

The Workload Process:  
Modelling, Inference and Applications.

M.Sc. Research Proposal

Michael Reich

Advisor: Prof. Avishai Mandelbaum  
The Faculty Of Industrial Engineering and Management  
Technion - Israel Institute of Technology

December 28, 2007

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Basic Stationary Models with iid Customers and iid Servers</b>	<b>4</b>
2.1	Erlang-C: The Square Root Safety Staffing Rule . . . . .	5
2.2	Erlang-A: Summary of Some Results . . . . .	7
2.3	Erlang-A: Four Operational Regimes . . . . .	8
<b>3</b>	<b>Our Workload Process and the Offered-Load</b>	<b>9</b>
3.1	The Offered Load of $M_t/GI/N_t + GI$ . . . . .	10
3.2	The Gap\ Lag - Examples . . . . .	13
3.3	Staffing the $M_t/GI/N_t + GI$ Queue . . . . .	14
<b>4</b>	<b>Our Proposed Research</b>	<b>15</b>
4.1	Imputing Service Times . . . . .	16
4.2	Staffing Aiming to Immediately Answer All Calls [17] . . . . .	18
<b>5</b>	<b>Data Sources</b>	<b>18</b>

# 1 Introduction

During the last century and the beginning of the current one, the service sector has grown significantly and now accounts for approximately 70% of the national income in the United States. The service sector covers a wide spectrum of activities, e.g. education, professional services, financial services and government services. In this thesis we focus on telephone call centers and on the health care system.

Call centers are very commonly used by companies and organizations for managing their customer relationships. This covers both the public and private sector. For some of the companies, such as banks and cellular operators, their call centers are the main channel for maintaining contact with their customers. In general, call centers are becoming a vital part of the service-driven society nowadays. As a result call centers have also become an object for academic research.

For the analysis of call centers operations, queueing theory and statistics are being used. The problems that call centers operators are dealing with, are often related to statistical characteristics of the calls arrival and handling processes. Very often, about 70% of the operations costs are devoted to human resources. As a result, forecasting the calls arrival rate, understanding the handling process, setting the right service performance measures and staffing levels, are central problems at all call centers. These are issues that any call center manager deals with on a daily basis.

Another important area of the service sector is the health care system. Hospital managers are increasingly aware of the need to use their resources as efficiently as possible in order to continue to assure their institutions' survival and prosperity. Moreover, there is a consistent pressure, coming from the patients, to increase the quality of care by technique improvements and medical innovations and to shorten the sojourn time in the hospital. Green [7] describes the general background and issues involved in hospital capacity planning and provides examples of how Operations Research models can be used to provide important insights into operational strategies and practice.

Satisfactory customer service can be defined in many ways, based on various performance measures. Focusing on operational measures, a customer enjoys satisfactory service if her delay in queue is at most  $\tau$  seconds [15]. For an emergency medical service system, the main performance measure is the fraction of calls that are reached within some time standard. The response time is typically considered to begin at the instant the call was made and end when an ambulance reaches the call address. In North America, a typical target is to reach 90% of the most urgent calls within 9 minutes [2]. The measures of quality are not absolute. They differ among service system in accordance with the system's goals, its environment and the services it offers.

The staffing problem is key for providing satisfactory service. Managers of call centers must decide on how many agents to hire and get to work at any time in order to ensure satisfactory customer service. In emergency medical service and in hospitals the system is even more complicated. Staffing consists of scheduling decisions for ambulances and their crews, for doctors and nurses and it must also account for the number of inpatient beds.

Service environments are typically very complex. Some of the complexity factors are, for example: changes of the environment parameters (arrival/service rates), uncertainty of these parameters, due to either random variation over time or lack of information, variety of services with different requirements (i.e. time and skills). Naturally, staffing decisions must account for all this complexity, yet the challenge is to develop staffing rules that are simple and insightful enough for implementation. For example, the "square-root staffing" rule is such a rule, as will be surveyed below.

Mathematical models have been developed to model the complexity of the call centers environment. Their strength is their simplicity and the theoretical insights they provide. On the other hand, their modeling scope is limited, and an analytical knowledge is needed in order to apply them efficiently. These weaknesses could be the reason why these models are not being used as often as expected. The most commonly-used models are the  $M/M/N$  queue (Erlang-C) and  $M/M/N+M$  queue (Erlang-A), which will be described in section 2.

In our research we shall investigate the application of dynamic staffing in service systems, based on operational data. By dynamic staffing we mean, that at any point of time  $t$ , we exploit all relevant information that we have till time  $t$ , in order to obtain good estimates for the workload in the system in the near future beyond  $t$ . For this purpose, a prediction is required for the future workload in the system, which consist of several components, e.g.: the remaining holding time of calls that are in progress at time  $t$ , the number of new calls that will arrive and the service times of those who will come. This provides the basis for our staffing decision, which will be dependent also on the desired service performance. Whitt suggested in [17] methods to facilitate dynamic staffing in telephone call centers to aim at immediately answer all calls. We plan to modify [17] so as to accommodate the effects of customers' impatience.

Our main data source is a unique repository of call centers data, including a detailed call-by-call operational history of several sources. The data source is described in more details in Section 5. See also the extended documentation in [16].

The research proposal is structured as follows. In section 2 we describe some basic models of service systems: ( $M/M/N$  and  $M/M/N+M$ ), and survey the work of [5]. In section 3 we discuss issues related to time-dependent queueing models. A description of the research that we propose to undertake is presented in section 4 and we conclude in section 5 with a description of the data source on which we shall be relying in our research.

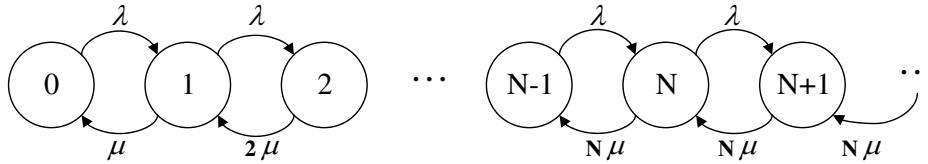
## 2 Basic Stationary Models with iid Customers and iid Servers

In this section, the two most common models that are used for call centers modeling and staffing are presented. The first model is Erlang-C: first developed around 1910 by Erlang [4], it has served until recently as the "working-horse" of call center staffing. Its main deficiency is that it ignores customers' impatience, which is remedied by the second model, namely Erlang-A. Impatience leads to the phenomenon of customers' abandonment, and, already around 1940, Palm [12] developed Erlang-A in order to capture it. We shall be using Erlang-A to motivate four operating regimes for medium-to-large call centers: one which emphasizes service quality, another that focuses on operational efficiency, a third that carefully balances these two goals of

quality and efficiency, and a fourth that is a refinement of the second by the third.

