The Palm/Erlang-A Queue, with Applications to Call Centers*

Avishai Mandelbaum and Sergey Zeltyn†

Faculty of Industrial Engineering & Management
Technion,
Haifa 32000, ISRAEL

emails: avim@tx.technion.ac.il, zeltyn@ie.technion.ac.il

June 19, 2005

Contents

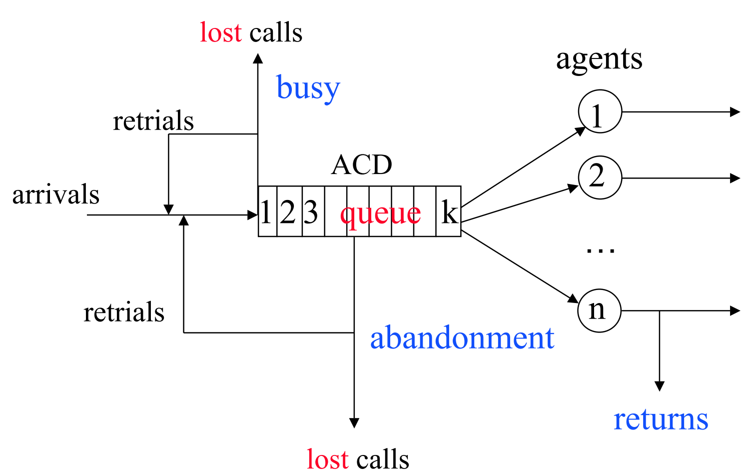
*Parts of the text are adapted from [11], [19], [22], [28] and [40]

†**Acknowledgements.** The research of both authors was supported by BSF (Binational Science Foundation) grant 2001685/2005175, ISF (Israeli Science Foundation) grants 388/99, 126/02 and 1046/04, the Niderzaksen Fund and by the Technion funds for the promotion of research and sponsored research.

1 Introduction

Service Engineering is a newly emerging discipline that seeks to develop scientifically-based engineering principles and tools, often culminating in software, which support the design and management of service operations. **Contact Centers** are service organizations that serve customers via the phone, fax, e-mail, chat or other tele-communication channels. A particularly important type of a contact center is the **Call Center**, which predominantly serves phone calls. Due to advances in Information and Communication Technology, the number, size and scope of contact centers, as well as the number of people who are employed there or use them as customers, grows explosively. For example, in the U.S. alone, the call center industry is estimated to employ several million agents which, in fact, outnumbers agriculture [22]. In Europe, the number of call center employees in 1999-2000 was estimated, for example, by 600,000 in the UK (2.3% of the total workforce) and 200,000 in Holland (almost 3%) [6]; Bittner et al. [8] assess that, in Germany in 2001, there were between 300,000 to 400,000 (1-2%) employed in the call center industry.

Figure 1: **Schematic representation of a telephone call center**



In a large performance-leader call center, many hundreds of agents could serve many thousands of calling customers per hour; agents' utilization levels exceed 90%, yet about 50% of the customers are answered immediately upon calling; callers who are delayed demand a response within seconds, the vast majority gets it, and scarcely few of the rest, say 1% of those calling, abandon during peak-congestion due to impatience. But most call centers are far from achieving such levels of performance. To these, scientific models are prerequisites for climbing the performance ladder, and the model described in this paper, namely **Palm/Erlang-A**, should constitute their starting point. See Gans, Koole and Mandelbaum [17] or Helber and Stolletz [23] for reviews of state-of-the-art of research on telephone call centers. In addition, Mandelbaum [26] provides a comprehensive bibliography, namely references plus abstracts, of call-center-related research papers.

Modelling a Call Center. A simplified representation of traffic flows in a call center is given in Figure 1. Incoming calls form a single queue, waiting for service from one of n statistically identical agents. There are $k + n$ telephone trunk-lines. These are connected to an Automatic Call Distributor (ACD) which manages the queue, connects customers to available agents, and also archives operational data. Customers arriving when all lines are occupied encounter a busy signal. Such customers might try again later ("retrial") or give up ("lost call"). Customers who succeed in getting through at a time when all agents are busy (that is, when there are at least n but fewer than $k + n$ customers within the call center), are placed in the queue. If waiting customers run out of patience before their service begins, they hang up ("abandon"). After abandoning, customers might try calling again later while others are lost. After service, there are "positive" returns of satisfied customers, or "negative" returns due to complaints.

Note that the model in Figure 1 ignores multiple service types and skilled-based routing that are present in many modern call centers. However, a lot of interesting questions still remain open (see Section 10) even for models with homogeneous servers/customers.

In basic models, the already simple representation in Figure 1 is simplified even further. Specifically, in the present paper we assume that there are enough trunk-lines to avoid busy signals ($k = \infty$). This assumption prevails in today's call centers. In addition, we assume out retrials and return calls, which corresponds to absorbing them within the primitive arrival process. (See, for example, Aguir et al. [2] for an analysis that takes retrials into account.) However, and unlike most models used in practice, here we do acknowledge and accommodate abandonment. The reasons for this will become clear momentarily.

2 Significance of abandonment in practice and modelling

The classical M/M/n queueing model, also called **Erlang-C**, is the model most frequently used in *workforce management* of call centers. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed service times with rate μ, and n independent statistically-identical agents. (Time-varying arrival rates are accommodated via piecewise constant approximations.) But Erlang-C does not acknowledge abandonment. This, as will now be argued, is a significant deficiency: customer abandonment is not a minor, let alone a negligible, aspect of call center operations.

According to the "Help Desk and Customer Support Practices Report, 1997" [24], more than 40% of the call centers set a target for the fraction of abandonment, but in most cases this target is not achieved. Moreover, the lack of understanding of the abandonment phenomenon and the scarcity of models that acknowledge it, has lead practitioners to ignore it altogether. (For example, this lead [13] to conclude that abandonment is "not a good indicator of call center performance".) But models which ignore abandonment either distort or fail to provide information that is important to call center managers. We now support this last statement, first qualitatively and then quantitatively.

- Abandonment statistics constitute the only ACD measurement that is customer-subjective: those who abandon declare that the service offered is not worth its wait. (Other ACD data, such as average waiting times, are "objective"; they also do not include the only other customer-subjective operational measures, namely retrial/return statistics.)

- Some call centers focus on the average waits of only those who get served, which does not acknowledge abandoning customers. But under such circumstances, the service-order that optimizes performance is LIFO = Last-In-First-Out [18], which clearly suggests that a distorted focus has been chosen.

- Ignoring abandonment can cause either under- or over-staffing: On the one hand, if service level is measured only for those customers who reach service, the result is unjustly optimistic - the effect of an abandonment is less delay for those further back in line, as well as for future arrivals. This would lead to under-staffing. On the other hand, using workforce management tools that ignore abandonment would result in over-staffing as actually fewer agents are needed in order to meet most abandonment-ignorant service goals.

The Palm/Erlang-A model: Palm [31] introduced a simple (tractable) way to model abandonment. He suggested to enrich Erlang-C (M/M/n) in the following manner. Associated with each arriving caller there is an exponentially distributed *patience time* with mean θ^{-1}. An arriving customer encounters an *offered waiting time*, which is defined as the time that this customer would have to wait given that her or his patience is infinite. If the offered wait exceeds the customer's patience time, the call is then abandoned, otherwise the customer awaits service. The patience parameter θ will be referred to as the individual *abandonment rate*. (We shall omit the "individual" when obvious.) Following Baccelli and Hebuterne [5], we denote this model by M/M/n+M, and refer to it as Palm/Erlang-A, or Erlang-A for short. Here the **A** stands for **A**bandonment, as well as for the fact that the model interpolates between Erlang-C and Erlang-B. (The latter is the M/M/n/n model, in which there are n trunk lines (k=0), hence customers that cannot be served immediately are blocked.)

With Erlang-A, the quantitative significance of abandonment can be demonstrated through simple numerical examples. We start with Figure 2, which shows the fraction of delayed customers and the average wait, when calculated via Erlang-C (M/M/n), and a corresponding Erlang-A (M/M/n+M) model. In both models, the arrival rate is 48 calls per minutes, the average service time equals 1 minute, and the number of agents is varied from 35 to 70. Average patience is taken to be 2 minutes for the Erlang-A model. Clearly, the two curves convey rather different pictures of what is happening in the system they depict, especially within the range of 40 to 50 agents: in particular, and as shown below, Erlang-C is stable only with 49 or more agents, while Erlang-A is always stable.

The above M/M/n and M/M/n+M models are further compared in Table 1. Note that exponential patience with an average of 2 minutes gives rise to 3.1% abandonment. Then note that the average wait and queue length are both strikingly shorter, and this with only 3.1% abandonment taking place. Indeed, "The fittest survive" and wait less - much less. (Significantly, this high-level performance is *not* achieved if merely the arrival rate to the M/M/n system is decreased by 3.1%; for example, the "average speed of answer" in such a case is 8.8 seconds, compared with 3.7 seconds. The reason is that abandonment reduce workload precisely when needed, namely when congestion is high.) Finally, note that system performance in such heavy traffic is very sensitive to staffing levels. In our example, adding 3 or 4 agents (from 50 to say 54) to M/M/n would result in M/M/n+M performance, as emerging from the horizontal distance between the graphs in Figure 2. Nonetheless, since personnel costs are the major operational costs of running call centers (typical estimates run at about 60-75% of the total), even a 6%-8% reduction in personnel is economically significant (and much more so for large call centers that employ thousands of agents).

Figure 2: **Comparison between Erlang-A and Erlang-C** [19, 39]
48 calls per min., 1 min. average service time, 2 min. average patience

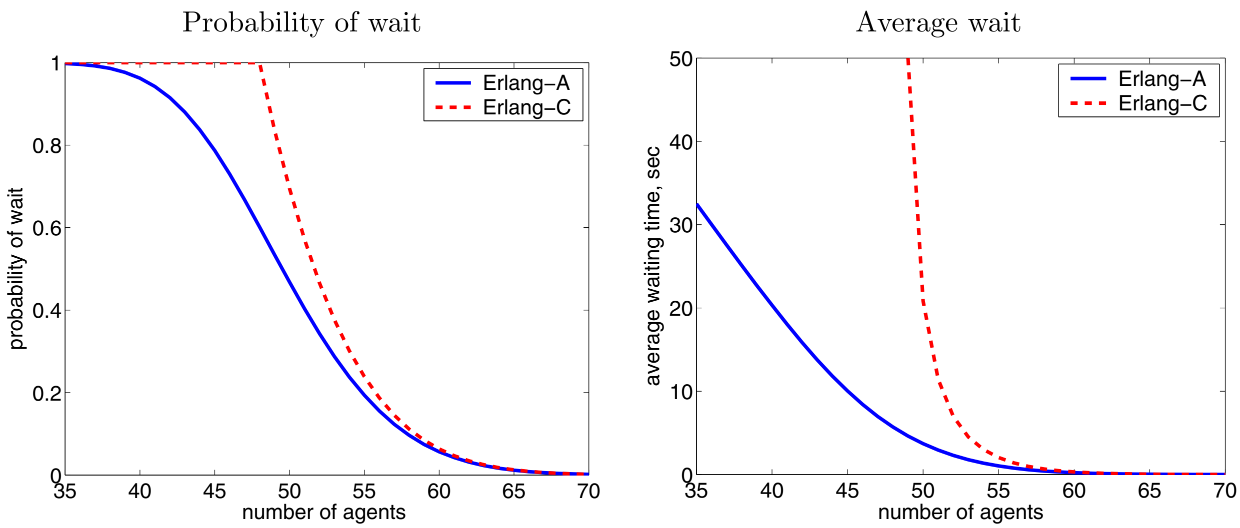


Table 1: **Comparing models with/without abandonment**
50 agents, 48 calls per min., 1 min. average service time, 2 min. average patience

	M/M/n	M/M/n+M	M/M/n, $\lambda \downarrow 3.1\%$
Fraction abandoning	–	3.1%	-
Average waiting time	20.8 sec	3.7 sec	8.8 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec	28.2 sec
Average queue length	17	3	7
Agents' utilization	96%	93%	93%

As a final demonstration of the significance of abandonment, we now use it to explain a phenomenon that has puzzled queueing theorists: It is the observation that, in practice, simple deterministic approaches often lead to surprisingly good results. For example, consider a call center with averages of 6000 calls per hour and service time of 4 minutes. Such a call center gets an average of $(6000 : 60) \cdot 4 = 400$ minutes of work per minute. The deterministic approach then prescribes 400 service agents to cope with this load (1 agent-minute per 1 work-minute), which is a questionable recommendation according to standard queueing models. For example, Erlang-C would then be unstable, and its waiting times and queue-lengths would increase indefinitely. But now assume that customers abandon, as they actually do, and assign a reasonable parameter to their average patience, say equal to the average service time. Then, under Erlang-A, about 50% of the customers would be answered immediately upon calling, the average wait would be a mere 5 seconds, agents' utilization would be 98%, and all this at the cost of 2% abandonment – a remarkable performance indeed. (See Remark 6.1 in Section 6 for a more formal explanation.)

