.18) then follows from setting L equal to the values 0, q and ∞ in (4.17).

We now observe that \mathcal{L} , when we ignore the terms involving c and d , is a piecewise linear function of K and L . Moreover, these terms are linear over the regions given by Figure (5). It follows that the Lagrangian is piecewise convex over these regions. Now we invoke the maximum modulus property of convex functions and the limiting behavior at infinity to show that the maximum must occur at one of the three vertices identified by the figure, namely $(0, 0)$, $(q, 0)$ or $(0, q)$. ■

At a given time t , we define the largest of these three Lagrangians to be the one that is *dominant*. We now have the following optimal scheduling policy for the fluid penalty model.

Theorem 4.4 Given p and q , construct K^* and L^* as follows:

$$K^*(t) = \begin{cases} q(t) & \text{if } \mathcal{L}_2 \text{ is dominant,} \\ 0 & \text{otherwise.} \end{cases} \quad (4.20)$$

and

$$L^*(t) = \begin{cases} q(t) & \text{if } \mathcal{L}_3 \text{ is dominant,} \\ 0 & \text{otherwise.} \end{cases} \quad (4.21)$$

If p and q maximize $\int_0^T \mathcal{L} dt$, then we have:

1. When $\mathcal{L}_1(p, q, \dot{q})$ is dominant, then p and q solve the Euler-Lagrange equations

$$\dot{p} = (p - \tau) \gamma \quad \text{and} \quad \dot{q} = \lambda - \gamma q. \quad (4.22)$$

2. When $\mathcal{L}_2(p, q, \dot{q})$ is dominant, then p and q solve the Euler-Lagrange equations

$$\dot{p} = (p - \sigma) \beta - d'(q) \quad \text{and} \quad \dot{q} = \lambda - \beta q. \quad (4.23)$$

3. Finally, when $\mathcal{L}_3(p, q, \dot{q})$ is dominant, then p and q solve the Euler-Lagrange equations

$$\dot{p} = (p + r) \mu - (c + d)'(q) \quad \text{and} \quad \dot{q} = \lambda - \mu q. \quad (4.24)$$

One immediate consequence of this theorem is that $K^*(t)$ and $L^*(t)$ are *complementary* variables, i.e. $K^*(t) \cdot L^*(t) = 0$ for all $0 \leq t \leq T$.

If we compare (4.11)-(4.13) with (2.8)-(2.10), we see that

$$\mathcal{L}_i(p, q, \dot{q}) = p \cdot (\dot{q} - \lambda) + \ell_i(p, q) \quad (4.25)$$

for $i = 1, 2, 3$. This means that the determination of which \mathcal{L}_i is dominant is equivalent to finding which ℓ_i is dominant. Moreover, for the special case of c and d being linear, we can subtract $c(0) + d(0)$ from all the \mathcal{L}_i 's and divide them by the always positive factor q . This reduces the comparisons of the competing Lagrangians to the comparisons of *Lagrangian lines* that are purely linear functions of p , as presented in Hampshire and Massey [18].

Summarizing the Lagrangian analysis in Table 2, we see that our fluid model has three operational modes of behavior. When $K^*(t) = L^*(t) = 0$, the model is in “busy signal” mode and the dynamics of q are the same as an infinite server queue with service rate γ . When $K^*(t) = q(t)$ and $L^*(t) = 0$, the model is in “music” mode and the dynamics of q are the same as an infinite server queue with service rate β . Finally, when $K^*(t) = 0$ and $L^*(t) = q(t)$, the model is in “agent” mode and the dynamics of q are the same as an infinite server queue with service rate μ . It is precisely these operational modes that we use when formulating our fluid modified offered load technique in the next section.

Operational Mode	Dominant Lagrangian	Optimal K^*	Optimal L^*	Dynamics of p	Dynamics of q
“busy signal”	$\mathcal{L}_1(p, q, \dot{q})$	0	0	$\dot{p} = (p - \tau) \gamma$	$\dot{q} = \lambda - \gamma q$
“music”	$\mathcal{L}_2(p, q, \dot{q})$	q	0	$\dot{p} = (p - \sigma) \beta - d'(q)$	$\dot{q} = \lambda - \beta q$
“agent”	$\mathcal{L}_3(p, q, \dot{q})$	0	q	$\dot{p} = (p + r) \mu - (c + d)'(q)$	$\dot{q} = \lambda - \mu q$

Table 2: Competing Lagrangian, Fluid Scheduling Analysis.

4.4 Fluid Modified Offered Load

The operational modes of our fluid model for optimal scheduling suggest that we can find a near optimal schedule by having our original call center model with blocking follow the same dynamic where $K(t) \cdot L(t) = 0$ holds for all t . Our call center model is then in busy mode when $K(t) = L(t) = 0$ holds for all t , music mode when only $L(t) = 0$ holds for all t , and agent mode when only $K(t) = 0$ holds for all t . Observe that the stochastic behavior of $Q_{L/0}$ or $Q_{0/K}$ is then the same as for an Erlang loss model with non-homogeneous Poisson arrival rate λ . We have service rate μ for $Q_{L/0}$ but our “service rate” for $Q_{0/K}$ is the regular abandonment rate β .

