Designing a Call Center with an IVR (Interactive Voice Response)

Polyna Khudyakov, Paul Feigin, Avishai Mandelbaum

Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

emails: polyna@tx.technion.ac.il, paulf@ie.technion.ac.il, avim@tx.technion.ac.il

October 14, 2010

Abstract

A call center is a service operation that caters to customer needs via the telephone. Call centers typically consist of agents that serve customers, telephone lines, an Interactive Voice Response (IVR) unit, and a switch that routes calls to agents.

In this paper we study a Markovian model for a call center with an IVR. We calculate operational performance measures, such as the probability for a busy signal and the average wait time for an agent. Exact calculations of these measures are cumbersome and they lack insight. We thus approximate the measures in an asymptotic regime known as QED (Quality & Efficiency Driven) or the Halfin-Whitt regime, which accommodates moderate to large call centers. The approximations are both insightful and easy to apply (for up to 1000's of agents). They yield, as special cases, known and novel approximations for the M/M/N/N (Erlang-B), M/M/S (Erlang-C) and M/M/S/N queue.[1]

Key words. Queues, Closed Queueing Networks; Call or Contact Centers, Impatience, Busy Signals; IVR, VRU; QED or Halfin-Whitt regime; Asymptotic Analysis.

1 Introduction

More than $300 billion is spent annually on call centers around the world [8]. Increased competition, deregulation and rising customer acquisition costs highlight the importance of high-quality customer service and high-efficiency operations; and to achieve *both*, most leading companies are deploying new technologies, such us enhanced Interactive Voice Response (IVR), natural speech self-service options and others. IVR systems are specialized technologies designed to enable self-service of callers, without

[1] **Acknowledgements.** The research was supported by BSF (Binational Science Foundation) grant 2001685/2005175, ISF (Israeli Science Foundation) grants 388/99, 126/02 and 1046/04 and by the Technion funds for the promotion of research and sponsored research. The authors thank Valery Trofimov, from the Technion's SEE Laboratory, for his help, interest and valuable hints in data analysis.

the assistance of human agents. The IVR technology helps call centers prevent costs from rising (and often reduce costs), while hopefully improve service levels, revenue and hence profits.

A typical call center spends about two thirds of its operational costs [21] on salaries. However, it would be over simplistic to drive costs down by reducing the number of agents, because small changes in staffing levels can effect dramatically the service level. Thus, a main goal of a call center manager is to establish an appropriate tradeoff between staffing cost and service level. Here queueing models come to the rescue, by yielding performance-analysis tools that support this tradeoff. Our paper develops and analyzes such a model, specifically one for a call center with an IVR.

The mathematical framework considered here is a many-server heavy-traffic asymptotic regime, which is referred to as the QED (Quality and Efficiency Driven) regime. Systems that operate in the QED regime enjoy a combination of very high efficiency together with very high quality of service; see Gans, Koole and Mandelbaum [6]. A mathematical characterization of the QED regime for the GI/M/S queue was established by Halfin and Whitt [10] as having a non-trivial limit (within (0,1)) of the fraction of delayed customers, with S increasing indefinitely. This characterization carries over to GI/D/S [12], M/M/S with exponential patience [7] and with general patience [19]. But the QED regime was explicitly recognized already in Erlang's 1923 paper (that appeared in [5]), which addresses both Erlang-B (M/M/S/S) and Erlang-C (M/M/S) models. Later on, extensive related work took place in various telecom companies but little has been openly documented. A precise characterization of the asymptotic expansion of the blocking probability, for Erlang-B in the QED regime, was given by Jagerman [11]; see also Whitt [25], and then Massey and Wallace [18] for the analysis of finite buffers.

Erlang's characterization of the QED regime was in terms of the *square-root staffing principle*; see the expression for S, in (10) below. This principle is formulated in terms of the offered load λ/μ, where λ is the arrival rate, μ is the service rate; then λ/μ denotes the amount of work, measured in units of service-time, that arrives to the system per unit of time. The square-root principle has two parts to it: first, the conceptual observation that the safety staffing level is proportional to the square-root of the offered load; and second, the explicit calculation of the proportionality coefficient. Borst, Mandelbaum and Reiman [1] developed a framework that accommodates both of these needs. More important, however, is the fact that their approach and framework allow an arbitrary cost structure, having the potential to generalize beyond Erlang-C. The square-root staffing principle arises also for the M/M/S/N queue [18], for M/M/S+M [7], and other models, as surveyed in Gans et al. [6].

Analytical models of a Call Center with an IVR were developed by Brandt, Brandt, Spahl and Weber [2]. They show (and we shall use this fact later on) that it is possible to replace the semi-open network of their model with a closed Jackson network. Such a network has the well-known product form solution for its stationary distribution. This product-form distribution was used by Srinivasan, Talim and Wang [22] in order to calculate expressions for the probability to find all lines busy and the conditional distribution function of the waiting time before service.

In this paper we first consider, in Section 2, the model of a Call Center with an IVR, as proposed by Srinivasan et al. [22]. In Section 3, we derive QED approximations of frequently used performance measures, which supports decision-making for call center managers and helps in resolving the staffing problem. Then, in Section 4, we equip customers with finite patience. This gives rise to IVR models with impatient customers (who abandon) for which we derive QED approximations as well. (A special case is the $M/M/S/N+M$ queue, which adds a busy-signal feature to Erlang-A [7].) In Section 5, we analyze the behavior of QED performance measures. In Section 6, we validate our approximations against data from a real call center, thus establishing their applicability. We conclude, in Section

7, with optimizing (asymptotically) the number of trunk lines and number of servers, subject to constraints on system performance.

2 A model for a call center with an IVR

We consider the following model of a call center, as depicted in Figure 1. The arrival process is a Poisson process with rate λ. There are N trunk lines and S agents in the system ($S \leq N$). Arriving customers enter the system only if there is an idle trunk line. When this is the case, the customer is first served by an IVR processor. We assume that the IVR processing times are independent and identically distributed exponential random variables with rate θ. After the IVR, a call may leave the system with probability $1-p$ or proceed to request service from an agent with probability p.

We assume for now that there are no abandonments in our model. (Abandonment will be incorporated in Section 4.2). Agents' service times are considered as independent identically distributed exponential random variables with rate μ, which are independent of the arrival times and IVR processing times. As mentioned, if a call finds the system full, i.e., all N trunk lines are busy, it is lost.