## 2.1 Erlang-C: The Square Root Safety Staffing Rule

The classical  $M/M/N$  (Erlang-C) queueing model is characterized by Poisson arrivals at rate  $\lambda$ , iid exponential service times with an expected duration  $1/\mu$ , and  $N$  servers working independently in parallel. One can view this model as a simple Birth-Death process where  $\lambda$  is the (constant) birth rate and  $\mu$  is the (constant) death rate. A Markovian-state description of the process is the total number of customers in the system, either served or queued. The corresponding transition rate diagram is then the following:



**Figure 1:** Erlang- C as a Birth-Death Process

Erlang-C is ergodic if and only if its traffic intensity  $\rho = \frac{\lambda}{\mu} < 1$ ;  $\rho$  is then the servers' utilization, namely the long-run fraction of time that a server is busy.

The stationary/limit distribution is defined as

$$\pi_j = \lim_{t \rightarrow \infty} P\{L(t) = j\}, \quad j \geq 0,$$

where  $L(t)$  is the system state at time  $t$ , namely, the total number of customers in the system. Solution of the following steady-state equations yields the probabilities  $\pi_j$  of being at any state  $j$  during steady-state:

$$\begin{cases} \lambda\pi_j = (j+1)\mu_{j+1} & 0 \leq j \leq N-1 \\ \lambda\pi_j = (N\mu)\pi_{j+1} & j \geq N \end{cases} \quad (1)$$

The probability that in steady-state all the servers are busy is given by  $\sum_{j \geq N} \pi_j$ , the stationary probability of being in one of the states  $\{N, N+1, \dots\}$ . This probability is sometimes referred to as the *Erlang-C formula*. It is denoted  $E_{2,N}$  and is given by

$$E_{2,N} = \sum_{j \geq N} \pi_j = \frac{(\lambda/\mu)^N}{N!(1-\rho)} \left[ \sum_{i=1}^{N-1} \frac{(\lambda/\mu)^i}{i!} + \frac{(\lambda/\mu)^N}{N!(1-\rho)} \right]^{-1} \quad (2)$$

The Poisson distribution of arrivals has an important and useful consequence, known as *PASTA* (Poisson Arrivals See Time Averages): it implies that the probability  $E_{2,N}$  is in fact also the probability that a customer is delayed in the queue (as opposed to being served immediately upon arrival).

Using formula (2) and its relatives, one can calculate staffing levels  $N$  for any desired service level, given the arrival rate and average service time. This can be easily done even with a spreadsheet. However, such a solution does not provide any insight on the dependence of  $N$  on model parameters, for example, how should  $N$  change if the load was doubled. Such insight comes out of a staffing rule that goes back to as early as Erlang [4], where he derived it via marginal analysis of the benefit of adding a server. (Erlang indicated that the rule had been practiced actually since 1913.) This is the *square-root safety- staffing rule*, which we now describe.

Let  $R = \lambda/\mu$  denote the average *offered load*. Then the square root safety-staffing rule states the following: for moderate to large values of  $R$ , the appropriate staffing level is of the form

$$N = R + \beta\sqrt{R} \tag{3}$$

where  $\beta$  is a positive constant that depends on the desired level of service;  $\beta$  will be referred to as the *Quality-of-Service (QOS)* parameter: the larger the value of  $\beta$ , the higher is the service quality. The second term on the right side of (3) is the excess (safety) capacity, beyond the nominal requirement  $R$ , which is needed in order to achieve an accepted service level under stochastic uncertainty.

The form of (3) carries with it a very important insight. Let  $\Delta = \beta\sqrt{R}$  denote the safety staffing level (above the minimum  $R = \lambda/\mu$ .) Then, if  $\beta$  is fixed, an  $n$ -fold increase in the offered load  $R$  requires that the safety staffing  $\Delta$  increases by only  $\sqrt{n}$ -fold, which constitutes significant economies of scale.

What does (3) guarantee as far as *QOS* is concerned? For Erlang-C, this is the subject of the seminal paper by Halfin and Whitt [8], where they provided the following answer:

**Theorem 1.** : *Consider a sequence of  $M/M/N$  queues, indexed by  $N=1,2,\dots$ . As the number of servers  $N$  grows to infinity, the square-root safety-staffing rule applies asymptotically if and only if the delay probability converges to a constant  $\alpha$  ( $0 < \alpha < 1$ ), in which case the relation between  $\alpha$  and  $\beta$  is given by the Halfin-Whitt function*

$$\alpha = [1 + \beta/h(-\beta)]^{-1} \tag{4}$$

here  $h(x) = \phi(x)/(1 - \Phi(x))$  is the *hazard rate* of the standard *normal distribution*  $N(0,1)$ .

*Note that (3) applies if and only if  $\sqrt{N}(1 - \rho_N)$  converges to  $\beta$  ( $\beta > 0$ ). Indeed, formally the Theorem of Halfin-Whitt reads:*

$$\text{As } N \uparrow \infty, P(\text{Wait} > 0) = E_{2,N} \rightarrow \alpha \quad (0 < \alpha < 1)$$

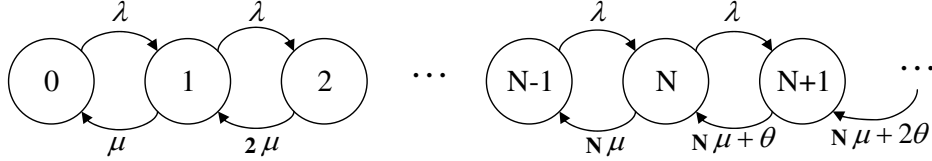
$$\text{iff } \sqrt{N}(1 - \rho_N) \rightarrow \beta \quad (0 < \beta < \infty)$$

$$(\text{equivalently } N \approx R_N + \beta\sqrt{R_N}) .$$

In practice, this rule makes the life of a call-center manager easier: he or she can actually specify the desired delay probability and achieve it by following the square-root safety staffing rule (3), simply choosing the right  $\beta$ .