We estimate the distribution of the time varying Erlang loss system by the modified offered load (MOL) approximation of Jagerman [21]. We simply apply the mean of the associated infinite server queue, which has a Poisson distribution, to the Erlang blocking formula, which is the conditional probability of a Poisson random variable or

$$P \{Q_{L/0}(t) = \ell\} \approx P \{Q_\infty(t) = \ell | Q_\infty(t) \leq L\}, \quad (4.26)$$

where $\ell = 0, 1, \dots, L$. Massey and Whitt [28] provide bounds on the accuracy of this approximation.

In telecommunications, the mean of Q_∞ is referred to as the *offered load* and the mean of $Q_{L/0}$ is called the *carried load*. Typically a fluid approximation of a queueing system can only give us a good estimate of the mean queueing behavior. The MOL approximation gives us a technique to circumvent this problem and allow us to estimate the probability of blocking in terms of the offered load. Moreover, the offered load is the mean of a Poisson random variable, so the square root of the offered load is also the standard deviation for this infinite server queue.

Our fluid model analysis suggests that we can approximate the distribution of the call center blocking model by using the MOL approximation with $q(t)$ as the offered load and initially $\lceil q(t) \rceil$ as the number of agents $L(t)$ in agent mode or buffer spaces $K(t)$ in music mode. This is not adequate for estimating the average number of customers who are blocked from

the system since this involves multiplying the arrival rate by the probability of blocking. The fluid approximates the mean of the underlying queueing distribution. To use it to estimate the distribution however, we need to add to the mean some multiple of an approximation to the standard deviation for the distribution. Moreover, this is an important issue for our call center model since our revenue comes only when the system is in agent mode. Hence, a good estimate of the blocking probability is critical to issues of profit maximization.

Motivated by square root staffing (Jennings et al. [22]), our fluid modified offered load approximation starts with the solution of the fluid model with penalties. The first application of the modified offered load method to refining our scheduling algorithm is to replace the isoperimetric constraints of (4.3) with

$$\int_{\{\text{music mode}\}} \beta \cdot q(t) \cdot (1 - b(0, q(t))) dt = \mathcal{E}_a \quad (4.27)$$

and

$$\int_{\{\text{busy signal mode}\}} \lambda(t) dt + \int_{\{\text{music mode}\}} \lambda(t) \cdot b(0, q(t)) dt + \int_{\{\text{agent mode}\}} \lambda(t) \cdot b(\chi(t), q(t)) dt = \mathcal{E}_b \quad (4.28)$$

where for all a and $q(t) \geq 0$, $b(a, q(t))$ equals the *Erlang blocking formula* with offered load $q(t)$ and $\lceil q(t) + a\sqrt{q(t)} \rceil$ telephone lines or

$$b(a, q(t)) \equiv \frac{q(t)^{\lceil q(t) + a\sqrt{q(t)} \rceil}}{\lceil q(t) + a\sqrt{q(t)} \rceil!} \bigg/ \sum_{i=0}^{\lceil q(t) + a\sqrt{q(t)} \rceil} \frac{q(t)^i}{i!}. \quad (4.29)$$

Based on our MOL approximations, we are estimating the SLA for the blocking call center model. We now use these isoperimetric conditions for the calibration of σ and τ , once we select an appropriate value for a .

The choice of a is a refinement of our staffing policy by adding agents during the agent mode of our staffing policy so that $L(t) = \lceil q(t) + a\sqrt{q(t)} \rceil$. In the language of offered load we are perturbing the offered load mean behavior by some multiple of its standard deviation. We think of square root staffing as another variation to the theme of modified offered load. Using a limit theorem due to Jagerman [21], we have an asymptotic estimate for the Erlang blocking formula with square root staffing

$$\lim_{z \rightarrow \infty} \sqrt{z} \cdot b(a, z) = \frac{\phi(a)}{\Phi(a)}$$

where ϕ and Φ , given by (2.15), are respectively the density and the cumulative distribution function for a mean zero, unit variance Gaussian random variable.

Our selection of a is driven by the goal of profit maximization. When we are in agent mode, we have a pure loss system with no delay, hence no abandonment. As a consequence, all admitted customers immediately receive service and generate r units of revenue. This means that $r\lambda(t) \cdot P\{Q_{L/0}(t) > 0\} - (c + d)(L(t))$ equals the instantaneous profit rate at

any given time during agent mode. This is the arrival rate of customers admitted into the system minus the immediate provisioning cost rate for their service. Using our MOL approximations, we obtain

$$\begin{aligned}
& r\lambda(t) \cdot P \{Q_{L/0}(t) > 0\} - (c+d)(L(t)) \\
&= r\lambda \cdot P \left\{ Q_{\lceil q(t)+a\sqrt{q(t)} \rceil/0}(t) > 0 \right\} - (c+d) \left(\lceil q(t) + a\sqrt{q(t)} \rceil \right) \\
&\approx r\lambda(t) \cdot (1 - b(a, q(t))) - (c+d) \left(q(t) + a\sqrt{q(t)} \right) \\
&\approx r\lambda(t) \cdot \left(1 - \frac{1}{\sqrt{q(t)}} \frac{\phi(a)}{\Phi(a)} \right) - (c+d) \left(q(t) + a\sqrt{q(t)} \right)
\end{aligned}$$