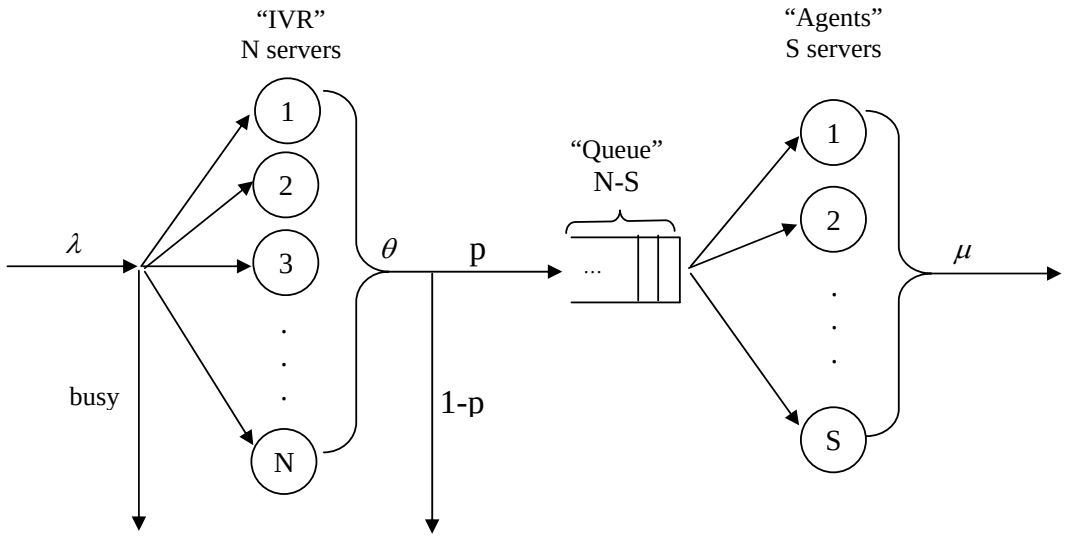

Figure 1: **Schematic model of a queueing system with an IVR, S agents and N trunk lines.**

We now view our model as a system with two multi-server queues connected in series (Figure 1). The first one represents the IVR processor. This processor can handle at most N jobs at a time, where N is the total number of trunk lines available. The second queue represents the agents' pool which can handle at most S incoming calls at a time. The number of agents is naturally less than the number of trunk lines available, i.e. $S \leq N$. Moreover, N is also an upper bound for the total number of customers within the system: at the IVR plus waiting to be served plus being served by agents.

Let $Q(t) = (Q_1(t), Q_2(t))$ represent the number of calls at the IVR processor and at the agents' pool at time t, respectively. Since there are only N trunk lines then $Q_1(t) + Q_2(t) \leq N$, for all $t \geq 0$.

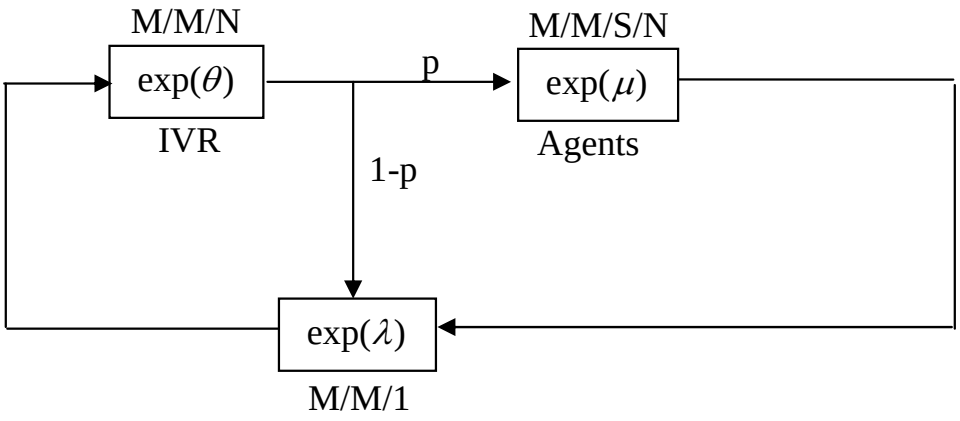

Figure 2: **Schematic model of a corresponding closed Jackson network.**

Note that the stochastic process $Q = \{Q(t),\ t \geq 0\}$ is a finite-state continuous-time Markov chain. We shall denote its states by the pairs $\{(i,j) \mid\ i+j \leq N,\ \ i,j \geq 0\}$.

As shown in [2] (see [14] for details), one can consider our model as 2 stations within a 3-station closed Jackson network, by introducing a fictitious state-dependent queue: there are N entities circulating in the network; service times in the first, second, and third stations are exponential with rates θ, μ and λ respectively; and the number of servers are N, S, and 1, respectively. This 3-station closed Jackson network has a product form solution for its stationary distribution (see Figure 2).

By normalization, which yields the stationary distribution of our model of a call center with an IVR, we deduce the stationary probabilities $\pi(i,j)$ of having i calls at the IVR and j calls at the agents' station. This can be written in a normalized product form as follows:

$$\pi(i,j) = \begin{cases} \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{j!}\left(\dfrac{p\lambda}{\mu}\right)^j, & j \leq S,\ \ 0 \leq i+j \leq N; \\ \pi_0 \dfrac{1}{i!}\left(\dfrac{\lambda}{\theta}\right)^i \dfrac{1}{S!S^{j-S}}\left(\dfrac{p\lambda}{\mu}\right)^j & j \geq S,\ \ 0 \leq i+j \leq N; \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where

$$\pi_0 = \left(\sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}}\left(\frac{p\lambda}{\mu}\right)^j + \sum_{i+j \leq N, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{p\lambda}{\mu}\right)^j \right)^{-1}. \tag{2}$$

Formally, for all states (i,j), we have $\pi(i,j) = \lim_{t\to\infty} P\{Q_1(t) = i,\ Q_2(t) = j\}$.