## 2.2 Erlang-A: Summary of Some Results

Trying to make the  $M/M/N$  model more realistic and useful, the following assumption is added: each customer has limited patience, that is, as the waiting time in the queue grows the customer may abandon. We assume that patience is distributed exponentially with mean  $1/\theta$ . This model is referred to as Erlang-A (A for Abandonment). It is also a Birth-Death process, and its transition rate diagram is depicted below.



**Figure 2:** Erlang-A as a Birth-Death Process

The steady-state distribution is found by solving of the following equations:

$$\begin{cases} \lambda\pi_j = (j+1)\mu\pi_{j+1} & 0 \leq j \leq N-1 \\ \lambda\pi_j = (N\mu + (j+1-N)\theta)\pi_{j+1} & j \geq N \end{cases}$$

The solution of the steady-state equations is given by:

$$\pi_j = \begin{cases} \frac{(\lambda/\mu)^j}{j!} \pi_0, & 0 \leq j \leq N \\ \prod_{k=N+1}^j \left( \frac{\lambda}{N\mu + (k-N)\theta} \right) \frac{(\lambda/\mu)^N}{N!} \pi_0, & j \geq N+1 \end{cases} \quad (5)$$

where

$$\pi_0 = \left[ \sum_{j=0}^N \frac{(\lambda/\mu)^j}{j!} + \sum_{j=N+1}^{\infty} \prod_{k=N+1}^j \left( \frac{\lambda}{N\mu + (k-N)\theta} \right) \frac{(\lambda/\mu)^N}{N!} \right]^{-1}$$

Just as with Erlang-C, appropriate staffing levels in Erlang-A can be found using the solution to (4), but again, the disadvantage is that this does not provide any insights. To this end, the Erlang -A analogue of Theorem 1 was proved in [6], and it is given as follows:

**Theorem 2. :** *Consider a sequence of  $M/M/N+M$  queues, indexed by  $N=1,2,\dots$ . As the number of servers  $N$  grows to infinity, the square-root safety-staffing rule (3) applies asymptotically if and only if the delay probability converges to a constant  $\alpha$  ( $0 < \alpha < 1$ ), in which case the relation between  $\alpha$  and  $\beta$  is given by the Garnett function*

$$\alpha = \left[ 1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta} \right]^{-1}, \quad -\infty < \beta < \infty. \quad (6)$$

Moreover, the above conditions apply if and only if  $\sqrt{n}P(\text{Abandon})$  converges to some positive constant  $\gamma$  that is given by

$$\gamma = \alpha\beta \left[ \frac{h(\delta)}{\delta} - 1 \right], \quad \delta = \beta\sqrt{\mu/\theta}.$$

Formally, Theorem 2 reads:

$$\text{As } N \uparrow \infty, P(\text{Wait} > 0) \rightarrow \alpha \quad (0 < \alpha < 1)$$

$$\text{iff } \sqrt{N}(1 - \rho_N) \rightarrow \beta \quad (-\infty < \beta < \infty)$$

$$(\text{equivalently } N \approx R_N + \beta\sqrt{R_N})$$

$$\text{iff } \sqrt{N}P(\text{Abandon}) \rightarrow \gamma \quad (0 < \gamma < \infty).$$

An important feature of Erlang-A is that, unlike Erlang-C, it is always stable whenever the abandonment rate  $\theta$  is positive.

Theorem 2 demonstrates that the square-root safety-staffing rule prevails for Erlang-A as well. The QOS parameter  $\beta$  now depends on both the abandonment rate  $\theta$  and the delay probability  $\alpha$ . It is significant that here  $\beta$  may take also negative values (since Erlang-A is always stable).

### 2.3 Erlang-A: Four Operational Regimes

Each organization has its own preferences in everyday functioning. Some of them try to get the most from the available resources, while others see customers' satisfaction as the most important target. Depending on organizational preferences, three different operational regimes arise [1]:

- Efficiency driven (ED).
- Quality driven (QD).
- Quality-Efficiency driven (QED).

A fourth regime turns out useful as well: introduced in [10], it is a QED refinement of the ED regime.

As the number of servers increases, which is relevant for moderate to large call centers, these regimes can be formally characterized by relating the number of servers to the offered load. Recall that the offered load is given by  $R = \lambda/\mu$ .

**Efficiency Driven (ED) Regime:** The efficiency driven regime is characterized by very high servers utilization ( $\sim 100\%$ ) and relatively high abandonment rate (around 10%). In the ED regime, the offered load is noticeably larger than the number of agents  $N$ . This means that the system would collapse unless abandonment take place. The formal characterization of the ED regime is in terms of the following relationship between  $N$  and  $R$ :

$$N \approx R \cdot (1 - \varepsilon),$$

where  $\varepsilon > 0$  is a *QOS* parameter: a larger value of  $\varepsilon$  implies longer waiting times and more abandonment.

**Quality Driven (QD) Regime:** In the quality driven regime the emphasis is given to customers' service quality. This regime can be characterized by relatively low servers utilizations (for large call centers below 90%, and for smaller ones around 80% and perhaps less) and very low abandonment rate. Formally, this regime is characterized by:

$$N \approx R \cdot (1 + \varepsilon), \quad \varepsilon > 0.$$

**Quality and Efficiency Driven (QED) Regime:** This regime is the most relevant for call centers operation, since it combines a relatively high utilization of servers (around 95%) and low abandonment rate (1%-3%). As noted already, the square-root safety-staffing rule (3) remains valid in the model with abandonment, although now  $\beta$  depends not only on the service level  $\alpha = P\{W > 0\}$  but also on the abandonment rate  $\theta$ . The number of servers in this regime is given by the square-root formula

$$N \approx R + \beta \cdot \sqrt{R}, \quad -\infty < \beta < \infty, \quad (7)$$

where  $\beta$  is, as usual, the *QOS* parameter; note that  $\beta$  can now take negative values, while in the case of Erlang-C it is restricted to be positive in order to ensure stability.