For our fluid modified offered load approximation, we set $a = \chi(t)$, the first positive number such that

$$\begin{aligned}
& r\lambda(t) \cdot \left(1 - \frac{1}{\sqrt{q(t)}} \cdot \frac{\phi(\chi(t))}{\Phi(\chi(t))} \right) - (c+d) \left(q(t) + \chi(t)\sqrt{q(t)} \right) \\
&= \max_{a \geq 0} \left(r\lambda(t) \cdot \left(1 - \frac{1}{\sqrt{q(t)}} \cdot \frac{\phi(a)}{\Phi(a)} \right) - (c+d) \left(q(t) + a\sqrt{q(t)} \right) \right) \quad (4.30)
\end{aligned}$$

We define $\chi(t)$ to be our *square root staffing factor*. By (4.30) we see that it gives us a staffing level that maximizes the approximate profit rate for our call center model during agent mode which reduces to the equivalent statement (2.14). When $\chi > 0$, it also solves the equation

$$\frac{r\lambda(t)}{q(t)} \cdot \frac{\phi(\chi(t))}{\Phi(\chi(t))} \cdot \left(\chi(t) + \frac{\phi(\chi(t))}{\Phi(\chi(t))} \right) = (c+d)' \left(q(t) + \chi(t)\sqrt{q(t)} \right), \quad (4.31)$$

since

$$\begin{aligned}
0 &= \frac{d}{da} \Big|_{a=\chi(t)} \left[r\lambda(t) \cdot \left(1 - \frac{1}{\sqrt{q(t)}} \cdot \frac{\phi(a)}{\Phi(a)} \right) - (c+d) \left(q(t) + a\sqrt{q(t)} \right) \right] \\
&= \frac{-r\lambda(t)}{\sqrt{q(t)}} \cdot \frac{d}{da} \Big|_{a=\chi(t)} \frac{\phi(a)}{\Phi(a)} - (c+d)' \left(q(t) + \chi\sqrt{q(t)} \right) \cdot \sqrt{q(t)} \\
&= \frac{r\lambda(t)}{\sqrt{q(t)}} \cdot \frac{\phi(\chi(t))}{\Phi(\chi(t))} \cdot \left(\chi(t) + \frac{\phi(\chi(t))}{\Phi(\chi(t))} \right) - (c+d)' \left(q(t) + \chi(t)\sqrt{q(t)} \right) \cdot \sqrt{q(t)}.
\end{aligned}$$

We summarize the FMOL algorithm in Table 3. The FMOL approximation of the revenue is

$$r\mu \int_{\{\text{agent mode}\}} q(t) \cdot (1 - b(\chi(t), q(t))) dt. \quad (4.32)$$

Operational Mode	Dominant Lagrangian Term	Optimal K^*	Optimal L^*	Abandonment Rate	Blocking Rate
“busy signal”	$\ell_1(p, q)$	0	0	0	λ
“music”	$\ell_2(p, q)$	$\lceil q \rceil$	0	$\beta \cdot q \cdot (1 - b(0, q))$	$\lambda \cdot b(0, q)$
“agent”	$\ell_3(p, q)$	0	$\lceil q + \chi\sqrt{q} \rceil$	0	$\lambda \cdot b(\chi, q)$

Table 3: Fluid Modified Offered Load Scheduling.

5 Numerical Implementation of the Algorithm

Numerically, the goal of our algorithm is to find positive penalty values σ and τ such that we satisfy the isoperimetric constraints of (2.11) and (2.12) where q paired with p solves the competing Lagrangian equations (2.5)-(2.10) and χ is the smallest positive number that satisfies (2.14) or solves (4.31).

We reduce the complexity of the search for the optimal pair by computing a gradient. In particular the gradient, with respect to this penalty pair, of the action or the integral of the optimal fluid Lagrangian with penalties. Our next candidate for σ and τ , then comes from updating the old pair by adding some scalar multiple of this gradient. Using our competing Lagrangian result and the Envelope Lemma [8], we give an analytic expression for the derivative of the action with respect to τ ,

$$\begin{aligned}
& \frac{d}{d\tau} \int_0^T \mathcal{L} \left(t, K(t), L(t), p(t), q(t), \dot{q}(t) \right) dt \\
&= \int_0^T \frac{\partial}{\partial \tau} \left(\mathcal{L}_1 \left(t, p(t), q(t), \dot{q}(t) \right) \vee \mathcal{L}_2 \left(t, p(t), q(t), \dot{q}(t) \right) \vee \mathcal{L}_3 \left(t, p(t), q(t), \dot{q}(t) \right) \right) dt \\
&= \int_{\{\text{busy signal mode}\}} \frac{\partial}{\partial \tau} \mathcal{L}_1 \left(t, p(t), q(t), \dot{q}(t) \right) dt \\
&= -\gamma \int_{\{\text{busy signal mode}\}} q(t) dt. \tag{5.1}
\end{aligned}$$

By a similar argument, we obtain a formula for the derivative in the σ direction

$$\frac{d}{d\sigma} \int_0^T \mathcal{L} \left(t, K(t), L(t), p(t), q(t), \dot{q}(t) \right) dt = -\beta \int_{\{\text{music mode}\}} q(t) dt. \tag{5.2}$$