Define the waiting time W as the time spent by customers, who opt for service, from just after they leave the IVR until they start service by an agent. We say that the system is in state (k,j), $0 \leq j \leq k \leq N$, when it accommodates exactly k calls, and j is the number of calls in the agents' station (queued or served); hence, $k-j$ is the number of calls in the IVR. The distribution function of the waiting time and the probability that a call starts its service immediately after leaving the IVR were derived by Srinivasan, Talim and Wang in [22]; for our future use, these are given by

$$P(W \le t) \triangleq 1 - \sum_{k=S+1}^{N} \sum_{j=S}^{k-1} \chi(k,j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!}, \qquad for \quad all \quad t > 0; \tag{3}$$

$$P(W = 0) \triangleq \sum_{k=1}^{N} \sum_{j=0}^{min(k,S)-1} \chi(k,j). \tag{4}$$

Here $\chi(k,j)$, $0 \le j < k \le N$, is the probability that the system is in state (k,j), given that a call (among the $k-j$ customers) is about to complete its IVR service:

$$\chi(k,j) = \frac{(k-j)\,\pi(k-j,j)}{\sum_{l=0}^{N} \sum_{m=0}^{l} (l-m)\,\pi(l-m,m)}. \tag{5}$$

The *expected waiting time* $E[W]$ (or, as it is called in practice, *Average Speed of Answer (ASA)*) can be derived from (3) via the tail's formula, which yields

$$E[W] = \frac{1}{\mu S} \sum_{\hat{k}=S+1}^{N} \sum_{j=S}^{\hat{k}-1} \chi(k,j)(j-S+1). \tag{6}$$

The fraction of the customers that wait in queue, which corresponds to the *delay probability*, is given by

$$P(W > 0) = \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j). \tag{7}$$

More specifically, equation (7) gives the conditional probability that a calling customer does not immediately reach an agent, given that the calling customer is not blocked, i.e., $P(W > 0)$ is the *delay probability for served customers*. This conditional probability can be reduced to an unconditional probability via the "Arrival Theorem" [4]. Specifically, for the system with N trunk lines and S agents, the fraction of customers that are required to wait, after their IVR service, coincides with the probability that a system with $N-1$ trunk lines and S agents has all its agents busy, namely

$$P_N(W > 0) = P_{N-1}(Q_2(\infty) \ge S). \tag{8}$$

Another measure for the service level of a call center is the probability that an arriving call finds all trunk lines busy, which results in a busy-signal. It was found in [22] and has the following form:

$$P(block) = \pi_0 \left(\frac{\lambda^N}{N!} \left(\frac{1}{\theta} + \frac{p}{\mu} \right)^N + \sum_{j=S+1}^{N} \frac{1}{(N-j)!} \left(\frac{1}{S! S^{j-S}} - \frac{1}{j!} \right) \left(\frac{\lambda}{\theta} \right)^{N-j} \left(\frac{p\lambda}{\mu} \right)^j \right). \tag{9}$$

Formulae (1) - (3) and (6) - (9) provide the basis for calculating all other measures of operational service quality - for more details, readers are referred to our Internet Supplement [14].

3 Asymptotic analysis in the QED regime

The ultimate goal of this section is to establish rules of thumb for solving the dimensioning (staffing and trunking) problem for a call center with an IVR. This will be done analogously to Halfin and Whitt [10] and Massey and Wallace [18].

3.1 Our asymptotic regime

All the following approximations will be derived when the arrival rate λ tends to infinity. In order for the system to not become overloaded, one must let the number of agents S and the number of trunk lines N tend to infinity as well.

To motivate our asymptotic regime, we view the model of a call center with an IVR as an extension of the M/M/S/N queue. The latter was investigated by Massey and Wallace [18], who performed QED asymptotic analysis as λ, S and N tend to ∞ simultaneously, under the following assumptions:

$$\begin{aligned} &(i) \quad N - S = \eta\sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), \quad 0 < \eta < \infty; \\ &(ii) \quad S = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o\left(\sqrt{\lambda}\right), \quad -\infty < \beta < \infty. \end{aligned} \tag{10}$$

(Actually, β in [18] was assumed positive because of the reliance on the M/M/S queue in the analysis. We shall dispose of this assumption momentarily.) For our call center with an IVR, in view of (10), we need $S + \eta_1\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right)$ trunk lines for the agents' station and $\frac{\lambda}{\theta} + \eta_2\sqrt{\frac{\lambda}{\theta}} + o\left(\sqrt{\lambda}\right)$ trunk lines for the IVR station, for some constants η_1 and η_2. This gives rise to the following QED conditions for our system: let λ, S and N tend to ∞ simultaneously, so that

$$\begin{aligned} &(i) \quad N - S = \eta_1\sqrt{\frac{\lambda p}{\mu}} + \frac{\lambda}{\theta} + \eta_2\sqrt{\frac{\lambda}{\theta}} + o\left(\sqrt{\lambda}\right), \quad -\infty < \eta_1, \eta_2 < \infty; \\ &(ii) \quad S = \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + o\left(\sqrt{\lambda}\right), \quad -\infty < \beta < \infty. \end{aligned} \tag{11}$$

Note that we have three parameters η_1, η_2 and β in (11). However, one can reduce the number of parameters to two, denoted β and η, by introducing $\eta = \eta_1\sqrt{\frac{p\theta}{\mu}} + \eta_2$. Indeed, as λ, S and N tend to ∞ simultaneous, conditions (11) have also the following equivalent form:

$$\textbf{QED:} \begin{cases} (i) & \lim\limits_{\lambda\to\infty} \frac{N - S - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ (ii) & \lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty\ . \end{cases} \tag{12}$$

Conditions (12) constitute a square-root safety-staffing principle, which recommends the number of agents S to be the offered load $\left(\frac{\lambda p}{\mu S}\right)$, plus safety-staffing $\beta\sqrt{\frac{\lambda p}{\mu S}}$ against stochastic variability. Analogously, the number of lines N, by rule (12), is the sum of the number of agents S and offered load in the IVR, λ/θ, plus an addition of safety-trunking $\eta\sqrt{\frac{\lambda}{\theta}}$.

3.2 QED approximations

Our asymptotic analysis is based on representing the performance measures (3), (6) and (7) in terms of building blocks. The asymptotic behavior of these blocks then determines that of the original measures. An example of such an analysis appears in the Appendix. The full proof is given in our Internet Supplement [14].