3 Birth-and-death process representation; Steady-state

The Erlang-A model is characterized by four parameters:

- λ – arrival rate (calls per unit of time);
- μ – service rate ($1/\mu$ is the average duration of service);
- n – number of servers/agents;
- θ – individual abandonment rate ($1/\theta$ is the average patience).

Figure 3: **Schematic representation of the Erlang-A model**

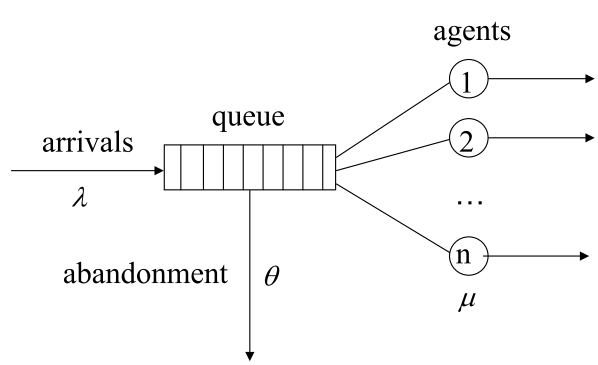


Figure 3 provides a representation of the traffic flows in Erlang-A, and a comparison with Figure 1 clearly reveals its limitations. (Nevertheless, and as we hope to demonstrate, Erlang-A still turns out very useful and insightful, both theoretically and practically.)

More formally, in the Erlang-A model customers arrive to the queueing system according to a Poisson(λ) process. Customers are equipped with *patience times* τ that are exp(θ), i.i.d.

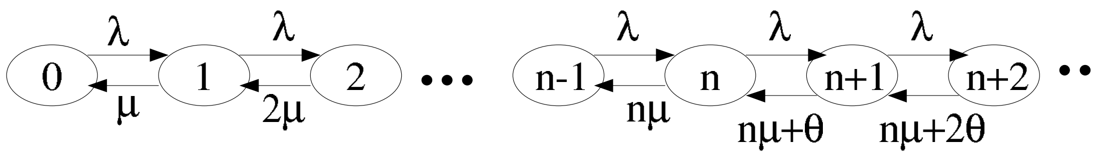
across customers. And service times are i.i.d. $\exp(\mu)$. Finally, the processes of arrivals, patience and service are mutually independent.

For a given customer, the patience time τ is the time that the customer is *willing* to wait for service - a wait that reaches τ results in an abandonment. Let V denote the *offered waiting time* - the time a customer, equipped with infinite patience, *must* wait in order to get service. The actual waiting/queueing time then equals

$$W = \min\{V, \tau\}\,.$$

Denote by $L(t)$ the total number-in-system at time t (served plus queued). Then $L = \{L(t), t \geq 0\}$ is a Markov birth-and-death process, with the following transition-rate diagram:

Figure 4: **Transition-rate diagram of the Erlang-A model**



Remark 3.1 Let d_j stand for the death-rate in state j, $0 \leq j < \infty$. Then

$$j \cdot \min(\mu, \theta) \leq d_j \leq j \cdot \max(\mu, \theta)\,. \tag{3.1}$$

The bounds on the left-hand and right-hand sides of (3.1) correspond to death-rates of an M/M/∞ queue with service rates $\min(\mu, \theta)$ and $\max(\mu, \theta)$, respectively. In some sense, which can be made precise via *stochastic orders* between distributions, these two M/M/∞ queues provide lower and upper (stochastic) bounds for the Erlang-A system. (The lower bound will be used later to prove that Erlang-A *always* reaches steady-state.) In the special case of equal service and abandonment rates ($\mu = \theta$), the Erlang-A and M/M/∞ models in fact coincide. ■

As customary, define the limiting distribution of L by:

$$\pi_j \;=\; \lim_{t\to\infty} \mathrm{P}\{L(t) = j\}, \qquad j \geq 0\,.$$

When existing, the limit distribution is also a steady-state (or stationary) distribution, which is calculated via the following version of the steady-state equations:

$$\begin{cases} \lambda\pi_j &= (j+1)\cdot\mu\pi_{j+1}, \quad 0 \le j \le n-1 \\ \lambda\pi_j &= (n\mu + (j+1-n)\theta)\cdot\pi_{j+1}, \quad j \ge n. \end{cases} \tag{3.2}$$

It is now straightforward to derive the "recipe" solution:

$$\pi_j = \begin{cases} \frac{(\lambda/\mu)^j}{j!}\pi_0\,, & 0 \le j \le n \\ \prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\pi_0\,, & j \ge n+1, \end{cases} \tag{3.3}$$

where

$$\pi_0 = \left[\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty}\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!}\right]^{-1}. \tag{3.4}$$

Our solution makes sense - equivalently the Markov process L is ergodic - if the infinite sum in (3.4) converges, which is a consequence of the lower bound in (3.1):

$$\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=n+1}^{\infty}\prod_{k=n+1}^{j}\left(\frac{\lambda}{n\mu+(k-n)\theta}\right)\frac{(\lambda/\mu)^n}{n!} \le \sum_{j=0}^{\infty}\frac{(\lambda/\min(\mu,\theta))^j}{j!} = e^{-\lambda/\min(\mu,\theta)}.$$

Formulae (3.3) and (3.4) include infinite sums that can cause numerical problems. To overcome these, Palm [31] represented the Erlang-A steady-state distribution, and some of its important performance measures, in terms of special functions. To this end, define the *Gamma function*

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt\,, \qquad x > 0,$$

and the *incomplete Gamma function*

$$\gamma(x,y) \triangleq \int_0^y t^{x-1}e^{-t}dt\,, \qquad x > 0,\ y \ge 0.$$

(Both functions can be calculated in Matlab.) Let

$$A(x,y) \triangleq \frac{xe^y}{y^x}\cdot\gamma(x,y) = 1 + \sum_{j=1}^{\infty}\frac{y^j}{\prod_{k=1}^{j}(x+k)}\,, \qquad x > 0,\ y \ge 0. \tag{3.5}$$

(The second equality is taken from Palm [31].)

In addition, let $E_{1,n}$ denote the *blocking probability* in the M/M/n/n (Erlang-B) system, and recall the classic Erlang-B formula:

$$E_{1,n} = \frac{\frac{(\lambda/\mu)^n}{n!}}{\sum_{j=0}^{n}\frac{(\lambda/\mu)^j}{j!}}\,. \tag{3.6}$$

A simple way for calculating $E_{1,n}$ is the recursion

$$E_{1,0} \;=\; 0; \qquad E_{1,n} \;=\; \frac{\rho E_{1,n-1}}{1+\rho E_{1,n-1}}\,, \qquad n \geq 1,$$

in which ρ is the offered load per agent, namely

$$\rho \;\triangleq\; \frac{\lambda}{n\mu}\,.$$

In the Appendix, we deduce from (3.5) the following solution for the steady-state distribution:

$$\pi_j \;=\; \begin{cases} \pi_n \cdot \dfrac{n!}{j! \cdot \left(\frac{\lambda}{\mu}\right)^{n-j}}\,, & 0 \leq j \leq n\,, \\ \pi_n \cdot \dfrac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta}+k\right)}\,, & j \geq n+1\,, \end{cases} \tag{3.7}$$

where

$$\pi_n \;=\; \frac{E_{1,n}}{1+\left[A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)-1\right]\cdot E_{1,n}}\,. \tag{3.8}$$

4 Operational measures of performance

In order to understand and apply the Erlang-A model, one must first define its measures of performance, and then be able to calculate them. Moreover, since call centers can get very large (thousands of agents), the implementation of these calculations must be both fast and numerically stable.

4.1 Practical measures: Waiting Time

The most popular measure of operational (positive) performance is the fraction of served customers that have been waiting less than some given time, or formally $\mathrm{P}\{W \leq T, \mathrm{Sr}\}$, where W is the (random) waiting time in steady-state, $\{\mathrm{Sr}\}$ is the event "customer gets service" and T is a target time that is determined by Management/Marketing. For example, in a call center that caters to emergency calls, $T = 0$ (or T very small) would be appropriate. A common rule of thumb (without any theoretical backing, as far as we know) is the goal that at least 80% of the customers be served within 20 seconds; formally, $P\{W \leq 20, \mathrm{Sr}\} \geq 0.8$. To this, one sometimes adds E[W], or $\mathrm{E}[W|W>0]$, as some measure of an average (negative) experience for those who waited.

An important measure that is rarely used in practice is $P\{W > 0\}$, the fraction of customers who encounter a delay. This is a useful stable measure of congestion. Its importance stems from the fact that it identifies an organization's operational focus, in the following sense:

- $P\{W > 0\}$ close to 0 indicates a Quality-Driven operation, where the focus is on *service quality*;
- $P\{W > 0\}$ close to 1 indicates an Efficiency-Driven operation, where the focus is on *servers' efficiency* (in the sense of high servers' utilization);
- $P\{W > 0\}$ strictly between 0 and 1 (for example 0.5) indicates a careful *balance* between Quality and Efficiency, which we abbreviate to **QED = Quality & Efficiency Driven** operational regime.

The above three-regime dichotomy is rather delicate. For example, consider a system in which customers' average patience is close to the average service duration (for example, let both be equal to one minute), and assume that its offered load λ/μ is 100 Erlangs. Then, staffing of 100 servers would lead to the QED regime, with high levels of both service and efficiency that are balanced as follows: about 50% of the customers are served immediately upon arrival, the average wait is 2.3 seconds, 4% of the customers abandon due to their impatience, and servers' utilization levels are 96%. The QED regime still prevails at staffing levels between 95 and 105. With 90 servers, the system is efficiency-driven: 11% of the customers abandon, only 15% are served immediately, and utilization is over 99%. With 110 agents, it is quality-driven: abandonment is less than 1%, and 83% are served immediately.

In Section 6, we shall add details about the three operational regimes. This will be done in the context of describing regime-specific approximations for performance measures. However, there is much more to say about this important subject, and readers are referred to [19, 9] and Section 4 in the review [17] for details.

4.2 Practical measures: accounting for Abandonment

In a quality-driven service, $P\{W > 0\}$ seems the "right" measure of operational performance. We thus turn to alternative modes of operations and consider hereafter services in which $P\{W > 0\}$ is not close to vanishing.

As explained before, performance measures must take into account those customers who abandon. Indeed, if forced into choosing a *single* number as a proxy for operational performance, we recommend the probability to abandon $\mathbf{P\{Ab\}}$, the fraction of customers who explicitly declare that the service offered is not worth its wait. Some managers actually opt for a refinement that excludes those who abandon within a very short time, formally $\mathrm{P}\{W > \epsilon; \mathbf{Ab}\}$, for some small $\epsilon > 0$, for example $\epsilon = 3$ seconds. The justification is that those who abandon within 3 seconds can not be characterized as poorly served. There is also a practical rational that arises from physical limitations, specifically that such "immediate" abandonment could in fact be a malfunction or an inaccuracy of the measurement devices.

The single abandonment measure $\mathrm{P}\{\mathrm{Ab}\}$ can be in fact refined to account explicitly for those customers who were or were not well-served. Thus, we propose:

- $P\{W \le T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{\mathrm{Ab}\}$ - fraction of poorly-served.

A further refinement, that yields a four-dimensional service measure, could be:

- $P\{W \le T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{W > T; \mathrm{Sr}\}$ - fraction of served, with a potential for improvement (say, a higher priority on their next visit);
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served;
- $P\{W \le \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined - see the above for an elaboration.