The competing Lagrangian equations are solved by a backward-forward shooting method. From a computational perspective, it is best to view q as a process that evolves forwards in

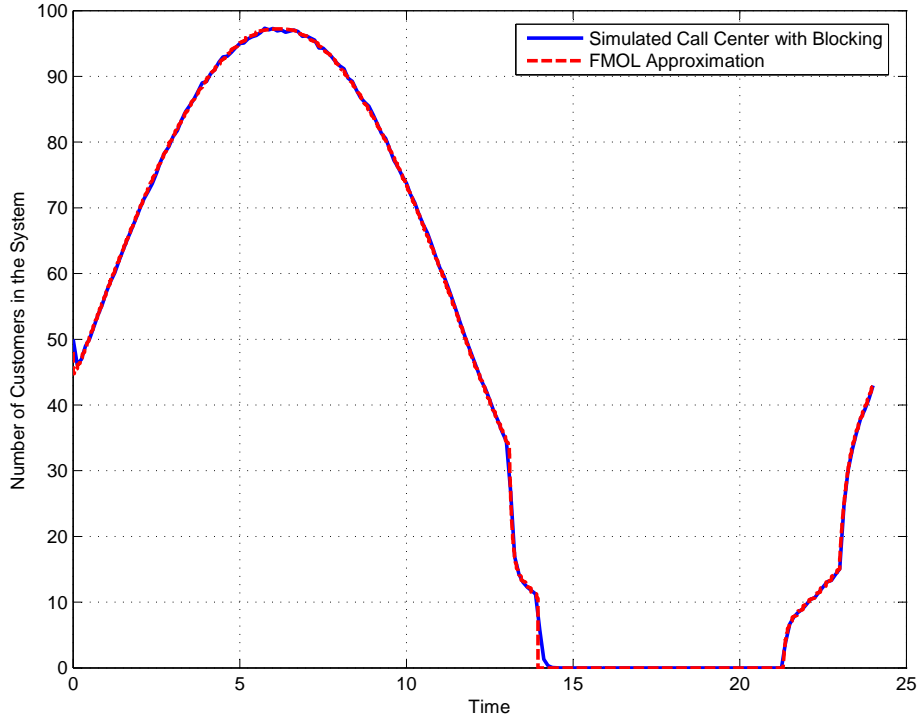


Figure 6: Comparison of the Simulated Call Center with Blocking to its FMOL Approximation.

time with a given initial value $q(0)$ but p as a process that evolves *backwards* in time with a terminal value $p(T) = 0$. For more background on shooting methods see Bryson and Ho [7].

5.1 Numerical Example

Now, we present a numerical example to demonstrate the algorithm. The original call center model is simulated using the near optimal schedule generated by our FMOL algorithm. The simulated performance quantities and profit are compared to their fluid approximations.

We consider an example where the arrival rate is $\lambda(t) = 300.0 + 300.0 \cdot \sin(2.0 \cdot \pi \cdot (t/24.0))$, service rate $\mu = 6.0$, and abandonment rate β equals 12.0. The target fractions of customer abandonment and blocking are $\epsilon_a = 0.10$ and $\epsilon_b = 0.05$ and respectively. The revenue per serviced customer r is normalized to equal 1.0. The cost rate for agents is $c(x) = 300.0 \cdot \log(1.0 + (x \cdot (\exp(1.0) - 1.0) / 50.0))$ and the cost rate for line usage is $d(x) = \log(1.0 + x)$. The initial number of customers in the system is $Q(0) = 50$. We achieve the abandonment and blocking SLA targets by setting the penalties to $\sigma = 0.10$ and $\tau = 0.12$. The resulting FMOL schedule visits each mode of operation at least once, a consequence of the dynamic demand setting.

We simulate 10,000 realizations of the original call center model under the FMOL schedule. Figure 6 compares the simulated mean queue length to the modified offered load approximation of the queue length under the near-optimal schedule. Table 4 compares the abandonment and blocking fractions as well as the profit of the simulated *original* finite buffer