Theorem 3.1. *[QED] Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions* (12)*, where μ, p, θ are fixed. Then the asymptotic behavior of our system, in Figure 1, is captured by the following performance measures:*

- *the probability $P(W > 0)$ that a served customer waits after the IVR:*

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1+\dfrac{\gamma}{\xi_1-\xi_2}\right)^{-1}, & \beta\neq 0, \\ \left(1+\dfrac{\sqrt{2\pi}}{c}\dfrac{\gamma}{\xi_3}\right)^{-1}, & \beta = 0; \end{cases} \tag{13}$$

- *the probability of a busy-signal:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \begin{cases} \dfrac{\nu}{\gamma+\xi_1-\xi_2}, & \beta\neq 0, \\ \dfrac{\nu}{\gamma+\xi_3}, & \beta = 0; \end{cases} \tag{14}$$

- *the expected waiting time:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W] = \begin{cases} \dfrac{1}{\mu}\dfrac{\xi_1-(\beta^2c^2-\beta\eta c-1)\xi_2}{\gamma+\xi_1-\xi_2}, & \beta\neq 0, \\ \dfrac{1}{2\mu}\dfrac{\eta c\xi_3}{\gamma+\xi_3}, & \beta = 0; \end{cases} \tag{15}$$

- *the conditional density function of waiting time:*

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \beta\mu e^{-\beta\mu t}\dfrac{\Phi(\eta-t\sqrt{p\mu\theta})}{\Phi(\eta)-exp\{\frac{1}{2}c^2\beta^2-\eta\beta\}\Phi(\eta-\beta/c)}, & \beta\neq 0, \\ \dfrac{\mu\Phi(\eta-t\sqrt{p\mu\theta})}{\xi_3}, & \beta = 0; \end{cases} \tag{16}$$

In the above, φ and Φ are, respectively, the density and distribution functions of the standard normal distribution, and

$$\gamma = \int_{-\infty}^{\beta} \Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \qquad c=\sqrt{\frac{\mu}{p\theta}}, \qquad \xi_1 = \frac{\varphi(\beta)\Phi(\eta)}{\beta},$$

$$\xi_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\{\frac{1}{2}(\eta-\beta/c)^2\}\Phi(\eta-\beta/c)}{\beta}, \qquad \xi_3 = \sqrt{\frac{1}{2\pi}}c\,(\eta\Phi(\eta)+\varphi(\eta)),$$

$$\nu = \frac{1}{\sqrt{1+c}}\varphi\left(\frac{\eta c+\beta}{\sqrt{1+c^2}}\right)\Phi\left(\frac{\beta c-\eta}{\sqrt{1+c^2}}\right) + \varphi(\sqrt{\eta^2+\beta^2})\exp\{\frac{1}{2}(\eta-\beta/c)^2\}\Phi(\eta-\beta/c).$$

Remark 3.1. *We must distinguish the cases when β equals or does not equal to zero, due to technical problems (division by zero).*

3.3 The M/M/S/N queue

Massey and Wallace [18] derived approximations for the following operational characteristics of the M/M/S/N queue:

- the probability to find the system busy $P(block)$;
- the probability to wait more than t units of time $P(W > t)$;

here λ, S and N tend to ∞ simultaneously so that (10) prevails.

The condition $\eta > 0$ in (10) is natural, because $N-S$ is the maximal queue length, but the condition $\beta > 0$ is not required. The reason of strict positivity of β in [18] is their using the M/M/S queue for calculating operational characteristics of M/M/S/N. In this section, we derive approximations for various performance measures, when $-\infty < \beta < \infty$. (As in the previous section, we distinguish the cases $\beta = 0$ and $\beta \neq 0$, to avoid division by zero.)

Theorem 3.2. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy conditions* (10) [2]*, where μ is fixed. Then the asymptotic behavior of the M/M/S/N system is described in terms of the following performance measures:*

- *the probability of waiting:*

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1+\dfrac{\beta\Phi(\beta)}{\varphi(\beta)(1-e^{-\eta\beta})}\right)^{-1}, & \beta \neq 0, \\ \left(1+\dfrac{\sqrt{\pi}}{\eta\sqrt{2}}\right)^{-1}, & \beta = 0; \end{cases} \tag{17}$$

- *the probability to encounter a full system:*

[2] Note that (10) can be also rewritten in the following form:

$$(i)\quad \lim_{\lambda\to\infty}\frac{N-S}{\sqrt{S}} = \eta, \quad 0<\eta<\infty;$$
$$(ii)\quad \lim_{\lambda\to\infty}\frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad -\infty<\beta<\infty.$$

$$\lim_{\lambda\to\infty}\sqrt{S}P(block) = \begin{cases} \dfrac{\beta\varphi(\beta)e^{-\eta\beta}}{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}, & \beta\neq 0; \\ \dfrac{1}{\sqrt{\dfrac{\pi}{2}}+\eta}, & \beta = 0; \end{cases} \tag{18}$$

- *the expected waiting time:*

$$\lim_{\lambda\to\infty}\sqrt{S}E[W] = \begin{cases} \dfrac{\dfrac{\varphi(\beta)}{\mu}\left[\dfrac{1-e^{-\eta\beta}}{\beta}-\eta e^{-\eta\beta}\right]}{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}, & \beta\neq 0; \\ \dfrac{\eta^2}{2\mu(\eta+\sqrt{\dfrac{\pi}{2}})}, & \beta = 0. \end{cases} \tag{19}$$

- *the density function of waiting time:*

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{S}}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \dfrac{\mu\beta e^{-\mu\beta t}}{(1-e^{-\eta\beta})}, & \mu t<\eta, \quad \beta\neq 0; \\ \dfrac{\mu}{\eta}, & \mu t<\eta, \quad \beta = 0; \\ 0, & \mu t\geq\eta. \end{cases} \tag{20}$$

The proof of Theorem 3.2 can be found in our Internet Supplement [14].

4 Adding abandonment

4.1 Model description

In this section, we add the feature of customers' patience or rather impatience, which could lead to abandonment from the queue prior to service. The modelling assumptions are the same as those in Section 2 (see Figure 1), which are characterized by the parameters (λ, N, θ, p, S, μ). The additional assumption is that if a call waits in the queue, it leaves the system after an exponentially distributed time with rate δ (impatience), or it is answered by an agent, whatever happens first.

As in the case without abandonment, one can consider the model with abandonment as a closed Jackson network, by introducing a fictitious state-dependent queue. The only difference is that the M/M/S/N queue in Figure 2 is now replaced by M/M/S/N+M, with impatience that is distributed $exp(\delta)$.