**ED+QED Regime:** As describes in [10], this operational regime combines the staffing rules of the ED and QED regimes. It arose from the need to accommodate the constraint  $P(W > T) \leq \alpha$ , assuming that  $T$  is in the order of the service time and  $\alpha$  is not too close to 0 or 1. In this regime, the number of servers is characterized by the following formula:

$$N \approx R \cdot (1 - \varepsilon) + \beta \cdot \sqrt{R}, \quad 0 < \varepsilon < 1, \quad -\infty < \beta < \infty. \quad (8)$$

One observes that ED+QED staffing amounts to QED fine-tuning of ED staffing.

### 3 Our Workload Process and the Offered-Load

In this section, we introduce the workload process and the offered-load of the  $M_t/GI/N_t + GI$  queue. Here  $M_t$  indicates that the arrival process is assumed to be non-homogenous Poisson with arrival rate  $\lambda(t)$ ,  $t \geq 0$ . The first  $GI$  indicates that the service times are iid with cdf  $G$ . The  $N_t$  notation indicates that the number of servers is finite and can be varied over time. Finally, the last  $GI$  indicates that each customer has patience with general cdf, iid across customers. This  $M_t/GI/N_t + GI$  model is likely to be much more appropriate for approximating real life service system.

In stationary models, we have shown that the staffing level is determined by the offered load  $R$ . We would like to understand the required staffing level in a time varying environment. While the  $M_t/GI/N_t + GI$  model is mathematically intractable, our goal is still to determine the required staffing level at time  $t$ , in order to achieve a specific service performance.

A common approach to handle time varying arrival rates is to use a *Pointwise Stationary Approximation* (PSA). This method provides a time dependent description of performance based on the steady-state behavior of a corresponding stationary model, using the parameters that prevail at the time at which we carry out the analysis. For example, we can approximate the performance of  $M_t/GI/N_t + GI$  at time  $t$  by the steady-state performance in the associated stationary  $M/GI/s + GI$  queue, with the same service time and patience distributions, but with the constant arrival rate and number of servers equal to the values  $\lambda(t)$  and  $N_t$  respectively. Clearly, a steady-state assumption could be problematic when the arrival rate changes over time, especially in time periods when the variation of  $\lambda(t)$  is large. Indeed, in practice, PSA often seems to be inappropriate, as demonstrated in [5].

A first step to solve the staffing problem for time-varying systems is to understand better the offered load at time  $t$ . To this end, we introduce the  $M_t/GI/\infty$  model. This model differs from the previous one by having infinitely many servers, which means that each customer who joins the system is facing a ready-to-answer server and does not need to wait. Consequently, there is no need to consider the patience of customers in this model. In our study, we use the  $M_t/GI/\infty$  model for several reasons: First, it is remarkably tractable, as will be shown later. Second, one can use the  $M_t/GI/\infty$  model to approximate a system that aims at immediately answering all calls or to analyze the level of required capacity. Moreover, we can use this model to get an upper bound on the performance that could be achieved if the staffing level was as high as needed. But most importantly, as explained momentarily, analyzing  $M_t/GI/\infty$  yields the "right" definition of offered load for  $M_t/GI/N_t + GI$ .

Applying the above assumptions to  $M_t/GI/\infty$ , the number of arrivals in the time interval  $[a, b]$  has a Poisson distribution with mean  $\int_a^b \lambda(u) du$ . In the special case where  $\lambda(t) = \lambda$ , the steady-state number of busy servers has a Poisson distribution with mean  $\lambda \cdot E[S]$ , independently of the service time distribution beyond its mean (insensitivity). We will show that, under a time-varying rate  $\lambda(t)$ , the number of busy servers at time  $t$ ,  $L(t)$ , also has a Poisson distribution with a time-varying mean,  $R(t) = E[L(t)]$ . The measure  $R(t)$  will be interpreted as the **offered-load** at time  $t$ , and will be discussed in the next section.

### 3.1 The Offered Load of $M_t/GI/N_t + GI$

For the  $M_t/GI/N_t + GI$  queue, the *offered load*  $R = \{R(t), t \geq 0\}$  is given by  $R(t) = E[L(t)]$ , where  $L(t)$  is the number of customers (number of busy servers) at time  $t$ , in a corresponding  $M_t/GI/\infty$  queue (same arrivals and services). The stochastic process  $L = \{L(t), t \geq 0\}$  will be referred to as the *workload* process.

The following theorem provides three representations for  $R(t)$ , with respect to the service time distribution. They are proved in [3], and we re-derive them here through a different approach (based on discrete approximations and biased sampling).

**Theorem 3.** For each  $t$ ,  $L(t)$  has a Poisson distribution with mean

$$R(t) = E[L(t)] = E[\lambda(t - S_e)] \cdot E[S] = E\left[\int_{t-S}^t \lambda(u) du\right] = \int_{-\infty}^t [1 - G(t - u)]\lambda(u) du, \quad (9)$$

where

$S$  is a generic service time;

$S_e$  is a generic excess service time, with the following cdf:

$$P(S_e \leq t) = \frac{1}{E(S)} \int_0^t [1 - G(u)] du, \quad t \geq 0. \quad (10)$$

**Proof:** In order to prove the theorem, we shall prove the following lemmas, each holding for all  $t$ :

**Lemma 1.** :  $L(t)$  is Poisson distributed with mean  $R(t) = \int_{-\infty}^t [1 - G(t - u)]\lambda(u) du$ .

**Lemma 2.** :  $\int_{-\infty}^t [1 - G(t - u)]\lambda(u) du = E\left[\int_{t-S}^t \lambda(u) du\right]$ .

**Lemma 3.** :  $E\left[\int_{t-S}^t \lambda(u) du\right] = E[\lambda(t - S_e)] \cdot E[S]$

The combination of these three lemmas establishes Theorem 3.

**Proof of Lemma 1:** Let  $(x, y)$  be a coordinate that represent an arrival at time  $x$  with service time  $y$ . Then  $L(t)$  is the set of all points with  $x \leq t$  and  $x + y \geq t$ . Eick, Massey and Whitt have shown in [3] that the number of points in any finite collection of disjoint rectangles, on the two dimensional system of arrival time against service time, are independent Poisson variables. As a result,  $L(t)$  is Poisson distributed with mean

$$\begin{aligned} R(t) = E[L(t)] &= \int_{u=-\infty}^t \lambda(u) \cdot P(u + S > t) du = \int_{u=-\infty}^t \lambda(u) \cdot P(S > t - u) du = \\ &= \int_{u=-\infty}^t \lambda(u) \cdot [1 - G(t - u)] du. \quad \blacksquare \end{aligned}$$

**Proof of Lemma 2:** We start with  $E\left[\int_{t-S}^t \lambda(u) du\right] = \int_{s=0}^{\infty} \left[\int_{u=t-s}^t \lambda(u) du\right] dG(s)$ .