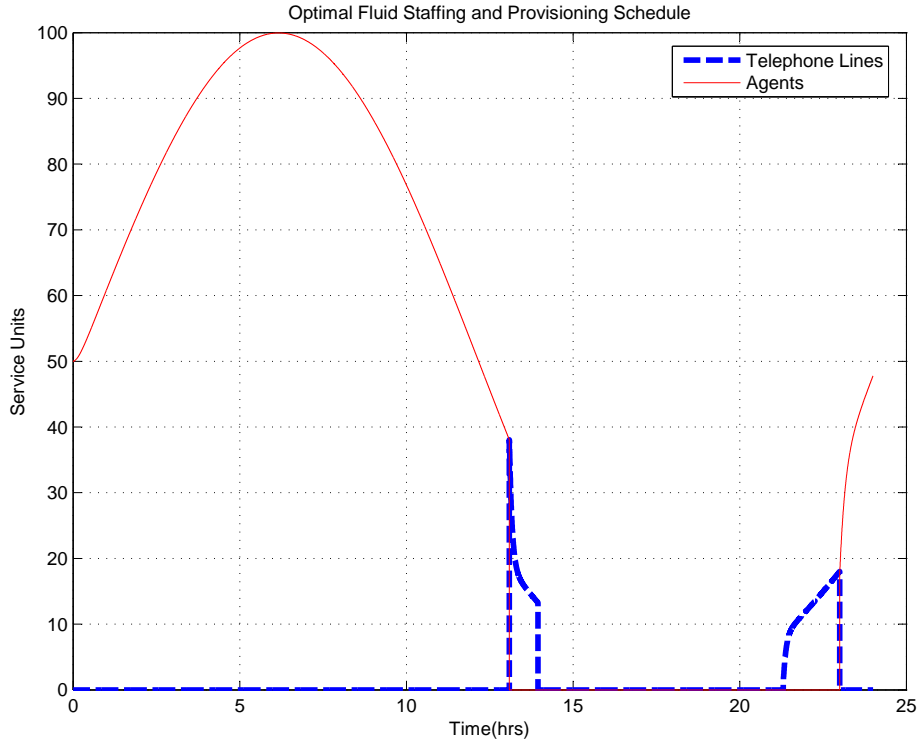


Figure 7: Fluid Optimal Schedule

call center model with the fluid modified offered load approximation of the same quantities under the FMOL schedule complete with the relative errors for all three quantities.

5.2 Fluid Dynamics

The left hand plot of Figure 7 displays the fluid optimal schedule of K and L versus time t . The right hand plot displays the phase space trajectory of p versus q . The schedule visits all three optimal staffing modes. In terms of the call center, the phase space diagram is a plot of the opportunity cost per customer p versus a fluid approximation of the number of customers in the system q . One indication of the Lagrangian being time dependent is that the phase space plot intersects itself. This would not be the case if our arrival rate λ were a constant.

	Abandonment Fraction	Blocking Fraction	Total Profit
Fluid Modified Offered Load Estimation	0.0498	0.0996	504.5459
Finite Buffer Call Center Simulation	0.0501	0.0993	501.3786
Relative Error	-0.006	0.003	0.006

Table 4: SLA and Profit Comparisons of the FMOL and Call Center with Blocking Simulation

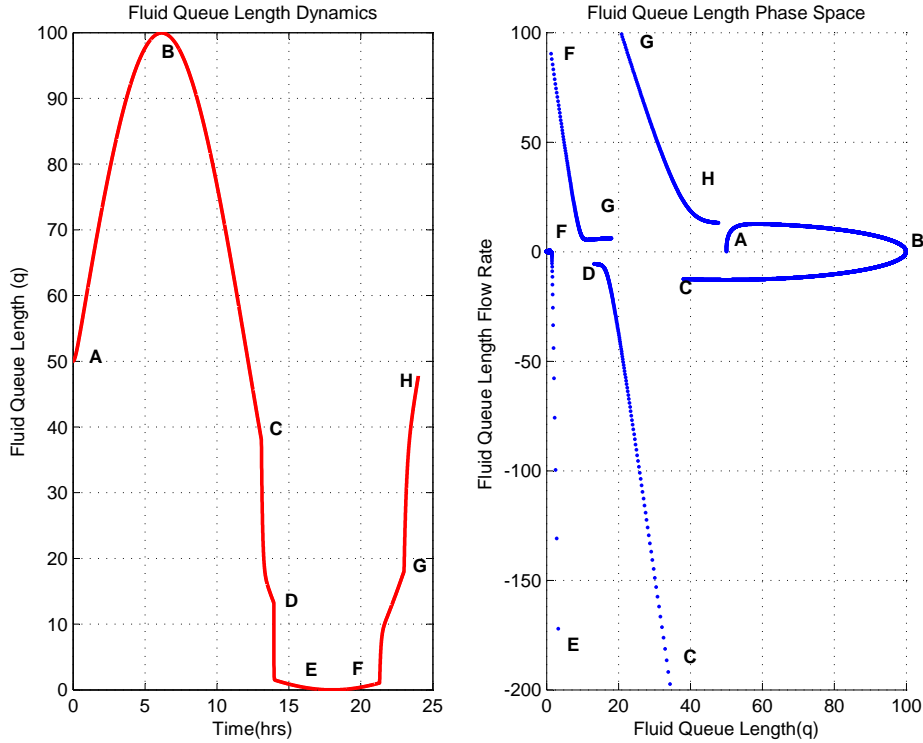


Figure 8: Dynamics of the Fluid Queue Length (q)

Figure 8 displays the dynamics and phase space plot of only the fluid queue length. The critical values of the fluid queue length are labeled on both the dynamics (q versus t) and phase space (q versus \dot{q}) plots. The labeled points on the phase space plot correspond in time to the labeled points on the plot of the dynamics. Curve segment AC corresponds to the agent mode, where the maximum number of agents occurs at B. The jump in the phase plot at C corresponds to switching from agent mode to music mode. The system is in music mode along curve segment CE. The jump in the phase plot at D is due to switching from music mode to busy signal mode. During the interval EF the busy signal model is optimal. Music mode then becomes optimal on the curve segment FG. The jump in the phase plot at F is due to switching from music mode to agent mode. Finally, we return to agent mode on the curve segment GH. Figure 9 displays the dynamics of the opportunity cost per customer. The analogous critical values of the opportunity cost per customer p are labeled. Notice that p peaks at time t during the busy signal mode only when $\dot{p}(t) = (p(t) - \tau)\gamma = 0$ which means that $p(t) = \tau = 0.12$. Moreover the derivative of a constant is zero so once p equals τ , it should stay there for the duration of the busy signal mode. All these dynamics are consistent with the plot of p over time.