We can thus consider our model as a three node closed Jackson network, when the stationary probabilities $\pi(i,j)$ of having i calls at the IVR and j calls at the agents station can be written in a

(normalized) product form (see [14] for details). We consider the probability of abandonment, given waiting

$$P(Ab|W>0)=\frac{\sum_{i=0}^{N-S}\sum_{j=S+1}^{N}\pi(i,j)(j-S)\delta}{\sum_{i=0}^{N-S}\sum_{j=S+1}^{N}\pi(i,j)\left(S\mu+(j-S)\delta\right)},$$

as the main performance characteristic that captures the abandonment phenomenon. Formulae (1) - (3) and (6) - (9) provide the basis for calculating all other measures of operational service quality. For all calculations we use the stationary probabilities $\pi(i,j)$ of the corresponding closed Jackson network.

4.2 A call center with abandonment and an IVR

As previously, our goal is to derive approximations for the case when the arrival rate λ tends to ∞. The asymptotic domain is the same as in the case without abandonment. (Now, however, the case $\beta=0$ does not require special treatment.) Analogously to the calculations in Section 3.2, we now introduce the approximations for performance measures with abandonments.

Theorem 4.1. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions* (12)*, where μ,p,θ are fixed. Then the asymptotic behavior of the system is described in terms of the following performance measures:*

- *the probability $P(W>0)$ that a customer waits after the IVR:*

$$\lim_{\lambda\to\infty}P(W>0)=\left(1+\frac{\gamma}{\xi_1-\xi_2}\right)^{-1};\tag{21}$$

- *the probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty}\sqrt{S}P(Ab|W>0)=\frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta)-\beta\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty}\Phi(\eta+(\beta\sqrt{\frac{\mu}{\delta}}-t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}{\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty}\Phi(\eta+(\beta\sqrt{\frac{\mu}{\delta}}-t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}.\tag{22}$$

In the above,

$$\xi_1=\sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\int\limits_{-\infty}^{\eta}\Phi\left((\eta-t)\sqrt{\frac{\delta}{p\theta}}+\beta\sqrt{\frac{\mu}{\delta}}\right)\varphi(t)dt,$$

$$\xi_2=\sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\Phi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta),\qquad\qquad\gamma=\int\limits_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt.$$

There exists a linear relation between $P(Ab|W>0)$ and $E[W|W>0]$, which applies for the M/M/S/N+M queue (see [19] for details):

$$E[W|W>0]=\frac{1}{\delta}P(Ab|W>0).\tag{23}$$

One can then deduce from equation (22) an approximation for the expected waiting time, given wait. An approximation to the probability of blocking can be calculated, similarly to the one presented in Theorem 3.1 (see our Internet supplement [14]).

4.3 The M/M/S/N+M queue

We now present approximations for performance characteristics of the M/M/S/N+M queue. As in [20], we apply the approximations under Conditions (10). The infinite-buffer case M/M/S+M, known as Erlang-A, was analyzed in [7]. It is becoming the standard queueing engine of work-force management systems, in telephone call centers. What we do here is to add the busy-signal feature to Erlang-A. The results are formalized in the following theorem.

Theorem 4.2. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy Conditions* (10)*, where μ is fixed* [3]*. Then the asymptotic behavior of the system is described in terms of the following performance measures:*

- *the probability to wait:*

$$\lim_{\lambda\to\infty} P(W>0) = \left(1 + \frac{\sqrt{\frac{\delta}{\mu}}\Phi(\beta)\varphi(\beta\sqrt{\frac{\mu}{\delta}})}{\varphi(\beta)\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]}\right)^{-1}, \tag{24}$$

- *the probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(Ab|W>0) = \frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}}) - \beta\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]}{\Phi(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})}. \tag{25}$$

The proofs of Theorems 4.1 and 4.2 are analogous to those of Theorems 3.1 and 3.2. They appear in our Internet supplement [14].

5 Rules of thumb

We derived approximations for performance measures in the QED regime (Quality and Efficiency Driven), as characterized by Conditions (12). A detailed comparison between exact and approximated performance was carried out in [13]. This analysis shows that the approximations are excellent, and often work perfectly, even *outside* the QED regime. In this section, we attempt to chart the boundary of this "outside". We summarize our findings through practical rules-of-thumb, expressed via the offered load $R = \frac{\lambda p}{\mu}$. These rules were derived via extensive numerical analysis (using Maple) of our analytical results; for an elaboration, readers are referred to [13], Chapter 5.

In the analysis of call center operations, it has been found useful to distinguish three operational regimes [1]:

[3] When $\eta = 0$, the M/M/S/N+M queue is equivalent to the M/M/S/S loss system. In this case, $P(W>0)$ and $P(Ab|W>0)$ are equal to 0 and their approximations become irrelevant.

(ED) Efficiency-Driven, meaning under-staffing with respect to the offered load, to achieve very high resource utilization;

(QD) Quality-Driven, meaning over-staffing with respect to the offered load, to achieve very high service level;

(QED) Quality-and Efficiency-Driven, meaning rationalized staffing that attempts to balance high levels of resource efficiency and service quality.

We shall use the characterization of the operational regimes, as formulated in [17], in order to specify numerical ranges for the parameters β and η, in the $M/M/S/N$ queue and in the model of an IVR with and without abandonment.

Table 1: Rules-of thumb for operational regimes. (AST stands for Average Service Time.)

	ED	**QED**	**QD**
Staffing	$S \approx R - \gamma R$	$S \approx R + \beta\sqrt{R}$	$S \approx R + \gamma R$
% Delayed	$\approx 100\%$	constant over time (25%-75%)	$\approx 0\%$
% Abandoned	10% - 25%	1% - 5%	≈ 0
Average Wait	$\geq 10\% \cdot AST$	$\leq 10\% \cdot AST$	≈ 0

The performance measures of a call center with an IVR, *without* abandonment, depend on β, η, $\frac{p\theta}{\mu}$ and S; in particular, large values of $\frac{p\theta}{\mu}$ and S improve performance (see [13] for an elaboration). When one is adding abandonment to the system, one adds a parameter δ describing customers' patience. Large values of δ, corresponding to highly impatient customers, decrease the probability to wait and the probability of blocking, but increase the probability of abandonment. Small values of δ have the opposite influence.