Remark 4.1 4CallCenters [16] calculates, for a given target time, both $P\{W \le T; \mathrm{Sr}\}$, the fraction of customers who are served within target, and $P\{W \le \epsilon; \mathrm{Ab}\}$, those who abandon within target. To calculate the other two measures, it suffices to have $\mathrm{P}\{\mathrm{Ab}\}$, also calculated by 4CallCenters. Indeed,

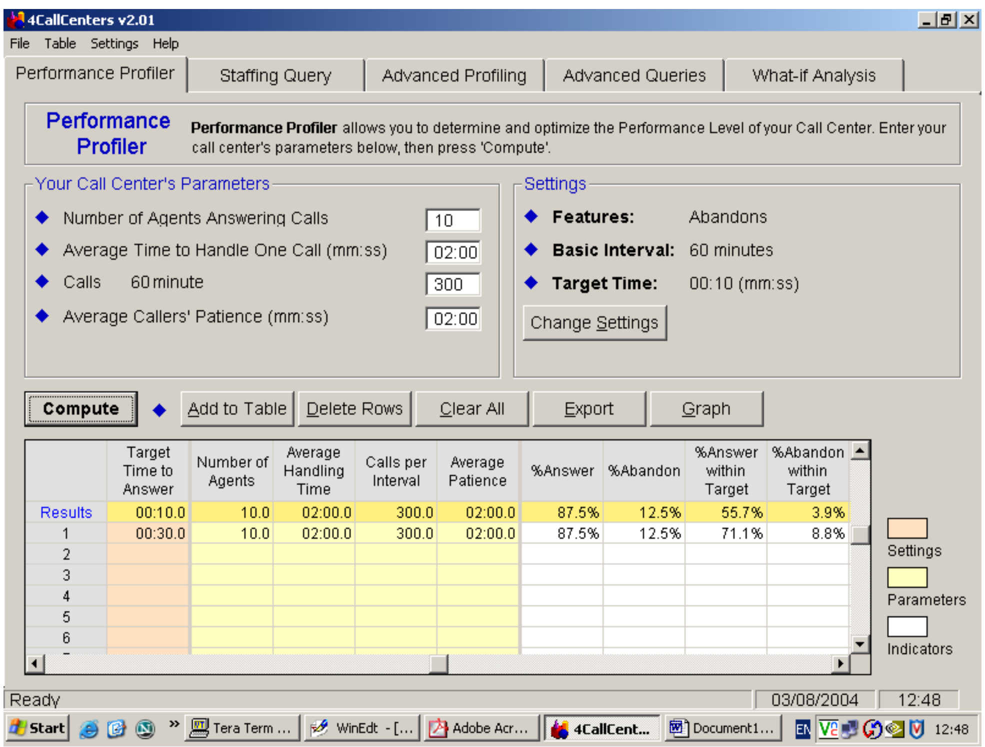
$$\begin{aligned}
\mathrm{P}\{W > T; \mathrm{Sr}\} &= 1 - \mathrm{P}\{\mathrm{Ab}\} - \mathrm{P}\{W \le T; \mathrm{Sr}\}\,, \\
\mathrm{P}\{W > \epsilon; \mathrm{Ab}\} &= \mathrm{P}\{\mathrm{Ab}\} - \mathrm{P}\{W \le \epsilon; \mathrm{Ab}\}\,.
\end{aligned}$$

Since a single target must be used ($T = \epsilon$ above), one must apply the program twice if different targets are required.

4.3 Calculations: the 4CallCenters software

Black-box Erlang-A calculations, as well as many other useful features, are provided by the free-to-use software 4CallCenters [16]. (This software is being regularly debugged and upgraded.) The calculation methods are described in Appendix B of [19]; they were developed in the Technion's M.Sc. thesis of the first author, Ofer Garnett.

Figure 5: **4Callcenters. Example of output.**



These calculations are in fact for measures of the form $\mathrm{E}[f(V,\tau)]$, for various functions f (Table 3 in [19]). For example,

$$\mathrm{E}[W] = \mathrm{E}\left[\min\{V,\tau\}\right] \ , \qquad \mathrm{P}\{\text{Abandon}\} = \mathrm{E}\left[1_{\{\tau < V\}}\right] \ .$$

Figure 5 displays a 4CallCenters output and demonstrates how to calculate the four-dimensional service measure, introduced in Subsection 4.2.

The values of the four Erlang-A parameters are displayed in the middle of the upper half of the screen: $n = 10$, $1/\mu = 2$ minutes, $\lambda = 300$ calls per hour, $1/\theta = 2$ minutes. Let $T = 30$

seconds and $\epsilon = 10$ seconds. Then one should perform computations twice: with *Target Time* 30 and 10 seconds. (Both computations appear in Figure 5.) We get:

- $P\{W \leq T; \text{Sr}\}$ - fraction of well-served is equal to 71.1%;
- $P\{W > T; \text{Sr}\}$ - fraction of served, with a potential for improvement, is 16.4% (87.5% − 71.1%);
- $P\{W > \epsilon; \text{Ab}\}$ - fraction of poorly-served is 8.6% (12.5% − 3.9%);
- $P\{W \leq \epsilon; \text{Ab}\}$ - fraction of those whose service-level is undetermined is 3.9%.

Note that the 4CallCenters output includes many more performance measures than those displayed in Figure 5: one could scroll the screen to values of agents' occupancy, average waiting time, average queue length, etc.

In Section 9 we describe several examples of the more advanced capabilities of 4CallCenters.

4.4 Delay probability P{W>0}

In this note, we content ourselves with few representative insightful calculations, based on conditioning and the incomplete gamma function introduced above. We start with the *delay probability* $\text{P}\{W > 0\}$, which represents the fraction of customers who are forced to actually wait for service. (The others are served immediately upon calling.) Recall that this measure identifies operational regimes of performance.

Following Palm [31], we show in the Appendix that the representations (3.5) and (3.7) immediately imply

$$\text{P}\{W > 0\} \;=\; \sum_{j=n}^{\infty} \pi_j \;=\; \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot E_{1,n}}{1 + \left(A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right) \cdot E_{1,n}}\,; \tag{4.1}$$

here, the first equality in (4.1) follows from PASTA.

4.5 Fraction abandoning P{Ab}

We proceed with calculating the probability to abandon, which represents the fraction abandoning. Define $\text{P}_j\{\text{Sr}\}$ to be the probability of ultimately getting served, for a customer that encounters all servers busy and j customers in queue, upon arrival (equivalently, $n + j$ in the system). "Competition among exponentials" now implies that

$$\text{P}_0\{\text{Sr}\} = \frac{n\mu}{n\mu + \theta}\ .$$

Then,

$$\mathrm{P}_1\{\mathrm{Sr}\} = \frac{n\mu+\theta}{n\mu+2\theta}\cdot \mathrm{P}_0\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+2\theta}\,,$$

where we conditioned on the first event, after an arrival that encounters all servers busy and a single customer in queue; this event is either a service completion (with probability $\frac{n\mu+\theta}{n\mu+2\theta}$) or an abandonment. More generally, via induction:

$$\mathrm{P}_j\{\mathrm{Sr}\} = \frac{n\mu+j\theta}{n\mu+(j+1)\theta}\cdot \mathrm{P}_{j-1}\{\mathrm{Sr}\} \;=\; \frac{n\mu}{n\mu+(j+1)\theta}\,, \qquad j\ge 1\,.$$

The probability to abandon service, given all servers busy and j customers in the queue upon arrival, finally equals

$$\mathrm{P}_j\{\mathrm{Ab}\} \;=\; 1-\mathrm{P}_j\{\mathrm{Sr}\} \;=\; \frac{(j+1)\theta}{n\mu+(j+1)\theta}\,, \qquad j\ge 0\,. \tag{4.2}$$

It follows that

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\sum_{j=n}^{\infty}\pi_j \mathrm{P}_{j-n}\{\mathrm{Ab}\}/\ \mathrm{P}\{W>0\} \;=\; \frac{1}{\rho A\left(\frac{n\mu}{\theta},\frac{\lambda}{\theta}\right)}+1-\frac{1}{\rho}\,. \tag{4.3}$$

The first equality in (4.3) is a consequence of PASTA, and the second is derived in the Appendix. The fraction abandoning, $\mathrm{P}\{\mathrm{Ab}\}$, is simply the product $\mathrm{P}[\mathrm{Ab}|W>0]\times \mathrm{P}\{W>0\}$.

4.6 Theoretical relations among P{Ab}, E(W), E(Q)

A remarkable property of Erlang-A, which in fact generalizes to other models with patience that is $exp(\theta)$, is the following linear relation between the fraction abandoning $\mathrm{P}\{\mathrm{Ab}\}$ and average wait $\mathrm{E}[W]$:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta\cdot \mathrm{E}[W]\,. \tag{4.4}$$

Proof: The proof is based on the balance equation

$$\theta\cdot \mathrm{E}[Q] \;=\; \lambda\cdot \mathrm{P}\{\mathrm{Ab}\}\,, \tag{4.5}$$

and on Little's formula

$$\mathrm{E}[Q] \;=\; \lambda\cdot \mathrm{E}[W]\,, \tag{4.6}$$

where Q is the steady-state queue length. The balance equation (4.5) is a steady-state equality between the rate that customers abandon the queue (left hand side) and the rate that abandoning customers (i.e. - customers who eventually abandon) enter the system. Substituting Little's formula (4.6) into (4.5) yields formula (4.4). ■

Observe that (4.4) is equivalent to

$$\mathrm{P}[\mathrm{Ab}|W>0] \;=\; \theta \cdot \mathrm{E}[W|W>0]\,. \tag{4.7}$$

Then, the average waiting time of delayed customers is computed via (4.3) and (4.7):

$$\mathrm{E}[W|W>0] \;=\; \frac{1}{\theta} \cdot \left[\frac{1}{\rho A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho} \right]. \tag{4.8}$$

The unconditional average wait $\mathrm{E}[W]$ equals the product of (4.1) with (4.8).

4.7 A general approach for computing operational performance measures

Expressions for additional performance measures of Erlang-A are derived in Riordan [32]. However, we recommend to use more general M/M/n+G formulae, as the main alternative to the 4CallCenters software. Indeed, M/M/n+G is a generalization of Erlang-A, in which patience times are generally distributed. A comprehensive list of M/M/n+G formulae, as well as guidance for their application, appears in Mandelbaum and Zeltyn [30]. The preparation of [30] was triggered by a request from a large U.S. bank. Consequently, this bank has been routinely applying Erlang-A in the workforce management of its 10,000 telephone agents, who handle close to 150 millions calls yearly. (In fact, Erlang-A replaced a simulation tool that had been used before.)

The handout [30] also explains how to adapt the M/M/n+G formulae to Erlang-A, in which patience is exponentially distributed:

$$G(x) = 1 - e^{-\theta x}\,, \qquad \theta > 0.$$

Specifically, see Sections 1,2 and 5 of [30].

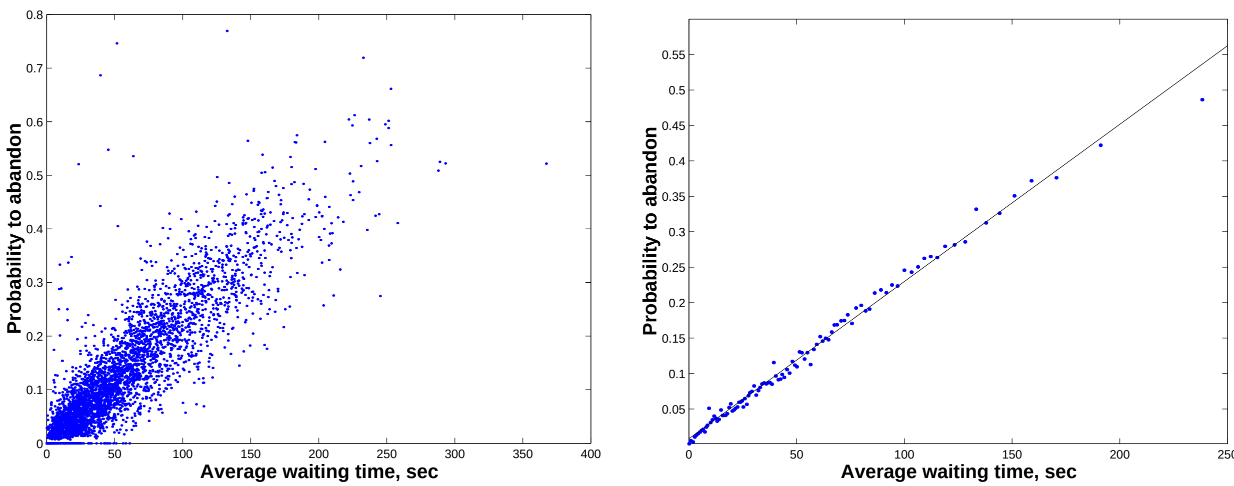
Finally, we explain how to calculate the four service measures from Section 4.2. The list on page 4 of [30] contains formulae for $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{P}\{W>T\}$ and $\mathrm{P}\{\mathrm{Ab}|W>T\}$. The product of the last two provides us with $\mathrm{P}\{W>T;\mathrm{Ab}\}$. The other three service measures are easily derived. For example,

$$\mathrm{P}\{W>T;\mathrm{Sr}\} \;=\; \mathrm{P}\{W>T\} - \mathrm{P}\{W>T;\mathrm{Ab}\}\,.$$

4.8 Empirical relations between E(W) and P{Ab}

Figure 6 illustrates the relation (4.4). It was plotted using yearly data of an Israeli bank call center [12]. (See also Brown et al. [11] for statistical analysis of this call center data.) First, P{Ab} and E[W] were computed for the 4158 hour intervals that constitute the year 1999. The left plot of Figure 6 presents the resulting "cloud" of points, as they scatter on the plane. For the right plot, we are using an aggregation procedure that is designed to emphasize dominating patterns. Specifically, the 4158 intervals were ordered according to their average waiting times, and adjacent groups of 40 points were aggregated (further averaged): this forms the 104 points of the second plot in Figure 6. (The last point of the aggregated plot is an average of only 38 hour intervals.)

Figure 6: **Probability to abandon vs. average waiting time**



We observe a convincing linear relation (line) between P{Ab} and E[W]. Based on (4.4) and Figure 6, the slope of this line is an estimate of average patience, which here equals 446 seconds. In Brown et al. [11] and Mandelbaum et al. [25] it is shown that some M/M/n+M assumptions do not prevail for the data [12]. Although arrivals are essentially Poisson, the service times are not exponential (in fact, they are very close to being lognormal). Patience times were shown to be non-exponential either. Yet, Erlang-A is proved useful for performance analysis, which we demonstrate further in Section 7.