5.3 Discrete Schedules, Continuity and Local Optimality

Our FMOL algorithm gives a call center schedule where we assume that the numbers of agents and lines are continually updated at any time. We now construct a series of schedules from the FMOL schedule that are only updated at constant, discrete time intervals. For any

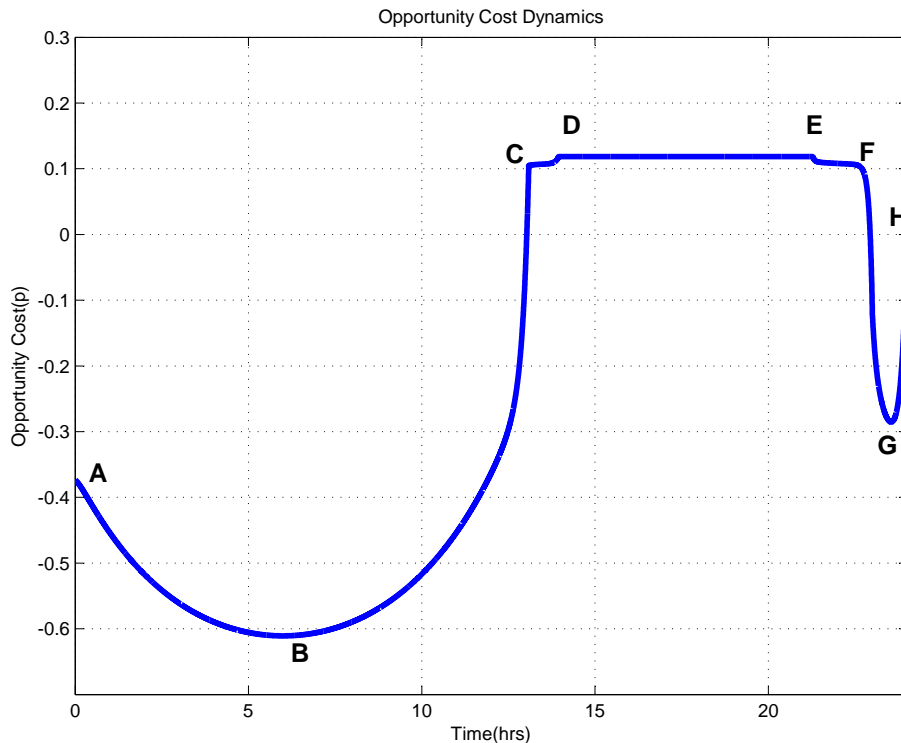


Figure 9: Dynamics of the Opportunity Cost (p)

positive integer n , we create a *dyadic partition* of $(0, T]$ where divide it up into 2^n disjoint intervals.

Next, we set the constant number of additional lines K over any given subinterval equal to the *minimum* of K^* over that same subinterval. Having fewer servers than needed reduces the average number of abandoning customers. Using a similar logic, we should also set the number of agents over any given subinterval equal to the *maximum* of L^* over that same subinterval. Here it is the case that the addition of agents also reduces the average number of blocked customers. This type of schedule, rounding downwards for L and rounding upwards for K , is similar to design methods discussed in Wallace [29] and [27].

The end result is a series of schedules that are *feasible*, i.e. their performance is bounded above by the SLA targets of $\epsilon_a = 0.05$ and $\epsilon_b = 0.10$. As we see in Figure 10, where n grows and the length of a subinterval equals $1/2^n$, the abandonment and blocking percentages (left and middle plots respectively) converge monotonically upwards to their SLA targets. Here, we are simulating an $M_t/M/L_t/K_t$ queue with abandonment according to these schedules. This suggests that the performance of the FMOL schedule can be approximated with arbitrary precision by a schedule with regular, discrete staffing (and provisioning) intervals. Thus our performance metrics behave as continuous functions of the scheduling. Moreover, observe that the mean profit (right plot) of these feasible schedules is also monotonically increasing in n . All these profits fall below the one given by the FMOL schedule. This suggests that FMOL is locally optimal for this sequence of schedules and possibly near-optimal in general.