Then we change variables by setting  $x = u - t$ . Keeping in mind that  $\lambda(t)$  is non-negative, an order-interchange of the integration is justified. We thus get

$$\begin{aligned} E\left[\int_{t-S}^t \lambda(u) du\right] &= \int_{s=0}^{\infty} \left[\int_{x=-s}^0 \lambda(x + t) dx\right] dG(s) = \int_{x=-\infty}^0 \int_{s=-x}^{\infty} \lambda(x + t) dx dG(s) = \\ &= \int_{x=-\infty}^0 \lambda(x + t) \cdot [1 - G(-x)] dx = \int_{u=-\infty}^t \lambda(u) \cdot [1 - G(t - u)] du. \quad \blacksquare \end{aligned}$$

**Proof of Lemma 3:** Suppose that the service time is deterministic  $S=D$ . Then  $S_e$  is uniformly distributed on the interval  $[0, D]$ . In this case

$$\begin{aligned} E[\lambda(t - S_e)] \cdot E[S] &= D \cdot E[\lambda(t - S_e)] = D \cdot \int_0^D \frac{1}{D} \cdot \lambda(t - x) dx = \int_0^D \lambda(t - x) dx = \\ &= E \left[ \int_{t-D}^t \lambda(u) du \right]. \end{aligned} \quad (11)$$

For a general  $S$  we will use the fact that any random variable can be approximated as close as required by a discrete random variable. We thus assume that  $S = D_i$  with probability  $p_i$ . From biased sampling we get  $S_e \stackrel{d}{=} Uni(0, D_i)$  w.p.  $\frac{p_i \cdot D_i}{E[S]}$ . If we denote  $U_i = Uni(0, D_i)$ , then:

$$\begin{aligned} E \left[ \int_{t-S}^t \lambda(u) du \right] &= \sum_i p_i \cdot \int_{t-D_i}^t \lambda(u) du \stackrel{(11)}{=} \sum_i p_i \cdot D_i \cdot E[\lambda(t - U_i)] = \\ &= E[S] \cdot \sum_i \frac{p_i \cdot D_i}{E[S]} E[\lambda(t - U_i)] = E[\lambda(t - S_e)] \cdot E[S], \end{aligned} \quad (12)$$

which completes the proof of Theorem 3 (up to rigorously justifying the approximation of an arbitrary random variable  $S$  by a discrete one, and taking limits in (12).) ■

From the result of Theorem 3, one observes that when  $\lambda(t) = \lambda$ , the expression for  $R(t)$  becomes the offered load of a homogeneous arrival rate  $\lambda \cdot E[S]$ . Hence, we deduce that the expression of  $R(t)$  is very similar to that of the homogeneous arrival case, except for a random time lag in  $\lambda(t)$ . In addition, if  $\lambda(t)$  changes very little before  $t$ , (in comparison to the average service time) then  $R(t) \approx \lambda \cdot E[S]$ . This is actually consistent with the result of PSA.

## Taylor-Series Approximations

We now consider the first representation in (9):  $R(t) = E[\lambda(t - S_e)] \cdot E[S]$ . The time lag  $S_e$  appears inside the arrival rate function  $\lambda(t)$ , which appears inside the expectation. In case that  $\lambda(t)$  is non-linear, the calculation of this expression might be complicated. If  $\lambda(t)$  is polynomial, then we can express  $R(t)$  directly in terms of moments of  $S_e$  (See Theorem 10 in [3]). However, in many cases the arrival rate function will not be polynomial. In this case, if  $\lambda(t)$  is smooth we can approximate it by a Taylor series. In this manner, a first order approximation for the arrival rate function in a time interval before  $t$  will be  $\lambda(t - u) \approx \lambda(t) - \lambda^{(1)}(t) \cdot u$ , where  $\lambda^{(k)}(t)$  denotes the  $k^{th}$  derivative of  $\lambda(t)$  evaluated at time  $t$ . The second order Taylor series approximation is  $\lambda(t - u) \approx \lambda(t) - \lambda^{(1)}(t) \cdot u + \lambda^{(2)}(t) \cdot \frac{u^2}{2}$ .

From the results above, one obtains, as shown in [3], the first and second order approximations for  $R(t)$  respectively:

1.  $R(t) \approx \lambda(t - E[S_e]) \cdot E[S]$  ;
2.  $R(t) \approx \lambda(t - E[S_e]) \cdot E[S] + \frac{\lambda^{(2)}(t)}{2} \cdot Var(S_e) \cdot E(S)$ .

The first order approximation for  $R(t)$  is similar to PSA, but with a backward time shift of  $E(S_e)$ . This is often related to *lagged-PSA* [18]. From the second order approximation, we notice that there is also a space shift by  $\frac{\lambda^{(2)}(t)}{2} \cdot \text{Var}(S_e) \cdot E(S)$ . Since  $\lambda^{(2)}(t)$  is negative at a local peak of  $\lambda(t)$ , and  $\text{Var}(S_e) \cdot E(S)$  is always positive, we conclude that the actual offered load at times of peak demand are shifted down.

### 3.2 The Gap\Lag - Examples

A good example for the effect of the time-gap  $S_e$ , described in Theorem 3, appears in [14]. In Section 6 of this thesis, empirical examples were presented to compare between several staffing methods. It has been shown that staffing according to PSA leads to over-staffing in periods when the offered load increases, and results in under-staffing in periods when it decreases. Another method that was applied was the *lagged PSA* which is described in the next subsection. The approximation to  $R(t)$  under this latter method is taken to be the PSA approximation with time lag of one mean service time,  $\lambda(t - E[S]) \cdot E[S]$ . This method performed better than PSA, but still encountered lack of stability in service performance. In general, PSA performs quite well for fast service rates (fast relative to variations of the arrival rate function). If service rates are slower, then the lagged PSA is more appropriate.