It is therefore important to understand the circumstances under which one can practically use simple relations that, theoretically, apply perhaps only to models with exponential patience. A recent paper of Mandelbaum and Zeltyn [28] addresses this question for (4.4), demonstrating that the linear relation is practically rather robust. See also [29] where we demonstrate a similar linear relation on another data set.

5 Parameter estimation and prediction in a call center environment

In order to apply Erlang-A, it is necessary to input values for its four parameters: λ, μ, θ and n. Typical applications use estimates, which are based on historical ACD data, and here we briefly outline procedures for their estimation/prediction. For a more detailed exposition, including some subtleties that occur in practice, readers are referred to [11].

In call centers, Erlang-A is used, or should be used, to support solutions of the *staffing problem*, namely: how many agents should be answering calls during a specified time period. Typically, the goal is to provide a satisfactory service level (for example, fraction abandoning less than 3%), but sometimes one optimizes an economic measure - minimize cost or maximize revenues.

Arrivals: Arrivals of incoming calls are typically assumed Poisson, with time-varying arrival rates. The goal is to estimate/predict these arrival rates, over short time-intervals (15, 30 minutes or one hour), chosen so that the rates are approximately constant during an interval. Then the time-homogeneous model is applied separately over each such interval.

The goal can be achieved in two stages. First, time-series algorithms are used to predict *daily* volumes, taking into account trends and special days (eg. holidays, "Mondays", special sales). Second, one uses (non)parametric regression techniques for predicting the *fraction* of arrivals per time-interval, out of the daily-total. This fraction, combined with the daily total, yields actual arrival rates per each time-interval. (See Section 4 in [11] for a detailed treatment.)

Services: Service durations are assumed exponential. Average service times tend to be relatively stable from day to day. (However, they often change depending on the time-of-day! See [11].)

In practice, service consists of several phases, mainly talk time, wrap-up time (after-call work), and what is sometimes referred to as auxiliary time. An easier-to-grasp notion is thus "idle-time", namely the time that an agent is immediately accessible for service. It is thus also possible and recommended to estimate the average service time during a time interval by:

$$\frac{\text{Total Working Time} - \text{Total Idle Time}}{\text{Number of Served Customers}},$$

where Total Working Time is the product of the Number of Agents by the Interval Duration.

(See Adler et al. [1] for an application of this approach, in the context of product development.)

Number of agents: In performance analysis, the number of agents n is an Erlang-A input. In staffing decisions, n is typically an output. In both cases, n is in fact the needed number of FTE's (Full Time Equivalent positions), and hence it must be adjusted by the *rostered staff factor (RSF)*, or *shrinkage factor*, which accounts for absenteeism, unscheduled breaks etc. (See Cleveland and Mayben [13]) . For example, if 100 agents are required for answering calls during a shift, in fact more agents (105, 110, ...) should be assigned, depending on RSF.

In addition, it should be carefully checked whether staffing-level data from ACD reports is reliable. Specifically, is planned or actual number of agents reported? Does the ACD data take into account time periods (both scheduled and not) when agents are taking breaks?

Patience: (Im)patience time is assumed exponential, say $exp(\theta)$. One must then estimate the individual abandonment rate θ, or equivalently, the average patience $(1/\theta)$. A difficulty arises from the fact that direct observations are censored - indeed, one can only measure the patience of customers who abandon the system before their service began. For the customers receiving service, their waiting time in queue is only a lower bound for their patience. To this end, statistical tools for "un-censoring" data have been developed, mostly within a scientific discipline that is called "Survival Analysis". (See the Appendix in [40] for a brief survey, customized to estimating the distribution of (im)patience.) These tools typically require call-by-call statistics, which is beyond the available in most ACD reports. However, (im)patience that is $\exp(\theta)$ enables far simpler and data-modest "uncensoring" methods, two of which will now be described.

The first method is based on the relation (4.4) between the probability to abandon and average wait. The average wait in queue, $\mathrm{E}[W]$, and the fraction of customers abandoning, $\mathrm{P}\{\mathrm{Ab}\}$, are in fact standard ACD data outputs, thus, providing the means for estimating θ as follows:

$$\hat{\theta} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\%\text{Abandonment}}{\text{Average Wait}}\,.$$

A second more general approach is to calculate some performance measure (see Section 4) and compare the result to the value derived from ACD data. (This approach is applied in [11].) The goal is to calibrate the patience parameter until these estimates closely match. One advantage of this method is the flexibility in choosing the performance measure being matched, which might depend on the given ACD data. The choice of this measure would depend (in ways that are worthy of research) on what is available in the ACD data, how reliable the measurements are, how important is the performance measure, etc. Furthermore, this calibration represents a form of validation of the model's assumptions, and can compensate for discrepancies.

Remark 5.1 (On estimation and prediction) We distinguish between two closely related, but different, statistical tasks: estimation and prediction. Estimation concerns the use of existing (historical) data to make inferences about the parameter values of a statistical model. Prediction concerns the use of the estimated parameters to forecast the behavior of a sample outside of the original data set (used to make the estimate). Predictions are "noisier" than estimates, because, in addition to uncertainty concerning the estimated parameters, they contain additional sources of potential errors. For example, future arrival rates to a call center of a bank could depend on stock trends or on unknown impacts of advertising. Therefore, in the case of arrivals, we have a clear-cut prediction task. The cases of service-times and customers' patience can typically be treated as estimation problems.

6 Approximations

Although exact formulae for the Erlang-A system are available and can be incorporated in software (see Sections 3 and 4), they are too complicated for providing guidelines and insights to call center researchers and managers. Consequently, some useful and insightful approximations have been developed, which we now review.

It has been found useful to distinguish three operational regimes, as in Garnett et al. [19] and Zeltyn and Mandelbaum [29]. Each regime represents a different philosophy for operating a call center. One regime is Efficiency-Driven (ED), another is service Quality-Driven (QD), and the third one rationalizes efficiency and quality, namely it is Quality and Efficiency-Driven (QED).

We are interested mainly in not-too-small call centers. Hence, we think of the service and abandonment rates, μ and θ, as fixed, and the arrival rate λ is not too small (formally, it increases indefinitely).

Actually, the regimes are determined by the *offered load* parameter R, which is defined as $R = \dfrac{\lambda}{\mu}$; R represents the amount of work, measured in time-units of service, that arrives to the system per unit of time. ($R = \lambda \cdot \dfrac{1}{\mu}$ is thus a more telling representation.) R is also the staffing level n that would be prescribed by the deterministic approach (see the end of Section 2), which ignores stochastic variability. An emphasis of efficiency (service quality) would conceivably lead to $n < R$ ($n > R$); the deviation of n from R then increases with the intensity of the emphasis. We now proceed with a formal description of the three operational regimes.

QED (Quality and Efficiency-Driven) regime.

$$n = R + \beta\sqrt{R} + o(\sqrt{R}), \qquad -\infty < \beta < \infty, \tag{6.1}$$

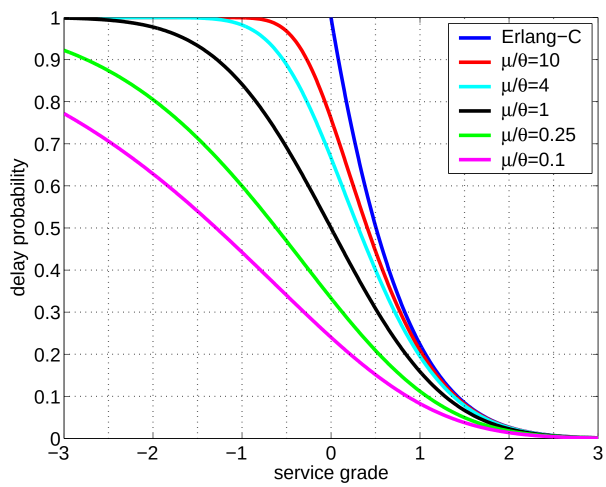
where β is a service-grade parameter – the larger it is, the better is the service-level. The staffing regime, described by (6.1), is governed by the so-called *Square Root Rule.* This rule was already described by Erlang [14], as early as 1924. He reported that it had been in use at the Copenhagen Telephone Company since 1913. A formal QED analysis for the Erlang-C queue appeared only in 1981, in the seminal paper of Halfin and Whitt [21]. (The service grade β must be positive in this case.) Garnett et al. [19] explored Erlang-A in the QED regime and Zeltyn and Mandelbaum [29] treated the M/M/n+G queue with a general patience distribution. (With abandonment, β can be also 0 or negative.)

In the QED regime, the delay probability $P\{W > 0\}$ converges to a constant that is a function of only the *service grade* β and the ratio μ/θ (or $\frac{1/\theta}{1/\mu}$, which is average patience that is measured in units of average service time). The probability to abandon and average wait vanish, as $\lambda, n \uparrow \infty$, both at rate $\dfrac{1}{\sqrt{n}}$. Formulae for different performance measures can be found in [19] or [29]. It is significant and useful to mention that QED approximations are often accurate over a wide range of parameter values, from the very large call centers (1000's of agents) to the moderate-size ones (10's of agents).

Figure 7 illustrates the dependence between β and $P\{W > 0\}$, for varying values of the ratio μ/θ. In addition, we plotted the curve for the Erlang-C queue, which is meaningful for positive β only. Note that for large values of μ/θ (very patient customers) the Erlang-A curves get close to the Erlang-C curve.

Remark 6.1 When $\beta = 0$ in (6.1), the staffing level corresponds to the simple rule that does not take into account stochastic considerations: simply assign the number of agents to be the *offered load* λ/μ. In Erlang-C, this "naive" approach would lead to system instability. However, in Erlang-A (which is a much better fit to the reality of call centers than Erlang-C) one would get a reasonable-to-good performance level. For example, if the service rate μ is equal to the individual abandonment rate θ, and $\beta = 0$, 50% of the customers would get service immediately upon arrival. (Check it in Figure 7. Note that for Erlang-C, 50% delay probability corresponds to $\beta = 0.5$.) This suggests why some call centers, that are managed using simplified deterministic models, actually perform at reasonable service levels. (One obtains the "right answer" from the "wrong reasons".)

Figure 7: **Asymptotic relations between the delay probability $\mathbf{P}\{W > 0\}$ and service grade $\boldsymbol{\beta}$**



The QED regime enables one to combine high levels of efficiency (agents' utilization close to 100%) and service quality (agents' accessibility) . The scatterplots in Figures 8 and 9 illustrate this point. The plots display data from ACD reports of two call centers: Italian and American, collected in half-hour intervals during a single working day. The service grade β is calculated via

$$\beta \;=\; \frac{n-R}{\sqrt{R}}\,.$$

We observe moderate-to-small values of abandonment for the service grade $-1 \leq \beta \leq 2$. The values of average wait in Figure 9 are also moderate, except the highly-loaded hours in the American call center.

ED (Efficiency-Driven) regime.

$$n = R\cdot(1-\gamma) + o(\sqrt{R})\,, \qquad \gamma > 0\,. \tag{6.2}$$

In this case, virtually all customers wait, the probability to abandon converges to γ and average wait is close to γ/θ. (See Mandelbaum and Zeltyn [30] for additional performance measures.)

The ED regime could be used if efficiency considerations are of main significance. Indeed, it has gained importance in recent research (see, for example, the papers of Whitt [34, 35]), motivated by the observation that ED could yield performance that is acceptable for many call centers, for example those operating in not-for-profit environments.