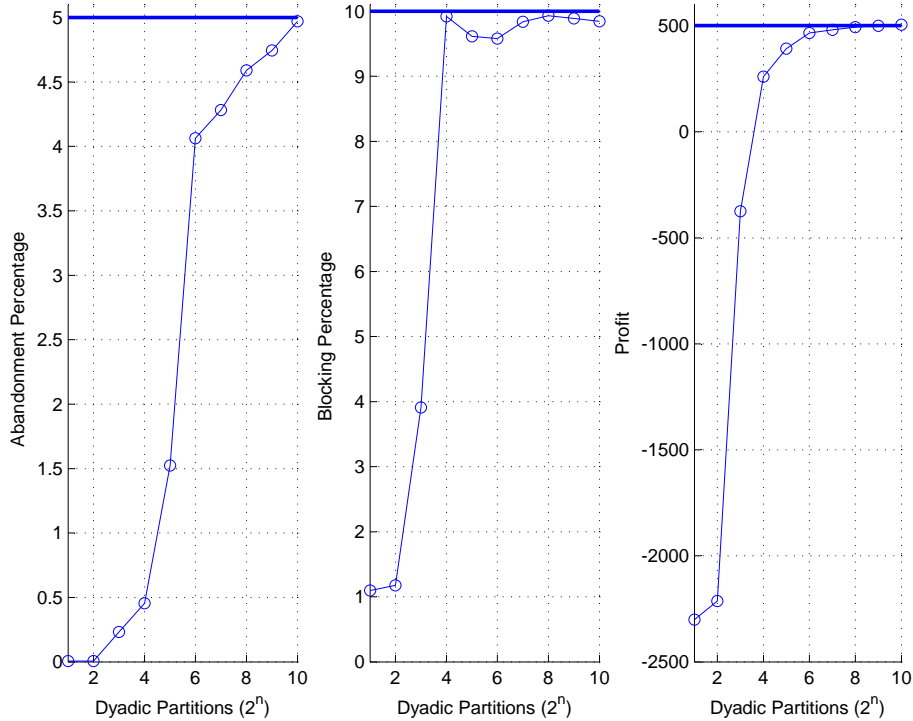


Figure 10: Plots of the Abandonment and Blocking Percentages as well as the Profit for Simulations of the Call Center Model with Blocking for a Dyadic Partition of the FMOL Schedule

6 Summary and Conclusions

We model a call center as a multi-server queue with finite additional waiting spaces and abandonment. We assume that there is a reward for every successful service completion, SLA target levels for the fractions of calls that are abandoned and blocked, and a cost rate for the number of agents and telephone lines used. The finite capacity model is then approximated by a Markovian service network, an infinite buffer, multi-server queue with regular abandonment with a notion of fast abandonment.

Motivated by growing a business to match a corresponding growth in customer demand, our MSN model converges to a deterministic “fluid” model that is a dynamical system. Moreover, this Lagrangian analysis shows that the optimal staffing and provisioning of the dynamic fluid model for the call center with SLA target levels is equivalent to a penalty formulation. By using the theory of dynamic optimization, we formulate a fluid optimal staffing schedule for profit optimality. The optimal staffing and provisioning schedules are generated by one of three optimal modes where we select the one with the dominant profit

rate. The three modes are:

1. No agents and no telephone lines (busy signal).
2. Telephone lines but no agents (music).
3. Agents but no additional telephone lines (talking to an agent).

Given that each of the three optimal staffing modes correspond to loss systems, the fluid model is enhanced by the modified offered load approximation. Insights from the modified offered load approximation lead to refining the fluid schedule by adding on a quantity that is proportional to the square root of the number of agents during the agent mode of operation given by the fluid model. The square root factor is then selected to balance the marginal revenue and marginal costs of adding additional agents and lines. Moreover, the modified offered load approximation also leads to accurate estimates of blocking rate, the abandonment rate, and the number of customers in the system. Thus these modified offered load approximation methods allow for the estimation of quantities typically beyond the scope of a fluid model.

We present a numerical example that visits all three operational modes. This occurs by exploiting the economies of scale. The cost per customer for providing the necessary resources for service decreases as the increase the total number of customers in service. We simulate the original call center model with blocking under the FMOL schedule. We use the FMOL approximations to obtain estimates of blocked customers, the number of abandoning customers, revenue, and the number of customers in the system. Finally, we simulate perturbed versions of the near-optimal schedules. The resulting perturbed schedules produce either lower profits or violate the service level agreements. These results suggest that our schedule is locally optimal.

Many connections are made between optimal call center staffing and classical mechanics. Table 1 shows the relationship between the call center quantities of interest and their classical mechanical counterparts. Viewing the call center optimization problem as a mechanics problem provides insight and additional intuition into the analysis and computation of the optimal staffing and provisioning schedules for the operations of this communications service.

Acknowledgements

The authors wish to thank the referees for their useful comments. Also thanks to Rudy L. Horne for bringing to our attention the connections between dynamic optimization and classical mechanics.

References

- [1] Armony, M., Shimkin, N. and Whitt, W. “The Impact of Delay Announcements in Many-Server Queues with Abandonment.” Forthcoming in *Operations Research*.
- [2] Baron, O. and Milner, J. “Staffing to Maximize Call Centers with Alternate Service Level Agreements.” Under review.