Analysis of the parameter values for η (which determines the number of trunk-lines) and β (determines the number of agents), which ensure QED performance, is presented in our Internet supplement [14], Section 7. Here, we mention only some main findings. From the definition of the QED regime for the $M/M/S/N$ queue, η must be strictly positive ($\eta > 0$), because otherwise there would be hardly any queue and, thus, no reason to be concerned with the probability to wait or to abandon the system. Our findings show that, when $\eta > 3$, the $M/M/S/N$ queue behaves like an $M/M/S$ queue (negligible blocking).

In the case of the system with an IVR, there are no mathematical restrictions for η to be non-negative, but we propose $\eta \geq 0$ because otherwise ($\eta < 0$), the probability of blocking is higher than 0.1. We believe that a call center cannot afford that 10% of its customers encounter a busy signal. Going the other way, a call center can extend the number of trunk lines to avoid the busy-line phenomenon altogether. When $\eta > 3$, the system with an IVR behaves as one with an infinite number of trunk lines.

For the QD and ED regimes (see Table 1), the number of agents can be specified via $0.1 \leq \gamma \leq 0.25$. In the case of QD, the number of agents is over-staffed; limiting the number of trunk lines will cause

unreasonable levels of agent idleness; hence $\eta \geq 3$ makes sense. In the case of ED, the number of agents is under-staffed, and we are interested in reducing the system's offered load. Therefore, we propose to take $\eta = 2$. This choice yields the probability of blocking to be approximately $\gamma/2$ (based on numerical experience).

Our rules of thumb demonstrate that for providing service in the QED regimes (in both cases: with and without an IVR) one requires the number of agents to be close to the system's offered load; the probability of blocking in the system with an IVR is always less than in the system without an IVR. One also observes that the existence of the abandonment phenomena considerably helps provide the same level of service as without abandonment, but with less agents. Moreover, as discussed in [14], it is possible to maintain operational service quality while reducing the number of agents by reducing access to the system. The cost is increased busy signals. Hence, such a solution must result from a tradeoff between the probability of blocking and the probability to abandon, which gives rise to a stochastic control problem such as in [24].

6 Model validation with real data

Our approximations can be of use in the operations management of a call center, for example when trying to maintain a pre-determined level of service quality. We analyzed approximations of a real call center by models with an IVR and without an IVR(M/M/S/N+M), in order to evaluate the value of adding an IVR. This evaluation is the goal of our empirical study, which is based on real data from a large call center. (The size of our call center, around 600 - 700 agents, forces one to apply our approximations, as opposed to exact calculations that are simply numerically prohibitive with that many agents.)

6.1 Data description

The data for the current analysis comes from[4] a call center of a large U.S. bank. The full database archives all the calls handled by the call center over the period of 30 months from March 2001 until September 2003. The call center consists of four different contact centers (nodes), which are connected through communication switches so that, in effect, they can be considered a single system. The call path can be described as follows. Customers, who make a call to the company, are first of all served in the IVR. After that, they either complete the call or opt to be served by an agent. In the latter case, customers typically listen to a message, after which they are routed as will be now described, to join the agents' queue.

The choice of routing is usually performed according to the customer's class, which is determined in the IVR. If all the agents are busy, customers wait in the queue; otherwise, they are served immediately. Customers may abandon the queue before receiving service. If they wait in the queue of a specific node (one of the four connected) for more than 10 seconds, the call is transferred to a common queue - so-called "inter-queue". This means that now the customer will be answered by the first available agent with the appropriate skill, from any of the four nodes. After service by an agent, customers may

[4]The data are available through the Technion's SEELab, after a registration:

`http://seeserver.iem.technion.ac.il/see-terminal/`.

For further details, readers are encouraged to write to see@ie.technion.ac.il

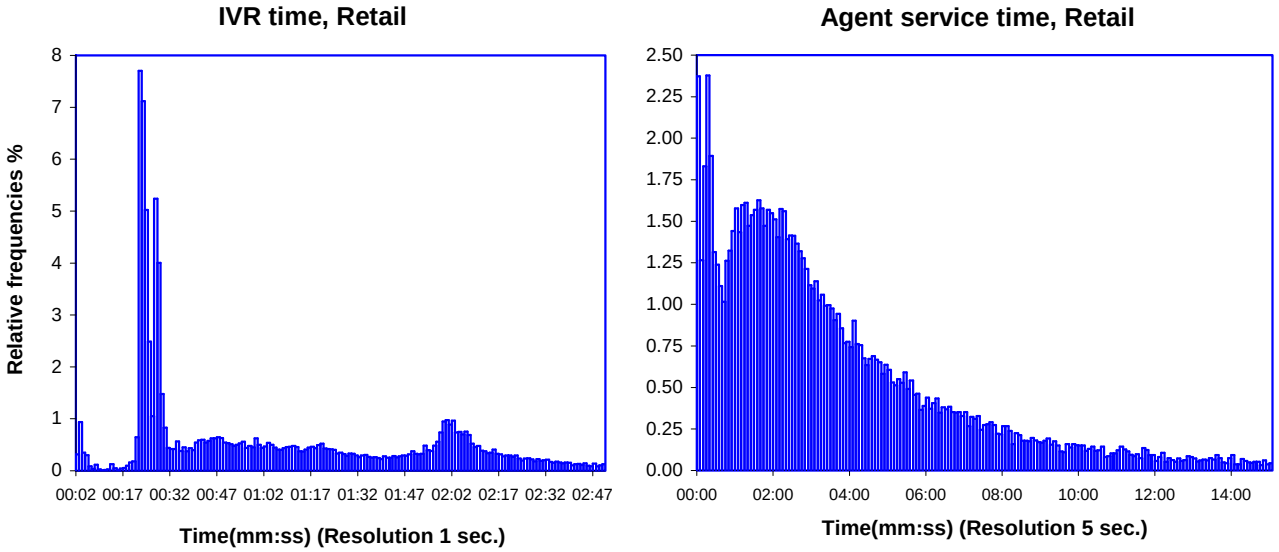

Figure 3: Histogram of the IVR service time and agent service time for "Retail" customers

either leave the system or return to the IVR, from which point a new *sub-call* ensues. The call center is relatively large with about 600 agents per shift, and it is staffed 7 days a week, 24 hours a day.

6.2 Fitting the theoretical model to a real system

The flow of our call center differs somewhat from the models described in Sections 2 and 4. The main difference is that it is possible for the customer to return to the IVR after being served by an agent. This is less common for so-called *Retail* customers who, almost as a rule, complete the call either after receiving service in the IVR or immediately after being served by an agent. We therefore neglect those few calls that return to the IVR. Retail Customers also constitute the vast majority of services (62%), and hence are worthy of special attention.