Figure 8: **Service grade for call centers - correlation with abandonment**

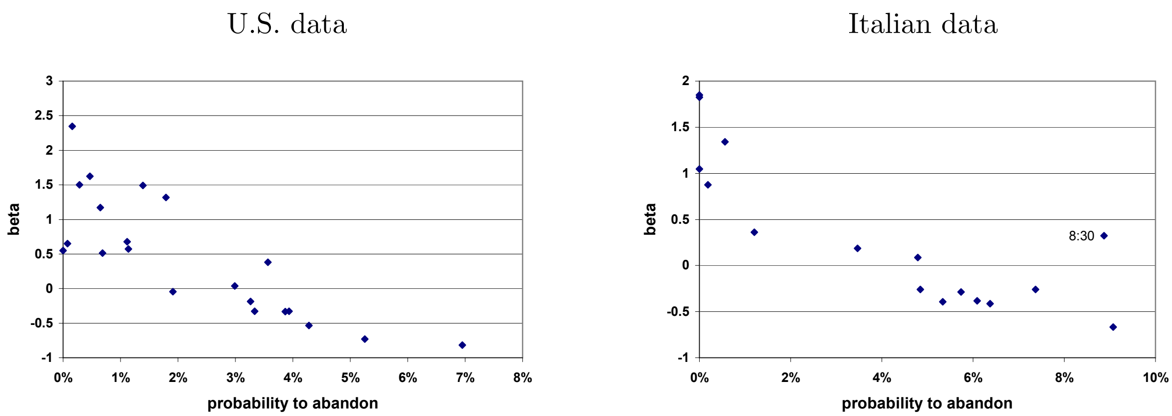
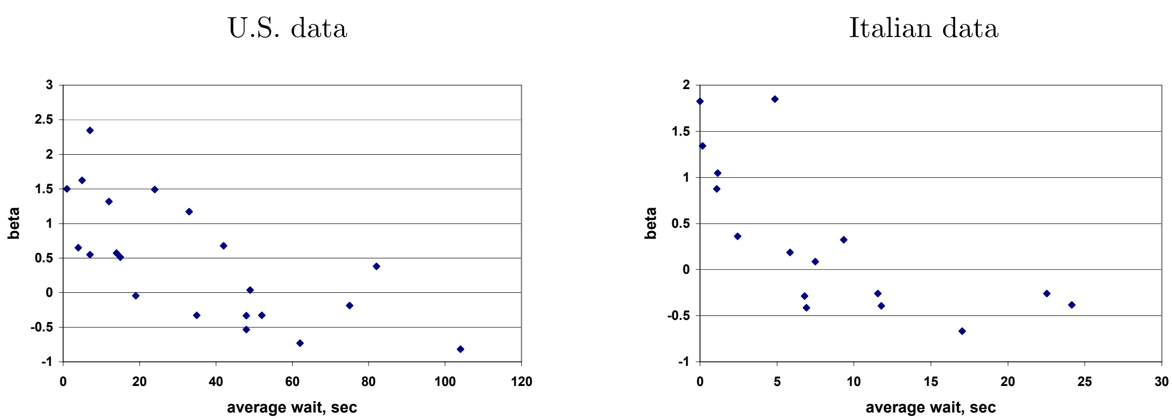


Figure 9: **Service grade for call centers - correlation with average wait**



QD (Quality-Driven) regime.

$$n = R \cdot (1+\gamma) + o(\sqrt{R}) \,, \qquad \gamma > 0 \,. \tag{6.3}$$

The staffing regime (6.3) should be implemented if quality considerations far dominate efficiency considerations (e.g. high-valued customers or emergency phones). Major performance measures (delay probability, fraction abandoning, average wait) vanish here at an exponential rate of n.

Remark 6.2 Above, we have considered steady-state performance measures. Process-limit results for the number-in-system process $L = \{L(t),\ t \geq 0\}$ are available, for the QED and ED regimes, in Garnett et al. [19] and Whitt [34], respectively.

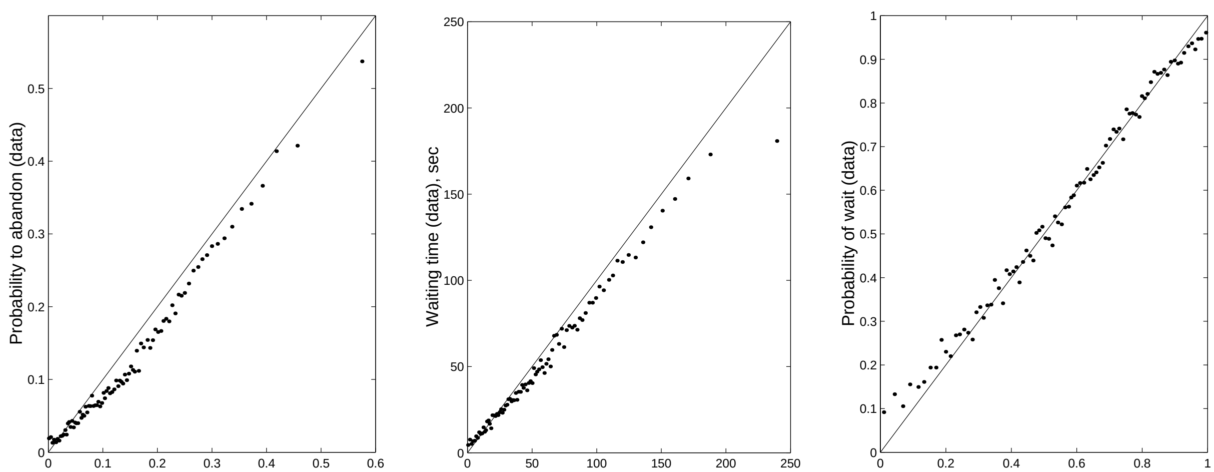
7 Applications in call centers

7.1 Erlang-A performance measures: comparison against real data

When modelling a real call center, the Erlang-A model is, most likely, a significant simplification of it. We now validate Erlang-A against the hourly data of the Israeli call center, already used for the examples in Section 4. Such validation is a prerequisite for any serious Erlang-A application.

Three performance measures are considered: probability to abandon, average waiting time and probability of wait. Their values are calculated for the hourly intervals using exact Erlang-A formulae. Then the results are aggregated along the same method employed in Figure 6. The resulting 86 points are compared against the line $y = x$: the better the fit the better Erlang-A describes reality.

Figure 10: **Erlang-A formulas vs. data averages [11]**



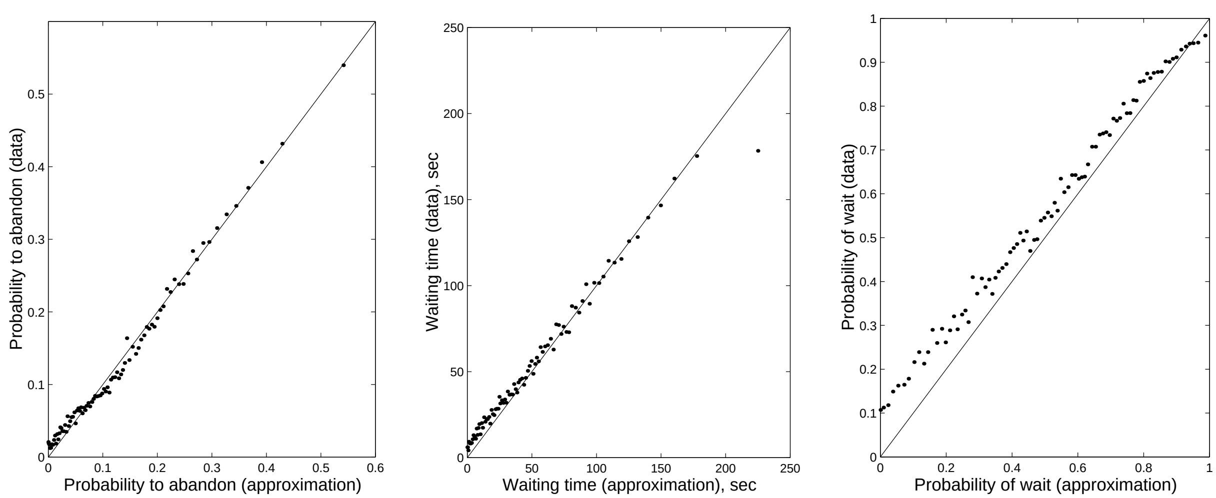
Computation of the Erlang-A parameters. Parameters λ and μ are calculated for every hourly interval. We also calculate each hour's average number of agents n. Because the resulting n's need not be integers, we apply a continuous extrapolation of the Erlang-A formulae, obtained from relationships developed in [31].

For θ, we use formula (4.4), valid for exponential patience, in order to compute 17 hourly estimates of $1/\theta = E(\tau)$ (for the 17 one-hour intervals 7am-8am, 8am-9am, ..., 11pm-12pm). The values for $E(\tau)$ ranged from 5.1 min (8am-9am) to 8.6 min (11pm-12pm). We judged this to be better than estimating θ individually for each hour (which would be very unreliable) or, at the other extreme, using a single value for all intervals (which would ignore possible variations in customers' patience over the time of day; see [40]).

The results are displayed in Figure 10. The figure's two left-hand graphs exhibit a relatively small yet consistent overestimation with respect to empirical values, for moderately and highly loaded hours. The right-hand graph shows a very good fit everywhere, except for very lightly and very heavily loaded hours. The underestimation for small values of $P\{W > 0\}$ can be probably attributed to violations of work conservation (idle agents do not always answer a call immediately). Summarizing, it seems that these Erlang-A estimates can be used as close *upper bounds* for the main performance characteristics of our call center.

7.2 Erlang-A approximations: comparison against real data

Figure 11: **Erlang-A approximations vs. data averages [11]**



In Section 6 we discussed approximations of various performance measures for the Erlang-A (M/M/n+M) model. Such approximations require significantly less computational effort than exact Erlang-A formulae. Figure 11, based on the same data as Figure 10, demonstrates a good fit between data averages and the approximations.

In fact, the fit for the probability of abandonment and average waiting time is even superior to the one in Figure 10 (the approximations provide somewhat larger values than the exact formulae). This phenomenon suggests two interrelated research questions of interest: explaining the overestimation in Figure 10 and better understanding the relationship between Erlang-A formulae and their approximations.

The empirical fit of both the simple Erlang-A model and its approximation turn out to be very (perhaps surprisingly) accurate. Thus, for the call center in consideration – and those like it – use of Erlang-A (exact or approximations) for workforce management could and should improve operational performance.

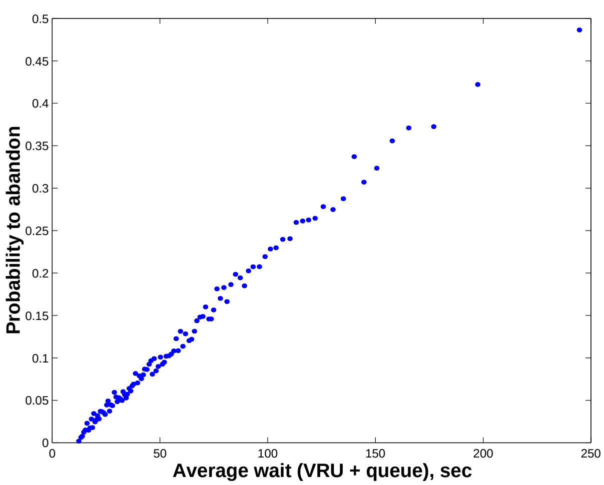
8 Human behavior

The Erlang-A model assumes exponential patience. In practice, this assumption is not necessarily valid. Below we present several examples of customers' behavior in tele-queues.

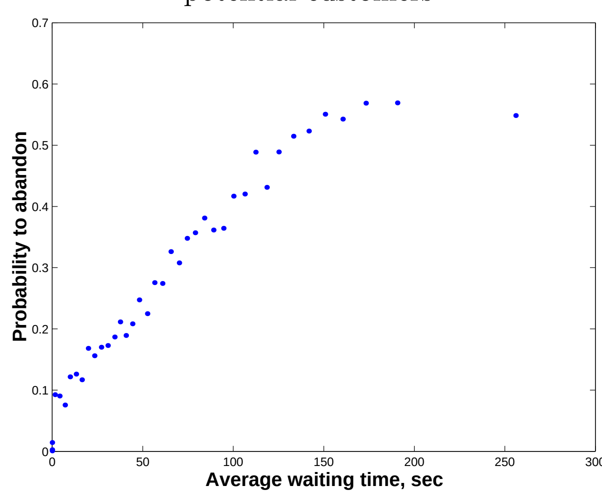
8.1 Balking and delayed impatience

Figure 12: **Probability to abandon vs. average waiting time**

VRU-time included in wait



potential customers



Two specific examples of a linear relation between the probability to abandon and average wait are displayed in Figure 12. Both are based on the same Israeli call center data, referred to above. The first plot takes into account all customers. It differs from Figure 6 by its definition of waiting time: here it includes also the time spent by customers in the VRU (Voice Response Unit). The second plot is for a specific type of customers: potential customers asking for information on available services (approximately 15% of phone calls over the year). Both plots aggregate data along the guidelines of the second graph in Figure 6.

The points on the first plot are approximated by a straight line that intersects the x-axis around 10 seconds (average time spent in the VRU). A similar pattern of positive x-intercept can arise if, for whatever reason, customers do not abandon during the first period of their wait (a so-called *delayed* distribution of patience).