- [3] Bassamboo, A., Zeevi, A. and Harrison, J. M. “Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based method.” *Operations Research*, Vol. 54, pp. 419-435, 2006.
- [4] Borst, S., Mandelbaum, A. and Reiman, M. I. “Dimensioning Large Call Centers,” *Operations Research*, Vol. 51, No. 1, pp. 17-34, 2004.
- [5] Bhandari, A., Harchol-Balter, M. and Scheller-Wolf, A. “ An Exact and Efficient Algorithm for the Constrained Dynamic Operator Staffing Problem for Call Centers”, *Management Science*, to Appear, 2007.
- [6] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. “Statistical analysis of a telephone call center: a queueing-science perspective.” *J. Amer. Statist. Assoc.* 100, pp. 36–50, 2005.
- [7] Bryson, A. and Ho, Y. *Applied Optimal Control*, Hemisphere Publishing Corp.,1975.
- [8] Dixit, A. K. *Optimization in Economic Theory*, Oxford University Press, 1990.
- [9] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. “Staffing of Time-Varying Queues to Achieve Time-Stable Performance”. To appear in *Management Science*, 2008.
- [10] Gans, N., Koole, G. and Mandelbaum, A. “Telephone Call Centers: Tutorial, Review and Research Prospects.” *Manufacturing Service Operations Management* 5(2), pp. 79-141, 2003.
- [11] Garnett, O., Mandelbaum, A. and Reiman, M. “Designing a Call Center with Impatient Customers,” *Manufacturing and Service Operations Management*, **4(3)**, 208–227, 2002.
- [12] Green, L., Kolesar, P., and Whitt, W. “Coping with Time-Varying Demand when Setting Staffing Requirement”, *Production and Operations Management*, Vol. 16, Issue 1, pp. 13-39, 2007.
- [13] Gregory, J. and Lin, C. *Constrained Optimization in the Calculus of Variations and Optimal Control Theory*, Von Nostrand Reinhold, 1992.
- [14] Gurvich, I. and Whitt, W. “Service-Level Differentiation in Many-Server Service Systems: A Solution Based on Fixed-Queue-Ratio Routing.” Under review.
- [15] Gurvich, I., Armony, M. and Maglaras, C. “Cross-Selling in a Call Center with a Heterogenous Customer Population.” Under review.
- [16] Halfin, S. and Whitt, W. “Heavy-Traffic Limits for Queues with Many Exponential Servers,” *Operations Research*, 29, pp. 567–587, 1981.
- [17] Hampshire, R. C. “Dynamic Queueing Models for the Operations Management of Communication Services.”. *Ph.D. Dissertation*, Princeton University, March 2007.

- [18] Hampshire, R. C. and Massey, W. A. “Variational Optimization for Call Center Staffing” (Extended Abstract). *Richard Tapia Celebration Of Diversity In Computing, Proceedings of the 2005 Conference on Diversity in Computing*. Albuquerque, New Mexico, USA, 2005.
- [19] Hampshire, R. C., Massey, W. A., Mitra, D. and Wang, Q. “Provisioning for Bandwidth Trading,” *Telecommunications Network Design and Management (Boca Raton, FL, 2002)*, 207–225, Oper. Res./Comput. Sci. Interfaces Ser., 23, Kluwer Acad. Publ., Boston, MA, 2003.
- [20] Harris, C. M., Hoffman, K. L. and Saunders, P. B. “Modeling the IRS Telephone Taxpayer Information System,” *Operations Research*. Volume 35, Number 4, pp. 504–523, July-August 1987.
- [21] Jagerman, D. L. “Nonstationary Blocking in Telephone Traffic,” *Bell System Technical Journal*; pp. 625–661, 1975.
- [22] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. “Server Staffing to Meet Time-Varying Demand,” *Management Science*, 42:10 (October 1996), pp. 1383–1394.
- [23] Koole, G. and Mandelbaum, A. “Queueing Models of Call Centers: An Introduction,” unpublished, Oct 2001, Downloadable from <http://www.cs.vu.nl/obp/callcenters> or <http://iew3.technion.ac.il/serveng>.
- [24] Lanczos, C. *The Variational Principles of Mechanics*, Fourth Edition, Dover Publications, 1970.
- [25] Mandelbaum, A. “Call Centers Research Bibliography with Abstracts,” unpublished, Version 2, Sept 2001, Downloadable from <http://iew3.technion.ac.il/serveng>. 17.
- [26] Mandelbaum, A., Massey, W. A. and Reiman, M. I. “Strong Approximations for Markovian Service Networks,” *Queueing Systems and Their Applications*, 30 (1998) pp. 149–201.
- [27] Massey, W. A. and Wallace, R. B. “An Optimal Design of the $M/M/C/K$ Queue for Call Centers,” under review.
- [28] Massey, W. A. and Whitt, W. “An Analysis of the Modified Offered Load Approximation for the Nonstationary Erlang Loss Model,” *The Annals of Applied Probability*, Vol.4, No.4, (1994) pp. 149–201 .
- [29] Wallace, R. B. “Performance Modeling of Call Centers with Skill-Based Routing,” Ph.D. Dissertation, George Washington University, 2003.
- [30] Whitaker, B. A. “Analysis and Optimal Design of a Multiserver, Multiqueue System with Finite Waiting Space in Each Queue,” *Bell Syst. Tech. J.* Volume 54, pp. 595–623, 1975.

- [31] Zeltyn, S. and Mandelbaum, A. “Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers”. Available at <http://iew3.technion.ac.il/serveng/References/references.html>., 2007.
- [32] Zeltyn, S. and Mandelbaum, A. “Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue.” *Queueing Systems*, 51 (3/4), pp. 361-402, 2005.