The possibility that queued customers abandon (hang up) without being answered is not acknowledged in our model from Section 2. We thus compare only the models from section 4 with the real system, namely the model of a call center with an IVR and abandonment and the M/M/S/N+M model.

Our theoretical model assumes exponentially distributed service times, in the IVR as well as for the agents. However, for the real data, neither of these service times have the exponential distribution. Figure 3, produced using the SEEStat program [23], displays the distributions of service time in the IVR and agents' service time, respectively.

Figure 3 exhibits three peaks in the histogram of the IVR service time. The first peak can be attributed to calls of customers who are well familiar with the IVR menu and move fast to Agents service; the second can be attributed to calls that, after an IVR announcement, opt for Agents service; the third peak can be related to the most common services in the IVR.

The distribution of the IVR service time is thus not exponential (see [9]). Additional empirical analysis of our call center was carried out in [23]. It is shown there that, similarly to IVR times, service

times are also not exponential. Indeed, as presented in Figure 3, service time turns out to be log-normal (up to a probability mass near the origin) for up to 93% of calls; the other 7% enjoy fast service for various reasons, for instance: mistaken calls, calls transferred to another service, unidentified calls sometimes transferred to an IVR, etc. (There are, incidentally, adverse reasons for short service times, for example agents "abandoning" their customers; see [3].)

The assumption that the arrival process is a homogeneous Poisson is also over simplistic. A more natural model for arrivals is an inhomogeneous Poisson process, as shown by Brown et al. [3], in fact modified to account for overdispersion (see [16]). However, and as done commonly in practice, if one divides the day into half-hour intervals, we get that within each interval the arrival rate is more or less constant and thus, within such intervals, we treat the arrivals as conforming to a Poisson process.

Even though most of our model assumptions do not prevail in practice, notably the Markovian assumptions, experience has shown that Markovian models still provide *very useful* descriptions of non-Markovian systems (for example, the Erlang-A model in [3]). We thus proceed to validate our models against the US Bank Call Center, and our results will indeed demonstrate that this is a worthwhile insightful undertaking.

6.3 Validation: Comparison of real and approximated performance measures

We consider the Retail operation on April 12, 2001, which is an example of an ordinary week day. The analysis was carried out for data from calls arriving between 07:00 and 18:00. This choice was made since we were interested in investigating the system during periods of a meaningful load. As stated above, time intervals of 30 minutes were considered. Since approximately 8000 calls are made during such intervals, we may expect that approximations for large λ would be appropriate. Moreover, system parameters seem to be reasonably constant over these intervals. It should be noted that, strictly speaking, we are not calculating the actual average arrival rate because we see only the calls that did not find all trunks busy, i.e. encountered a busy signal. Practically, the fraction of customers with busy-signals is very small and hence the difference between the real and approximated (calculated our way) arrival rate is not significant.

The average rate of customers' patience was estimated via the relation (23), which assumes a linear relation between $P(Ab|W > 0)$ and $E[W|W > 0]$. Our data analysis demonstrates that this assumption is not unreasonable for our call center (see an illustration in [14], Section 8).

Estimating the average rate of customers' patience for our data gave varying behavior of this parameter, for example at 14:30 its value is 5 minutes, at 15:00 it equals to 1, and at 15:30 it equals 4. It is not unreasonable that customers' patience does not vary dramatically over each 30-minute period, hence we smoothed the 30-minute values by using the R-function "smooth".

In order to validate our approximations, we must assign an appropriate value for N, the number of trunk lines, which is not available to us. We could consider the simplifying assumption that the number of trunk lines is unlimited. Certainly, call centers are typically designed so that the probability of finding the system busy is very small, but nevertheless it is positive. One approach is to assume that, because the system is heavily loaded, there must be calls that are blocked since there are no explosions. In such circumstances, a naive way of underestimating N for each 30-minute period[5] would

[5]Note that for the system with an IVR, N is calculated from the total sojourn of calls in the IVR, of the agents' queue and in service. For the system without IVR, it depends only on the total sojourn of calls in the agents' queue and service.

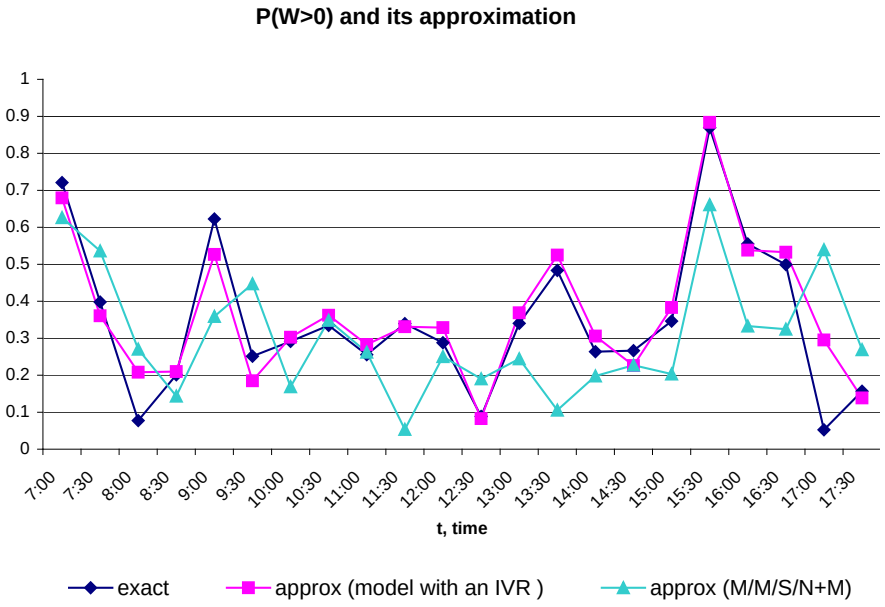

Figure 4: Comparison of approximate and real probabilities to wait.

be the total system sojourn time (in minutes) of all calls that arrived to the system divided by 30 minutes.

The calculation of the number of agents is also somewhat problematic, because the agents who serve retail customers may also serve other types of customers, and vice versa: if all Retail agents are busy, some other agent types may serve Retail customers (see [15] for details). Thus, it is practically impossible to determine the exact number of Retail agents; hence we use an averaged value, specifically the total agent service time divided by 30 minutes.