Another example, which we believe is related to the described phenomenon of the lag of the offered load behind the arrival rate, is introduced in [9]. This paper describes a study, based on data from two Massachusetts hospitals, on the diversion of ambulances from emergency departments due to congestion. Specifically, a diversion to another emergency department is made when the emergency department is fully occupied, and thus is not able to accept new patients. We focus on the results of the first hospital described in this paper (which is referred to in the study as "Hospital A"). Observations of the number of arrivals to the emergency department of this hospital and the number of diversions from it, per hour, were collected over period of 42 days. Then, an average of these measures by hour of the day was made over the 42 days. By this procedure, two series of 24 records were constructed: the average number of arrivals to the emergency department per hour of the day and average number of diversions from the emergency department per hour of the day.

The average service time was found to be approximately 6 hours and therefore the writers of the paper decided to build 6 more series, based on the series of the arrivals. The first series was constructed by the addition of the arrivals of each hour to the arrivals of the previous hour. The second series was constructed by the addition of the arrivals of the first series to the arrivals of two previous hours, and so on. In this manner, each point in the last series is actually the sum of the arrivals in the 6 hours prior to it.

The interesting insight emerging from this analysis is that even in the sixth series, where the arrivals of last 6 hours (which is the mean service time) are taken into account, there is a lag of the average number of diversions behind this series. This is another example of the fact that taking into account only the mean service time might not be good enough for the analysis of the current congestion in the system. This can be explained by the variance of the service time. Consider the relation  $R(t) = E \left[ \int_{t-S}^t \lambda(u) du \right]$  from Theorem 3, and compare it to

$\int_{t-E(S)}^t \lambda(u)du$ , which is what was used in [9]. If the service time was deterministic with the value of 6 hours, then under the described procedure, the cumulative arrival rate of the last 6 hour will be exactly the time-dependant offered load. In the cases, however (depending on the distribution of  $S$  and the function  $\lambda(t)$ ), this is totally different from  $R(t)$ .

### 3.3 Staffing the $M_t/GI/N_t + GI$ Queue

The staffing problem is to determine the minimal required number of agents subject to pre-specified satisfactory quality of service constraints. In the  $M_t/GI/N_t + GI$  model, one allows changing the number of servers as function of time. In practice, the staffing level cannot be changed continuously. We regard the period during which the staffing level must be kept constant as the *staffing interval*. In the analysis of the arrivals to call centers, we consider two sources of variability: *predictable variability* - changes over time of the expected number of arrivals and *stochastic variability* - random fluctuations around this expected number.

We already mentioned PSA, which treats the system at time  $t$  as if it is in steady state, with the arrival rate of time  $t$ . However, this approximation is designed to cope only with the stochastic variability. Hence, whenever the predicted variability becomes significant or the service times are not relatively short, PSA tends to be inappropriate. The first component of Theorem 3 shows that PSA is correct except for a random time lag  $S_e$  (the stationary-excess service time). A simple refinement for PSA, that relies on the time shift from the Taylor-series approximation, is the lagged PSA. In lagged PSA we approximate the offered load at time  $t$ ,  $R(t)$ , by  $\lambda(t - E[S_e]) \cdot E[S]$ .

Another approximation for the performance of the  $M_t/GI/N_t + GI$  model is the MOL approximation. The approximation applies the performance in an associated stationary  $M/GI/s + GI$  model. In the MOL approximation, we approximate the time-varying performance at time  $t$  by a *stationary* model where we take the offered load to be  $R(t)$ . Since the stationary offered load is  $\lambda \cdot E(S)$ , we use at time  $t$  the model with homogenous Poisson arrival process, namely  $M/GI/s + GI$ , with arrival rate  $\lambda_{MOL}(t) = \frac{R(t)}{E(S)}$ . Thus, the MOL method suggests staffing in the  $M_t/GI/N_t + GI$  model by the square-root formula  $N_t = R(t) + \beta \cdot \sqrt{R(t)}$ , where  $\beta$  is the related service quality parameter of the stationary model as given by the Garnett function (6). It is important to note that the validity of that refinement depends on the assumption of nonhomogeneous Poisson arrival process.

The MOL approximation is supported by the work in [5]. The authors of [5] introduced an effective staffing algorithm, called the *Iterative Staffing Algorithm* (ISA). ISA is designed to deal with both predictable and stochastic variability to achieve stable performance over the day. It is shown in [5] that to achieve time stable performance in the face of time-varying arriving rates it is enough to target a stable delay probability, which then yields stable performance of several other operational measures. More specifically, it is shown in [5] that for the time-varying  $M_t/GI/N_t + GI$  model, it is actually possible to reduce the hard problem of time-varying staffing to an associated stationary staffing problem. Indeed, a time-varying square-root staffing, i.e.  $N(t) = R(t) + \beta \cdot \sqrt{R(t)}$ , applies ( $R(t)$  is given in (9)). This staffing level gives rise to a remarkable time-stable performance in which the delay probability is constantly  $\alpha$ . In the case

of  $M_t/M/N_t + M$ , the relation between  $\alpha$  and  $\beta$  in the time-varying model is in fact the Garnett function (6), derived for the stationary Erlang-A model.

## 4 Our Proposed Research

One of our main goals in this work is to dynamically determine the required staffing level of a service system, ideally so that no customers queue up and no servers idle. *Achieving the latter two ideals amounts to knowing  $L(t)$ ,  $t \geq 0$  precisely.* For this purpose, we must first decide on the performance target. Then we must predict the amount of demand for service in the near future (the next point of time when we can change the staffing level). In this paper we regard the *lead time* of a prediction as the time period between the time when the prediction is made (current time) and the time point that is related to the prediction value (future). For example, assume that we are now at time  $t$  and wish to predict  $L(t + s)$ , then  $s$  is the lead time. We can take the lead time of the prediction to be very short, when we deal with a dynamic service system where, for example, the arrival rate variation is high or when the service parameters, such as staffing level, can be changed rapidly. On the other hand, it can be applied over longer time periods as well.

In order to predict the exact demand at time  $t$  (within the lead time), one could first estimate its mean value  $R(t)$ , then finely predict the number of required servers,  $L(t)$ , from the (approximated Poisson) distribution of  $L(t)$ . Under this approach, one should be careful with estimating  $R(t)$ , to take into account that (due to abandonment) not all customers' service times are observable from the historical data.