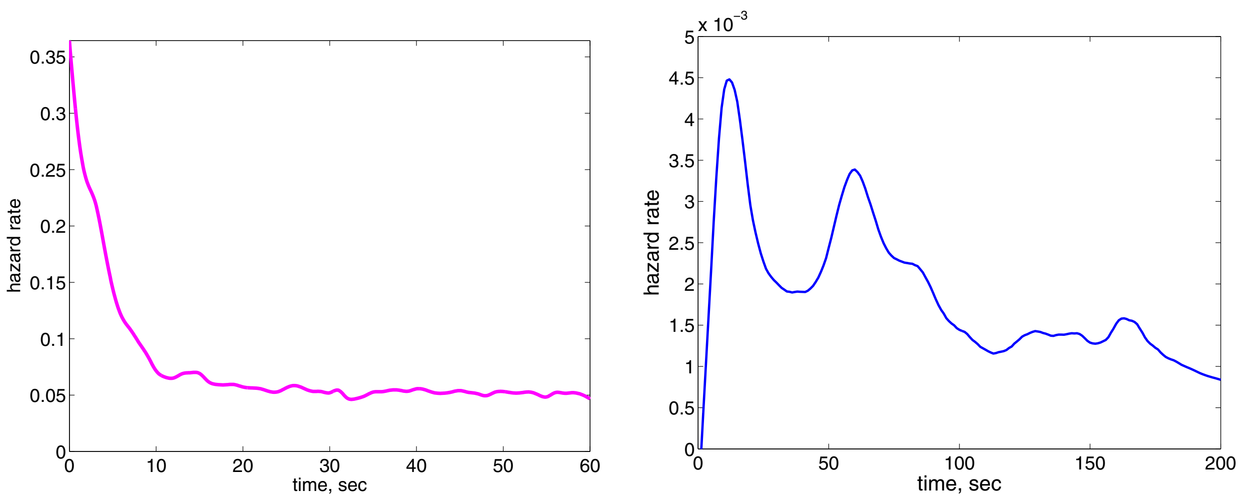
The pattern of the second plot is close to a straight line with a positive intercept of the y axis, approximately at 0.1 (except for a couple of points near the origin and the point with the largest wait). Theoretically, a similar relation arises in the case of exponential abandonment with *balking*: customers' patience is equal to zero with probability p (namely, immediate abandonment if wait is encountered) and it is exponential with probability $(1 - p)$. The relation between the theoretical balking model and the second empirical plot in Figure 12 can be partially clarified using [25]. It was shown there that potential customers are less patient than customers overall, and they often abandon during their first seconds of wait.

8.2 Examples of the patience-time hazard rate

Figure 13: **Bank data: hazard rates of patience times**

U.S. bank

Israeli bank



In Figure 13 we display estimates of the hazard rates of customers' patience for two banks: a large U.S. bank and a small Israeli one. In the two cases we observe different, but clearly non-exponential patterns. (Recall that the hazard rate of an exponential random variable is a constant.) The U.S. customers are very impatient at the beginning of their wait, but their patience stabilizes after approximately 10 seconds. In contrast, Israeli customers have two clear peaks of abandonment: approximately at 15 and at 60 seconds.

Both patterns can be related to human behavior. The U.S. data is, in fact, from a network of 3 call centers, with interflows among them. Roughly speaking, the interflow protocol prescribes, first, waiting of up to 10 seconds at the original call center to where the call is being made. After 10 seconds of wait, the system either identifies a free qualified agent in one of the other call centers, or it adds the call to their relevant queues. Hereafter, the call waits simultaneously in several queues, until being assigned to the first turning-free qualified agent. it follows that the "10 seconds mark" arises as a natural change-point of customers' behavior, but it is unclear to us why the change is manifested by a plateau of the impatience.

As for the Israeli data, it turns out that the two surges of abandonment take place immediately after two recorded messages to which customers are exposed: the first one when they enter the queue and the second after approximately 1 minute. Interestingly, such messages stimulate abandonment. This could be opposite of what the messages are typically designed to achieve, namely reduce abandonment by reducing anxiety via information. On the other hand, these messages provide customers with the opportunity of making an "informed choice", deciding to wait or to abandon now, instead of possibly abandoning later in frustration.

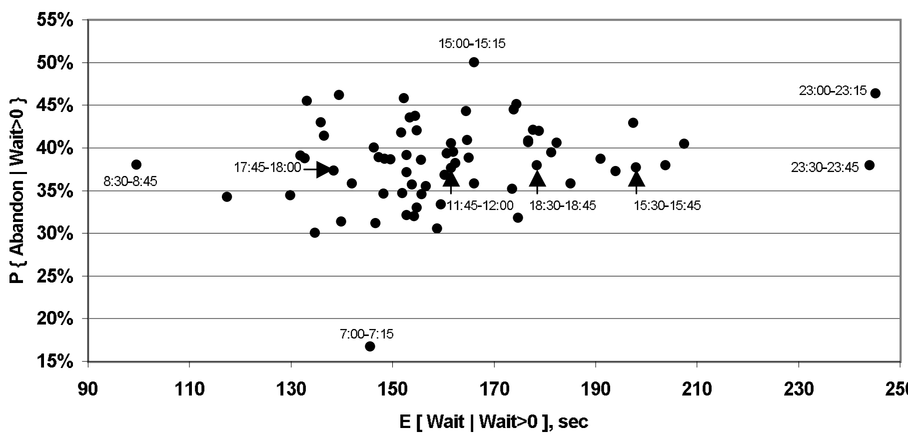
Therefore, at least in some applications, customers' patience times are non-exponential and applicability of the Erlang-A formulae to such systems should be validated. (Recall Section 7.)

8.3 Adaptive behavior of impatient customers

The prevalent assumption in traditional queueing theory is that patience-time ($\exp(\theta)$, in our case) is "assigned" to individual customers independently of system state. In particular, patience is unaltered by possible changes in congestion. Such models, however, can not accommodate the scatterplot in Figure 14, that exhibits remarkable patience-adaptivity: the more the congestion the higher is the patience to cope with it.

To elaborate, the data is again from the Israeli bank call center. We are scatterplotting abandonment fraction against average delay, for delayed customers (positive queueing time) who seek technical Internet-support. It is seen that average delay during 8:30-8:45am, 17:45-18:00, 18:30-18:45 and 23:30-23:45pm is about 100, 140, 180 and 240 seconds respectively. Nonetheless, the fraction of abandoning customers (among those delayed) is remarkably stable at 38%, for all periods. This stands in striking contrast to traditional queueing models, where patience is assumed unrelated to system performance: such models would predict a strict increase of the abandonment fraction with the waiting time, as in Figure 6. The behavior indicated in Figure 14 thus suggests that customers do adapt their patience to system performance. See the papers of Mandelbaum, Shimkin and Zohar [27, 33, 40] for an analysis of such adaptive behavior.

Figure 14: **Adaptive behavior of Internet-support customers**



Summarizing, the waiting experience of a customer can be broken down into the following five components:

- Time that a customer *expects* to wait;
- Time that a customer is *willing* to wait (R, patience or need);
- Time that a customer *must* wait (V, offered wait);
- Time that a customer *actually* waits (W, the minimum between R and V);
- Time that a customer *perceives* waiting.

For a customer that is "experienced", it is plausible that the expected wait is close to the offered wait and that the perceived wait is similar to the actual wait. Then, one is left with the triplet (R, V, W) from our basic model, introduced in Section 3: R is a model input and (V, W) are its output.

8.4 Patience index

In search for a better understanding of customers' (im)patience, we have found a *relative* definition to be of use. Specifically, we define the *patience index* to be

$$\begin{aligned}\text{Theoretical Patience Index} &\triangleq \frac{\text{time willing to wait}}{\text{time required to wait}} \\ &= \frac{\text{average patience}}{\text{average offered wait}} = \frac{\mathrm{E}[\tau]}{\mathrm{E}[V]}\,.\end{aligned}$$

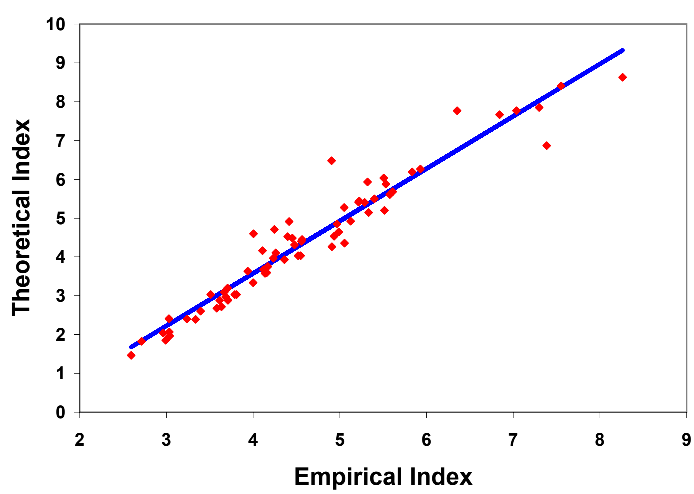
While this patience index makes sense intuitively, its calculation requires the application of survival analysis techniques to call-by-call data. Such data may not be available in certain circumstances. Therefore, we wish to introduce an *empirical* index which will serve as an auxiliary measure for the patience index, and the following has been found useful:

$$\text{Empirical Patience Index} \triangleq \frac{\%\ \text{served}}{\%\ \text{abandoned}}\,.$$

The empirical index is easily calculable since both the numbers of served and of abandoned calls are very easy to obtain from call-center reports.

Figure 15 demonstrates how well the empirical patience index estimates the theoretical patience index for the Israeli bank data [11]. (Aggregated data of 68 quarter hours between 7am and midnight is used.)

Figure 15: **Patience index – empirical vs. theoretical [12]**



Under certain circumstances, one can explain the closeness of the theoretical and empirical indices [11]. However, these explanations are incomplete and hence leave open this research direction.

9 Advanced features of the 4CallCenters software

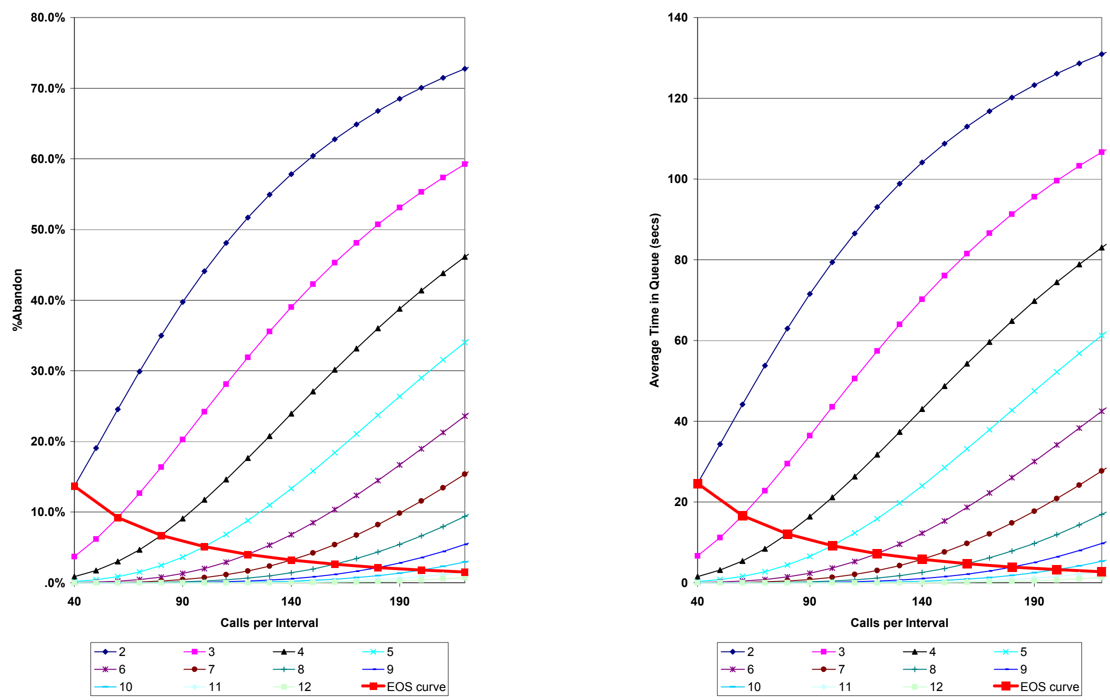
The 4CallCenters software [16] provides a valuable tool for implementing Erlang-A calculations. Its basic feature is “Performance Profiler” that enables calculation of all the useful performance measures, given the four Erlang-A parameters as input. In addition, 4CallCenters allows many advanced options: staffing queries, graphs, export and import of data and more.

Here we demonstrate, as an example, two advanced capabilities of 4CallCenters.