Another approach, which we hope that will yield better prediction of  $L(t)$ , is to divide the future demand at time  $t$  into two different components: those who are currently in system and will stay at least until time  $t$ , and those who will come before time  $t$  and will stay after it. Under the assumptions of our model, it is reasonable to regard these future demand components as independent. Moreover, as the time interval between staffing changes decreases (and respectively the lead time for our prediction is smaller), with respect to the average service time, then the significance of current calls in the total future demand increases [17]. From the other point of view, as the average service time increases, the current number of calls in the system becomes more significant for estimating future demand. The latter is relevant, for example, when considering healthcare services, where service times last several hours. The current calls in progress may also provide useful information for predicting their remaining service times (according to their class and service time distribution). In this sense, one must predict both the arrivals and service times (and remaining service times). We concentrate here on estimating the service times.

Consider the case where we have all required operational data of every call that reached our service system and we wish to predict  $L(t)$  within a certain lead time. A naive solution is to predict it from the available historical records. This could be good enough, unless the *potential service times of customers who abandoned* (and therefore their service times are not available) *have different distributions from those who waited and reached service.* This might be a very realistic possibility. Hence, we would like first to check the relationship between patience and service time. Then we shall impute the service times of those who abandoned, which cannot be

observed from the operational data. Finally, we shall try to predict the patience and (remaining) service times of the customers who are currently in the system or will come before the lead time is reached.

#### 4.1 Imputing Service Times

We start with the following notations:

- $S$  - The service time of a customer.
- $\tau$  - The (im)patience of a customer.
- $V$  - The virtual-waiting-time (or offered-waiting-time), namely the time a customer is required to wait before entering service. In other words, this is the time a customer with an infinite patience would have waited.
- $W$  - The waiting time of a customer. The waiting time is defined as the minimum between  $V$  and  $\tau$ , which is the actual time a customer waits, for both served and abandoning customers.

Assume that each customer has an associated pair of random variables  $\tau$  and  $S$  as above, that are characterized by the customer, whereas the random variable  $V$  is independent of the customer (as in telephone services, where the customer does not observe the queue) and it is determined by system conditions.

The following properties naturally follow from the above discussion:

- The pair  $(\tau, S)$  is independent of  $V$ .
- For those who abandon, we observe  $V$  censored by their (im)patience. For those served, we observe  $V$ .
- $W$  is observable for all customers.
- We observe  $S$  only for customers who have  $\tau > V$ .

To learn about the relation between patience and service time, we would like to estimate

$$E(S|\tau = w), \quad w > 0. \quad (13)$$

Following Ritov [13], we propose a non-linear regression method:

$$S_i = g(W_i) + \varepsilon_i, \quad (14)$$

where  $g(w) = E(S|\tau > W = w)$  is the mean service time of those who waited exactly  $w$  units of time and were served, and  $S_i$  and  $W_i$  are the service and waiting times of customer  $i$  respectively.

**Lemma 4.** A simpler expression for  $g$  is given by:

$$g(w) = E(S|\tau > w). \quad (15)$$

**Proof:** Note that  $g(w) = E(S|\tau > W = w) = E(S|\tau > w, W = w)$ , and that the events  $\{\tau > w, W = w\}$  and  $\{\tau > w, V = w\}$  are equivalent. Hence, we get  $g(w) = E(S|\tau > w, V = w)$ . For simplicity, we avoid events of probability zero by assuming that all variables are discrete. Then:

$$\begin{aligned} g(w) &= E(S|\tau > w, V = w) = \sum_s s \cdot P(S = s|\tau > w, V = w) = \\ &= \sum_s s \cdot \frac{P(S = s, \tau > w|V = w)}{P(\tau > w|V = w)} = \sum_s s \cdot \frac{P(S = s, \tau > w)}{P(\tau > w)}, \end{aligned}$$

where the last equality holds because of the independence of  $(\tau, S)$  and  $V$ . It is easy to see that the final expression here is exactly  $E(S|\tau > w)$ , which concludes our proof. ■

At this stage, we propose a method to estimate the service times of those who have patience  $\tau = w$ . We start with:

$$\begin{aligned} g(w) &= E(S|\tau > w) = \int_{s=0}^{\infty} s \cdot f_{S|\tau > w}(s|\tau > w) ds = \int_{s=0}^{\infty} s \cdot \frac{f_{S, \tau > w}(s, \tau > w)}{P(\tau > w)} ds = \\ &= \frac{\int_{s=0}^{\infty} s \cdot \int_{\tau=w}^{\infty} f_{S, \tau}(s, u) du ds}{P(\tau > w)} = \frac{\int_{s=0}^{\infty} s \cdot \int_{\tau=w}^{\infty} f_{S|\tau}(s|u) \cdot f_{\tau}(u) du ds}{\int_w^{\infty} f_{\tau}(u) du} \\ &= \frac{\int_{\tau=w}^{\infty} (\int_{s=0}^{\infty} s \cdot f_{S|\tau}(s|u) ds) \cdot f_{\tau}(u) du}{\int_w^{\infty} f_{\tau}(u) du} \\ &= \frac{\int_{\tau=w}^{\infty} f_{\tau}(u) \cdot E(S|\tau = u) du}{\int_{\tau=w}^{\infty} f_{\tau}(u) du}. \end{aligned}$$

We arrived at

$$g(w) = \frac{\int_{\tau=w}^{\infty} f_{\tau}(u) \cdot E(S|\tau = u) du}{\int_{\tau=w}^{\infty} f_{\tau}(u) du}.$$

By multiplying both sides by  $\int_{\tau=w}^{\infty} f_{\tau}(u) du$ , we get

$$g(w) \cdot \int_{\tau=w}^{\infty} f_{\tau}(u) du = \int_{\tau=w}^{\infty} f_{\tau}(u) \cdot E(S|\tau = u) du.$$

Differentiating both side of the last equation with respect to  $w$  yields

$$g'(w) \cdot \int_{\tau=w}^{\infty} f_{\tau}(u) du - g(w) \cdot f_{\tau}(w) = -f_{\tau}(w) \cdot E(S|\tau = w).$$

We conclude that

$$E(S|\tau = w) = g(w) - g'(w) \cdot \frac{\int_{\tau=w}^{\infty} f_{\tau}(u) du}{f_{\tau}(w)} = g(w) - \frac{g'(w)}{h_{\tau}(w)}, \quad (16)$$

where  $h_{\tau}(w)$  is the hazard rate function of  $\tau$ .