Figure 16: **4CallCenters. Advanced profiling.**

Probability to abandon

Average wait

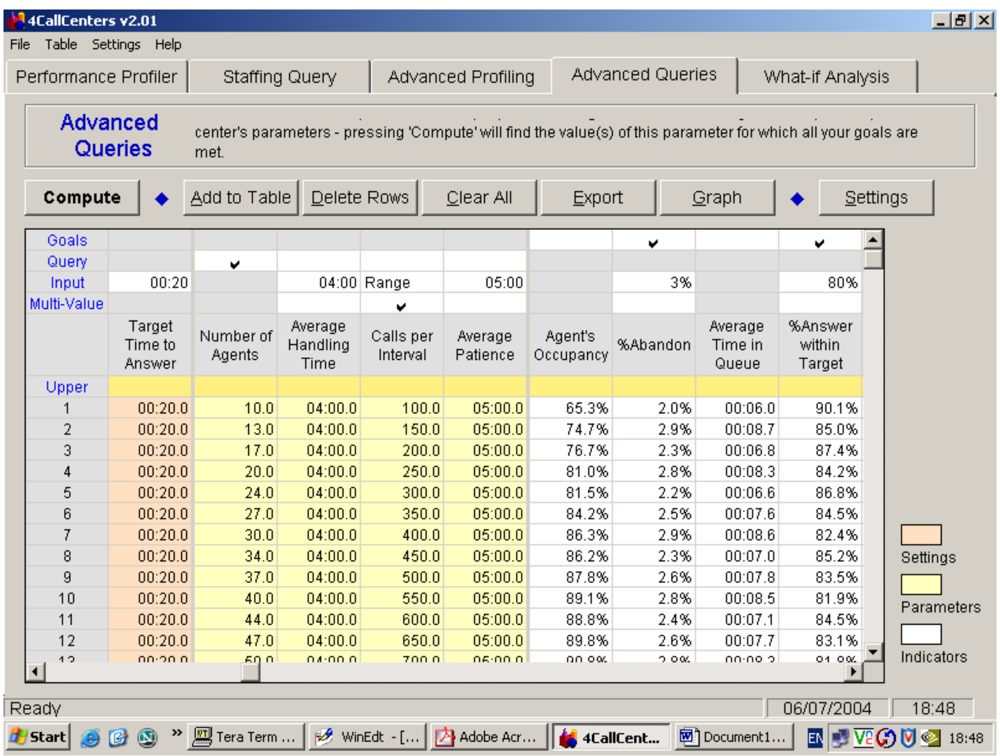


Example 1: Advanced profiling. One can vary any input parameters of the Erlang-A queue and display the corresponding model output (performance measures) either in a table or graphically. For example, let the average service time equal 2 minutes and average patience 3 minutes. Let the arrival rate vary from 40 to 230 calls per hour, in steps of 10, and the number of agents from 2 to 12. Then one can immediately produce a table that contains values of different performance measures for all combinations of the two input parameters. Based on this table, 4CallCenters can transform the data into EXCEL plots, such as those depicted in Figure 16.

Specifically, Figure 16 shows the dependence of the probability to abandon and average wait on different number of agents. Note that the two plots look identical: the reason is relation (4.4). In addition, the red curves on both plots in Figure 16 (which were added to the 4CallCenters plots) illustrate Economies of Scale (EOS): while offered load per server remains constant along this curve $\left(\frac{\lambda}{n\mu} = \frac{2}{3}\right)$, performance significantly improves as the number of agents increases. For example, the probability to abandon is equal to 13.7% for $n = 2$, 5.1% for $n = 5$ and 1.5% for $n = 12$. Finally, note that both P{Ab} and E[W] actually vanish as n gets large.

Example 2. Advanced staffing queries. 4CallCenters enables staffing queries with several performance goals. For example, assume that the average service time is equal to 4 minutes, and average patience is 5 minutes. Our goal is to calculate appropriate staffing levels for arrival-rate values that vary from 100 to 1200, in steps of 50. The performance targets are:

Figure 17: **4Callcenters. Advanced staffing queries.**



	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Agent's Occupancy	%Abandon	Average Time in Queue	%Answer within Target
Input	00:20		04:00	Range	05:00		3%		80%
1	00:20.0	10.0	04:00.0	100.0	05:00.0	65.3%	2.0%	00:06.0	90.1%
2	00:20.0	13.0	04:00.0	150.0	05:00.0	74.7%	2.9%	00:08.7	85.0%
3	00:20.0	17.0	04:00.0	200.0	05:00.0	76.7%	2.3%	00:06.8	87.4%
4	00:20.0	20.0	04:00.0	250.0	05:00.0	81.0%	2.8%	00:08.3	84.2%
5	00:20.0	24.0	04:00.0	300.0	05:00.0	81.5%	2.2%	00:06.6	86.8%
6	00:20.0	27.0	04:00.0	350.0	05:00.0	84.2%	2.5%	00:07.6	84.5%
7	00:20.0	30.0	04:00.0	400.0	05:00.0	86.3%	2.9%	00:08.6	82.4%
8	00:20.0	34.0	04:00.0	450.0	05:00.0	86.2%	2.3%	00:07.0	85.2%
9	00:20.0	37.0	04:00.0	500.0	05:00.0	87.8%	2.6%	00:07.8	83.5%
10	00:20.0	40.0	04:00.0	550.0	05:00.0	89.1%	2.8%	00:08.5	81.9%
11	00:20.0	44.0	04:00.0	600.0	05:00.0	88.8%	2.4%	00:07.1	84.5%
12	00:20.0	47.0	04:00.0	650.0	05:00.0	89.8%	2.6%	00:07.7	83.1%
13	00:20.0	50.0	04:00.0	700.0	05:00.0	[illegible]	[illegible]	[illegible]	[illegible]

- Probability to abandon less than 3%;
- 80% of customers served within 20 seconds.

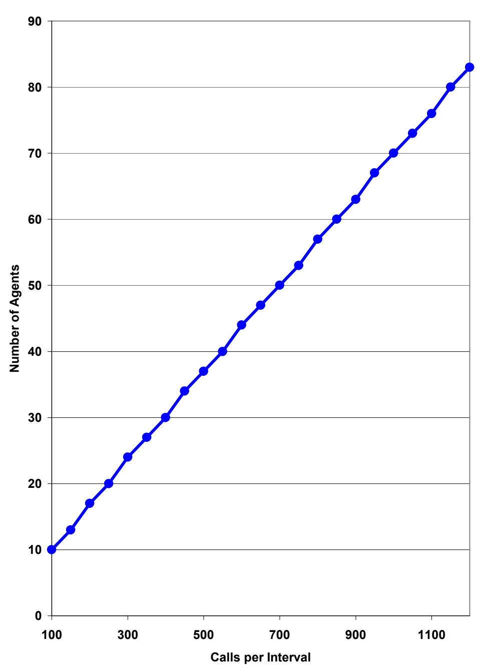
Figure 17 presents the screen output of 4CallCenters which, again, the software easily transforms into graphs.

The first plot of Figure 18 displays the minimal staffing level that adheres to both goals. The EOS phenomenon is observed here as well: 10 agents are needed for 100 calls per hour but only 83 (rather than $10 \cdot 12 = 120$) for 1200 calls per hour. (Despite its look, the curve in the first plot is not a straight line but it is rather concave.) The second plot displays the values of the two target performance measures. (This plot, unlike the first one, is not an immediate output graph of 4CallCenters but rather an edited version of it.)

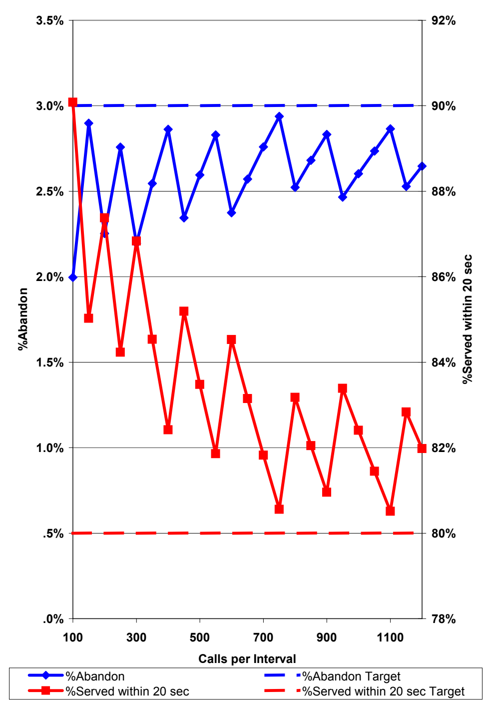
Remark 9.1 Since the number of agents must be an integer, we observe performance "zigzags" in the right plot of Figure 18.

Figure 18: **4CallCenters. Dynamics of staffing level and performance.**

Recommended staffing level



Target performance measures



10 Some open research topics

10.1 Dimensioning the Erlang-A queue

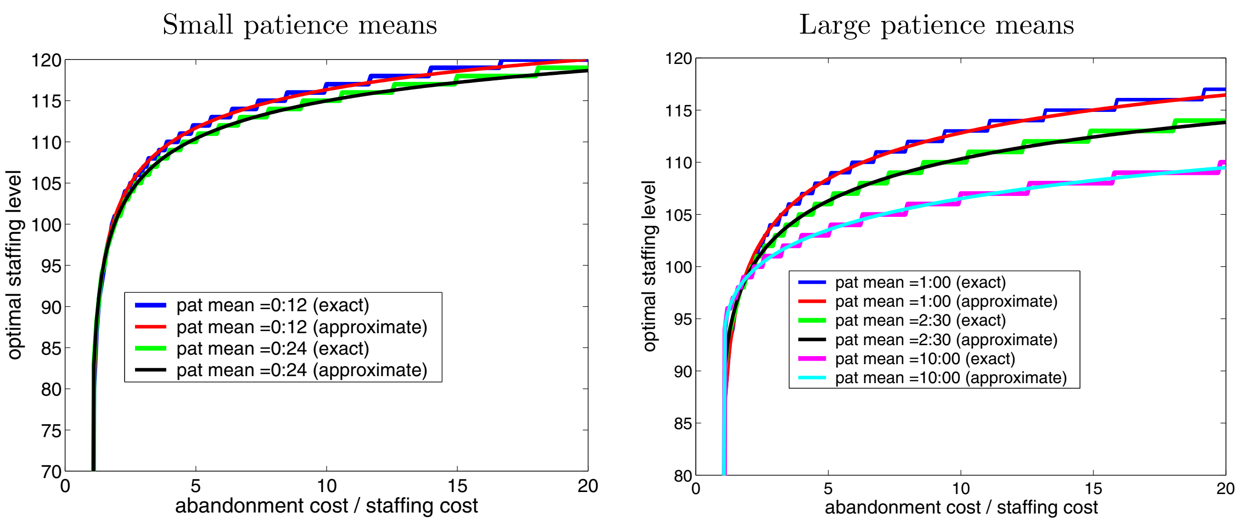
One can search for an economically optimal staffing level, given the trade-off between staffing cost, cost of customers' waiting and cost of abandonment. Borst et al. [9] referred to this problem as *dimensioning* and solved it for the Erlang-C queue (no abandonment): if the staffing cost and the cost of waiting are comparable, the optimal staffing was found to take place in the QED regime, described in Section 6 (with positive β in formula (6.1)). Ongoing research by Borst et al. [10] is dedicated to the same question for the Erlang-A queue. For example, let the average operational cost (per unit of time) be equal to

$$U(n, \lambda) \;=\; c \cdot n + \lambda a \cdot \mathrm{P}\{\mathrm{Ab}\}\,,$$

where c is the staffing cost, and a is the abandonment cost. Our goal is to find the staffing level n^* that minimizes cost. (Note that this is in fact mathematically equivalent to maximizing revenues.)

Figure 19: **Cost optimization. Approximation vs. exact optimum.**

Arrival rate $\lambda = 100$, service rate $\mu = 1$.



Define the abandonment/staffing cost ratio by $r \triangleq a/c$, and let $s \triangleq \sqrt{\mu/\theta}$. Assume that $a > c/\mu$. (Otherwise, the asymptotic optimal policy is $n^* = 0$: not to provide service at all.) Then we suggest that the asymptotic optimal staffing level is equal to

$$n^* \;=\; \left[R + y^*(r; s) \cdot \sqrt{R} \right] \,, \tag{10.1}$$

where the square brackets in (10.1) denote the nearest integer value, the function $y^*(\cdot)$ is defined by

$$y^*(r;s) \stackrel{\Delta}{=} \arg \min_{-\infty<y<\infty} \left\{ c \cdot y + a \cdot \theta s \cdot \left[1 + \frac{h(ys)}{sh(-y)}\right]^{-1} \cdot [h(ys) - ys] \right\} ,$$

and $h(\cdot) = \phi(\cdot)/(1 - \Phi(\cdot))$ is the hazard rate of the standard normal distribution ($\phi(\cdot)$ is its density function and $\Phi(\cdot)$ is the cumulative distribution function).

As in [9], Figure 19 compares the rule (10.1) with the exact optimal staffing values. We consider five exponential patience distributions with different means and perform comparisons by varying the value of the ratio r. A perfect fit is observed for *all* the special cases!

Numerical experiments for other cost optimization problems (e.g. with waiting cost, instead of abandonment cost) also demonstrate a very close correspondence between exact values and the corresponding analogs of (10.1). Hence the goal is to develop a theoretical framework, parallel to [9], that will support our experimental research. In addition, we are working on a *constraint satisfaction* version where one chooses the least number of agents that adheres to a given constraint on the waiting and/or the abandonment cost. This latter formulation is in fact closer to the way that managers perceive their staffing problems in practice.

10.2 Uncertainty in parameter values

In applications, one does not know the exact values of the four Erlang-A parameters. Therefore, it is essential to study the sensitivity of performance measures to their values. In a recent paper [37], Whitt calculates *elasticities* in Erlang-A, which measure the percentage change of a performance measure caused by a small percentage change in a parameter. Both exact numerical algorithms and several types of approximations are used. It turns out that Erlang-A performance is quite sensitive to small changes in the arrival rate, service rate, or number of agents, but relatively insensitive to small changes in the abandonment rate.

In staffing planning, of which the number of agents is the output, it is reasonable to assume knowledge of the service rate. Thus, the problem of uncertainty in the arrival rate surfaces as the most significant.

In Brown et al. [11] it was shown that the Poisson arrival rate in the Israeli call center actually varied from day to day and, hence, its prediction raised statistical and practical challenges. This motivates the study of queueing models in which the Poisson arrival rate Λ (the arrival-rate function) is a random variable (random process).

If $E(\Lambda) \to \infty$ and its standard deviation is of the order $\sqrt{E(\Lambda)}$, we expect that the QED operational regime and the square-root staffing rule will play a role that is similar to the one with known (deterministic) arrival rate; the offered load R in (6.1) will be replaced by the average offered load $E(\Lambda)/\mu$, and uncertainty will manifest itself through a different value of the service grade β. However, if $\sigma(\Lambda)$ is of the order $E(\Lambda)$, the "cruder" ED regime seems to be the most appropriate, as in Whitt [36], and Bassamboo, Harrison and Zeevi [7].

10.3 Additional topics

The call center environment is rich with additional features that are worthy of modelling. Some have been investigated, at least partially, as can be seen from [26]. Among these, we mention:

- Skill-based routing; see Subsection 5.1 in [17], as well as Armony and Mandelbaum [4], Gurvich [20], Armony et al. [3], Wallace and Whitt [38];
- Retrials; see Subsection 6.3.3 in [17], as well as Aguir et al. [2];
- Networked operations; see Subsection 5.3 in [17];
- Time-inhomogeneous arrival rate; see Subsection 4.3 in [17], as well as Feldman et al. [15].

References

[1] Adler P.S., Mandelbaum A., Nguyen V. and Schwerer E. (1995) From project to process management: An empirically-based framework for analyzing product development time. *Management Science*, 41, 458484. 5

[2] Aguir M.S., Karaesmen F., Aksin O.Z. and Chauvet F. (2004) The impact of retrials on call center performance, *OR Spectrum*, Special Issue on Call Center Management, 26(3),353-376. 1, 10.3

[3] Armony M., Gurvich I. and Mandelbaum A. (2004) Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers. Working paper. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 10.3

[4] Armony M. and Mandelbaum A. (2004) Design, staffing and control of large service systems: The case of a single customer class and multiple server types. Working paper. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 10.3

[5] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: F.J.Kylstra (Ed.), *Performance '81*. North-Holland Publishing Company, 159-179. 2

[6] Bain P. and Taylor P. (2002) Consolidation, "Cowboys" and the developing employment relationship in British, Dutch and US call centres. In: Holtgrewe U., Kerst C. and Shire K. (Ed.), *Re-Organising Service Work*. Ashgate Publishing Limited, 42-62. 1

[7] Bassamboo A., Harrison J.M. and Zeevi A. (2004) Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method. Submitted for publication. 10.2

[8] Bittner S., Schietinger M., Schroth J. and Weinkopf C. (2002) Call Centres in Germany: Employment, Training and Job Design. In: Holtgrewe U., Kerst C. and Shire K. (Ed.), *Re-Organising Service Work*. Ashgate Publishing Limited, 63-85. 1

[9] Borst S., Mandelbaum A., and Reiman M. (2004), Dimensioning large call centers, *Operations Research*, 52(1), 17-34. 4.1, 10.1, 10.1

[10] Borst S., Mandelbaum A., Reiman M. and Zeltyn S. (2004) Dimensioning call centers with abandonment. In prepatation. 10.1

[11] Brown L.D., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (2002) Statistical analysis of a telephone call center: a queueing science perspective. To be published in *JASA*. ∗, 4.8, 4.8, 5, 10, 11, 8.4, 8.4, 10.2

[12] Call Center Data (2002) Technion, Israel Institute of Technology. Available at `http://iew3.technion.ac.il/serveng/callcenterdata/index.html`. 4.8, 4.8, 15

[13] Cleveland B., Mayben J. (1997) *Call Center Management on Fast Forward*. Annapolis: Call Center Press. 2, 5

[14] Erlang A.K. (1948) On the rational determination of the number of circuits. In *The life and works of A.K.Erlang*. Brockmeyer E., Halstrom H.L. and Jensen A., eds. Copenhagen: The Copenhagen Telephone Company. 6

[15] Feldman Z., Mandelbaum A., Massey W. and Whitt W. (2004) Staffing of time-varying queues to achieve time-stable performance. Submitted to *Management Science*. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 10.3

[16] 4CallCenters Software (2002). Available at http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm. 4.1, 4.3, 9

[17] Gans N., Koole G. and Mandelbaum A. (2003) Telephone call centers: a tutorial and literature review. Invited review paper, *Manufacturing and Service Operations Management*, 5(2), 79-141. 1, 4.1, 10.3

[18] Garnett O. and Mandelbaum A. (2000) An Introduction to Skills-Based Routing and its Operational Complexities. Teaching note, Technion, Israel. Available at http://iew3.technion.ac.il/serveng2004/Lectures/SBR.pdf. 2

[19] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a telephone call-center with impatient customers. *Manufacturing and Service Operations Management*, 4,208-227. ∗, 2, 4.1, 4.3, 4.3, 6, 6, 6.2

[20] Gurvich I. (2004) Design and Control of the M/M/N Queue with Multi-Class Customers and Many Servers. M.Sc. Thesis, Technion, 2004. Available at http://iew3.technion.ac.il/serveng/References/references.html. 10.3

[21] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588. 6

[22] Helber S. and Mandelbaum A. (2004) GIF Research Proposal. ∗, 1

[23] Helber S. and Stolletz R. (2004) *Call Center Management in der Praxis*. Springer-Verlag, Berlin, Heidelberg. (In German) 1

[24] Help Desk and Customer Support Practice Report; May 1997 survey results. The Help Desk Institute, SOFTBANK Forums, 1997. 2

[25] Mandelbaum A., Sakov A. and Zeltyn S. (2001) Empirical analysis of a call center. Technical report, Technion. Available at http://iew3.technion.ac.il/serveng/References/references.html. 4.8, 8.1

[26] Mandelbaum A. (2004) Call Centers. Research Bibliography with Abstracts. Version 6. Available at http://iew3.technion.ac.il/serveng/References/references.html. 1, 10.3

[27] Mandelbaum A. and Shimkin N. (2000) A model for rational abandonment from invisible queues. *Queueing Systems: Theory and Applications (QUESTA)*, 36, 141-173. 8.3

[28] Mandelbaum A. and Zeltyn S. (2004) The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/N+G Queue. *OR Spectrum*, Special Issue on Call Center Management, 26(3), 377-411. $*$, 4.8

[29] Mandelbaum A. and Zeltyn S. (2004) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. Submitted to *QUESTA*. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 4.8, 6, 6

[30] Mandelbaum A. and Zeltyn S. (2004) M/M/n+G queue. Summary of performance measures. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 4.7, 6

[31] Palm C. (1957) Research on telephone traffic carried by full availability groups. Tele, vol.1, 107 pp. (English translation of results first published in 1946 in Swedish in the same journal, which was then entitled *Tekniska Meddelanden fran Kungl. Telegrafstyrelsen.*) 2, 3, 3, 4.4, 7.1

[32] Riordan J. (1962) Stochastic Service Systems, Wiley. 4.7

[33] Shimkin N. and Mandelbaum A. (2004) Rational abandonment from tele-queues: non-linear waiting costs with heterogeneous preferences. *Queueing Systems: Theory and Applications (QUESTA)*, 47, 117-146. 8.3

[34] Whitt W. (2004) Fluid Models for Many-Server Queues with Abandonments. Submitted to *Operations Research*. 6, 6.2

[35] Whitt W. (2004) Two Fluid Approximations for Multi-Server Queues with Abandonments. Submitted to *Operations Research Letters*. 6

[36] Whitt W. (2004) Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. Submitted to *Management Science*. 10.2

[37] Whitt W. (2004) Sensitivity of Performance in the Erlang A Model to Changes in the Model Parameters. Submitted to *Operations Research*. 10.2

[38] Wallace R.B. and Whitt W. (2004) A staffing algorithm for call centers with skill based routing. Submitted to *Manufacturing and Service Operations Management (M&SOM)*. 10.3

[39] Zeltyn S. (2004) Call centers with impatient customers: exact analysis and many-server asymptotics of the M/M/n+G queue, Ph.D. Thesis, Technion. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 2

[40] Zohar E., Mandelbaum A. and Shimkin N. (2002) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science*, 48, 566-583. ∗, 5, 7.1, 8.3

Appendix

A Derivation of some Erlang-A performance measures

Steady-state distribution. Using formulae (3.4), (3.5) and definition (3.6) one gets

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

$$= \frac{(\lambda/\mu)^n}{n!} \cdot \left[\frac{1}{E_{1,n}} + \sum_{j=1}^{\infty} \frac{(\lambda/\theta)^j}{\prod_{k=1}^{j} \left(\frac{n\mu}{\theta} + k \right)} \right] = \frac{(\lambda/\mu)^n}{n!} \cdot \left[\frac{1}{E_{1,n}} + A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right].$$

Hence

$$\pi_0 = \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right] \cdot E_{1,n}} \cdot \frac{n!}{(\lambda/\mu)^n}.$$

For $1 \le j \le n$

$$\pi_j = \pi_0 \cdot \frac{(\lambda/\mu)^j}{j!} = \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right] \cdot E_{1,n}} \cdot \frac{n!}{j! \cdot (\lambda/\mu)^{n-j}}.$$

Specifically,

$$\pi_n = \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right] \cdot E_{1,n}}. \tag{A.1}$$

Finally, for $j > n$,

$$\pi_j = \pi_n \cdot \frac{\lambda^{j-n}}{\prod_{k=1}^{j-n} (n\mu + k\theta)} = \frac{E_{1,n}}{1 + \left(A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta} \right) - 1 \right) \cdot E_{1,n}} \cdot \frac{\left(\frac{\lambda}{\theta} \right)^{j-n}}{\prod_{k=1}^{j-n} \left(\frac{n\mu}{\theta} + k \right)}. \tag{A.2}$$

Probability of wait. From PASTA, (A.1) and (A.2), the delay probability is equal to

$$\mathrm{P}\{W > 0\} \;=\; \sum_{j=n}^{\infty} \pi_j \;=\; \frac{E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}} \cdot \left[1 + \sum_{j=n+1}^{\infty} \frac{(\lambda/\theta)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta} + k\right)}\right]$$

$$= \frac{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) \cdot E_{1,n}}{1 + \left[A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right) - 1\right] \cdot E_{1,n}}\,. \tag{A.3}$$

Probability to abandon. First, we need to perform some preliminary calculations. Differentiating (3.5), we get

$$\frac{\partial}{\partial y} A(x,y) \;=\; \frac{\partial}{\partial y}\left[\frac{x e^y}{y^x}\gamma(x,y)\right] \;=\; \frac{x}{y} + \left(1 - \frac{x}{y}\right) \cdot A(x,y)\,.$$

Then, for $x > 0$, $y > 0$,

$$\sum_{j=0}^{\infty} \frac{(j+1)y^j}{\prod_{k=1}^{j+1}(x+k)} \;=\; \frac{\partial}{\partial y}\left[\sum_{j=1}^{\infty} \frac{y^j}{\prod_{k=1}^{j}(x+k)}\right]$$

$$= \frac{\partial}{\partial y}[A(x,y) - 1] \;=\; \frac{\partial}{\partial y} A(x,y) \;=\; \frac{x}{y} + \left(1 - \frac{x}{y}\right) \cdot A(x,y)\,. \tag{A.4}$$

Using (A.3) and (4.2), the conditional probability to abandon is equal to

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;=\; \frac{\sum_{j=n}^{\infty} \pi_j \cdot \mathrm{P}_{j-n}\{\mathrm{Ab}\}}{\mathrm{P}\{W > 0\}}$$

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \sum_{j=n}^{\infty} \frac{\left(\frac{\lambda}{\theta}\right)^{j-n}}{\prod_{k=1}^{j-n}\left(\frac{n\mu}{\theta} + k\right)} \cdot \frac{\theta(j+1-n)}{n\mu + \theta(j+1-n)}$$

(by convention, $\prod_{k=1}^{0}\left(\frac{n\mu}{\theta} + k\right) \triangleq 1$)

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda}{\theta}\right)^{j} \cdot (j+1)}{\prod_{k=1}^{j+1}\left(\frac{n\mu}{\theta} + k\right)} \;=\; \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \frac{\partial}{\partial y}\left[A\left(\frac{n\mu}{\theta}, y\right)\right]_{y=\lambda/\theta}$$

$$= \frac{1}{A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} \cdot \left[\frac{n\mu}{\lambda} + \left(1 - \frac{n\mu}{\lambda}\right) A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)\right] \;=\; \frac{1}{\rho A\left(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}\right)} + 1 - \frac{1}{\rho}\,,$$

where the last line follows from (A.4).