Examining equation (16) reveals that in order to estimate  $E(S|\tau = w)$ , it suffices to estimate the following expressions:

1. The hazard rate function  $h_\tau(w)$ .
2. The function  $g(w)$  and its derivative  $g'(w)$ .

We thus intend to carry out the two steps above and deduce  $E(S|\tau = w)$ ,  $w > 0$ .

## 4.2 Staffing Aiming to Immediately Answer All Calls [17]

After imputing the service times of those who abandoned, we would like to exploit our findings in order to determine the staffing level of the call center. We are interested in applying dynamic staffing aiming to *immediately* answer all calls. Of course, actually answering all calls immediately may be an unrealistic objective, but it is often possible to come very close to this goal by dynamically staffing based on recently updated information. As mentioned before, for this purpose, the use of infinite-server  $M_t/GI/\infty$  queueing model is quite natural.

In this part of our study we will base our research on the results of Whitt in [17]. The main idea here is to dynamically use all available information up to current time in order to more accurately predict the demand in a relatively short time  $t$  in the future. We will use the term *current time* for the time at which we want to make the staffing decision and the term *prediction time* for the time which we want to make the prediction for.

The prediction of future demand requires prediction of the arrival rate function, the service times of new calls that will come before the prediction time and the elapse holding time of calls that are currently in progress. In [17], a method to apply dynamic staffing is suggested, assuming that the cdf of the service time  $S$  and the elapse service time  $S_e$  can be estimated from historical data. Of course, a modification should be done to accommodate the distributions of possibly abandoning customers (as calculated in section 4.1). This amounts to imputing the values of the service times that could not be observed, due to abandonments, in the prediction of the future amount of work in the system. This modification is essential if the distribution of the service times of those who abandon will be found to differ from those who do not.

## 5 Data Sources

### DataMOCCA [16]

The DataMOCCA Project was initiated by researchers from the Technion and the Wharton School of the University of Pennsylvania. The mission of the project is to collect, pre-process, organize and analyze data from Telephone Call/Contact Centers. DataMOCCA (Data MOdel for Call Center Analysis) is a universal model for call center data that, together with its graphical user interface STATCCA (Statistical Call Center Analyzer), enables real-time statistical analysis spanning seconds-to-months resolutions. Currently, DataMOCCA covers call-by-call data of three large call centers: a U.S. bank and an Israeli Telecom company, both over periods of 2.5-3 years each, and an Israeli bank, over several months (and more coming). For example, the U.S. bank data has close to 220 million calls, out of which about 40 million were served by agents and the rest by the Interactive Voice Response (IVR) system.

## **Healthcare Data [11]**

Another data source of potential interest is that of Emergency Departments (ED) in 5 Israeli hospitals was gathered in a study on analyzing of emergency department performance [11]. ED data is relevant, as clear from our description of the ambulance-diversion, in section 3.2. This ED data source includes 16,250 records of procedures that the patients of the emergency departments had to go through. In addition to the data gathered through the study, the data includes also 24 month of historical patient data from 4 of the 5 hospitals computerized information.

## References

- [1] C. Borst, A. Mandelbaum, and M. Reiman, *Dimensioning large call centers*, Operations Research **52** (2004), no. 1, 17–34.
- [2] N. Channouf, P. L’Ecuyer, A. Ingolfsson, and A. Avramidis, *The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta*, Health Care Management Science **10** (2007), no. 1, 25–45(21).
- [3] S.G. Eick, W.A. Massey, and W. Whitt, *The physics of the  $M_t/G/\infty$  queue*, Operations Research **41** (1993), no. 4, 731–742.
- [4] A.K. Erlang, *The theory of probabilities and telephone conversations*, Nyt Tidsskrift Mat. D 20 (1909), 33–39.
- [5] Z. Feldman, A. Mandelbaum, W.A. Massey, and W. Whitt, *Staffing of time-varying queues to achieve time-stable performance*, Management Science (2007).
- [6] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, Manufacturing and Service Operations Management **4(3)** (2002), 208–227.
- [7] L. Green, *Capacity planning and management in hospitals*, Operations Research and Health Care (2004).
- [8] S. Halfin and W. Whitt, *Heavy-traffic limits for queues with many exponential servers*, Operations Research **29** (1981), no. 3, 567–588.
- [9] E. Litvak, M.L. McManus, and A. Cooper, *Root cause analysis of emergency department crowding and ambulance diversion in Massachusetts*, Boston University Program for Management Variability in Health Care Delivery (2002).
- [10] A. Mandelbaum and S. Zeltyn, *Staffing many-server queues with impatient customers: constraint satisfaction in call centers*, (2007), Under revision for Operations Research.
- [11] Y. Marmor, *Developing a simulation tool for analyzing emergency department performance*, M.Sc. Thesis, Technion (2003).
- [12] C. Palm, *Research on telephone traffic carried by full availability groups*, Tele **1** (1957), 107, (English translation of results first published in 1946 in Swedish in the same journal which was entitled Tekniska Meddelanden fran Kungl. Telegrafstyrelsen.).
- [13] Y. Ritov, Private communication.
- [14] L. Rozenshmidt, *On priority queues with impatient customers: Stationary and time-varying analysis*, M.Sc. Thesis, Technion (2007).
- [15] S. Steckley, S. Henderson, and V. Mehrotra, *Service system planning in the presence of a random arrival rate*, (2004), Downloadable from: <http://legacy.orie.cornell.edu/shane/pubs>.
- [16] V. Trifomov, P.D. Feigin, A. Mandelbaum, e. Ishay, and E. Nadjarov, *DataMOCCA: Model Description and Introduction to User Interface*, (2006), [http://iew3.technion.ac.il/serveng/References/DataMOCCA\\_Volume1-V2.3\(for%20Circulation\)\\_Ella.pdf](http://iew3.technion.ac.il/serveng/References/DataMOCCA_Volume1-V2.3(for%20Circulation)_Ella.pdf).

- [17] W. Whitt, *Dynamic staffing in a telephone call center aiming to immediately answer all calls*, Operations Research Letters **24** (1999), 205–212.
- [18] ———, *What you should know about queueing models to set staffing requirements in service systems*, Naval Research Logistics **54** (2007), no. 5, 476–484.