Figure 4 shows a comparison of the theoretical probability to wait, calculated with the help of the above estimated parameters, against the exact proportion of waiting customers, as estimated from real data. There are three curves: the blue one (line with diamonds) shows the fraction of customers that are delayed in queue before agent service. This fraction is calculated for each half-hour period. The lilac line (squares) shows the approximation from the model with an IVR of the probability of waiting, which is calculated for each half-hour period. The last line (triangles) corresponds to the probability to wait from the M/M/S/N+M queue model. As mentioned, the number of agents (600 and more) renders exact analysis impossible. Our theoretical calculations are based on the QED approximations in Theorem 4.1 (Call Center with an IVR) and Theorem 4.2 (M/M/S/N+M).

Figure 5 demonstrates the real and approximate theoretical conditional probability for customers to abandon. Overall, the approximation is useful, but sometimes we observe significant discrepancies. A possible explanation for these deviations is the sensitivity of this measure under heavy traffic, i.e. a little change of parameters can dramatically change the performance measure.

Considering the overall accuracy of our approximations, one observes that it is satisfactory, especially for the model with an IVR. The theoretical values for this model, in many intervals, are very close to the exact ones. In some intervals, the difference is about 10%, which can be attributed to the

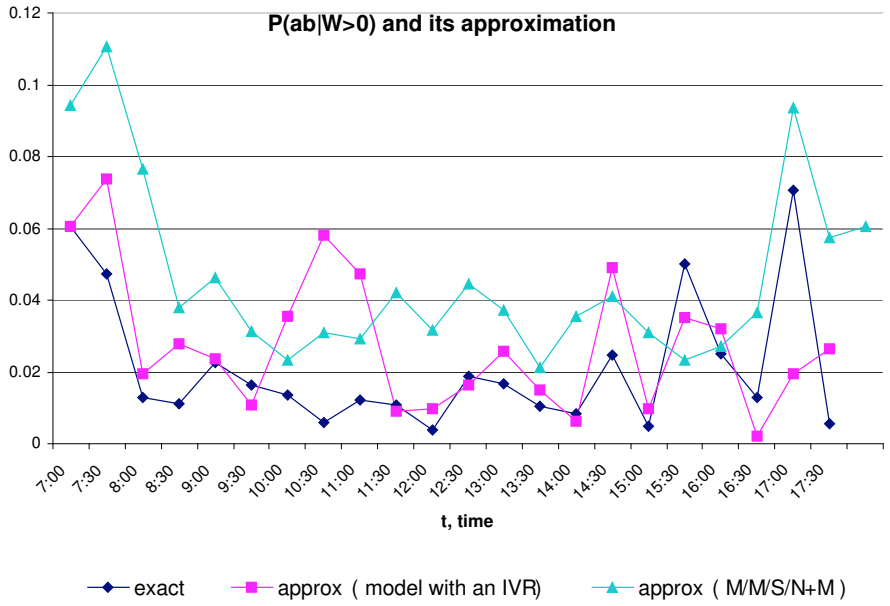

Figure 5: Comparison of approximate and real probabilities to wait.

non-perfect correspondence between the model and the real call center. An additional explanation is our estimation of parameters, such as N and S, which we estimate in a crude way. The approximation from the M/M/S/N+M queue works less well and sometimes does not even reflect the trends seen for the real values: namely, where the real values decrease the approximation increases and vice versa. The reasons for these discrepancies can be the same as previously stated, but most significantly, it is plausibly a consequence of ignoring the IVR influence.

A Appendix

As mentioned, our asymptotic analysis is based on representing the performance measures (3), (6), (7) and (9) in terms of building blocks ($A(\lambda)$ and $B(\lambda)$ below). The asymptotic behavior of these blocks then determines that of the measures. A detailed proof of Theorem 3.1 is presented in [14]. In this Appendix we restrict ourselves to the proof of the approximation (13), for the probability that a customer waits after the IVR. These calculations are representative of the others, all appearing in [14].

According to (1), (2), (5) and (7), the operational characteristic $P(W > 0)$ can be represented as follows:

$$P(W > 0) = \left(1 + \frac{A(\lambda)}{B(\lambda)}\right)^{-1}, \tag{26}$$

where

$$A(\lambda) = \sum_{i+j \le N-1, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \tag{27}$$

and

$$B(\lambda) = e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \sum_{i=0}^{N-S-1} \sum_{j=S}^{N-i-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!S^{j-S}} \left(\frac{\lambda p}{\mu}\right)^j. \tag{28}$$

We now derive QED approximations for $A(\lambda)$ and $B(\lambda)$, as λ, S and N tend to ∞, according to (12).

A.1 Approximating $A(\lambda)$

Consider a partition $\{S_j\}_{j=0}^{l}$ of the interval $[0, S]$:

$$S_j = S - j\Delta, \quad j = 0, 1, ..., l; \quad S_{l+1} = 0, \tag{29}$$

where $\Delta = [\varepsilon\sqrt{\frac{\lambda p}{\mu}}]$, ε is an arbitrary non-negative real and l is a positive integer.

If λ and S tend to infinity and satisfy the assumption (12)*(ii)*, then l is less than S/Δ for λ large enough and all the S_j belong to $[0, S]$, $j = 0, 1, ..., l$.

We emphasize that the length Δ of every interval $[S_{j-1}, S_j]$ depends on λ. The variable $A(\lambda)$ is given by formula (27), where the summation is taken over the trapezoid: $\{(i, j) \mid i \in [0, N-j] \ and \ j \in [0, S-1]\}$, presented in Figure 6. Consider the lower estimate for $A(\lambda)$, given by the following sum:

$$\begin{aligned} A(\lambda) \geq A_1(\lambda) &= \sum_{k=0}^{l} \sum_{j=S_{k+1}}^{S_k-1} \frac{1}{j!} \left(\frac{\lambda p}{\mu}\right)^j e^{-\frac{\lambda p}{\mu}} \sum_{i=0}^{N-S_k} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \\ &= \sum_{k=0}^{l} P(S_{k+1} \leq Z_\lambda < S_k) P(X_\lambda \leq N - S_k), \end{aligned} \tag{30}$$

where

$$Z_\lambda \stackrel{d}{=} Pois\left(\frac{\lambda p}{\mu}\right), \qquad and \qquad X_\lambda \stackrel{d}{=} Pois\left(\frac{\lambda}{\theta}\right). \tag{31}$$

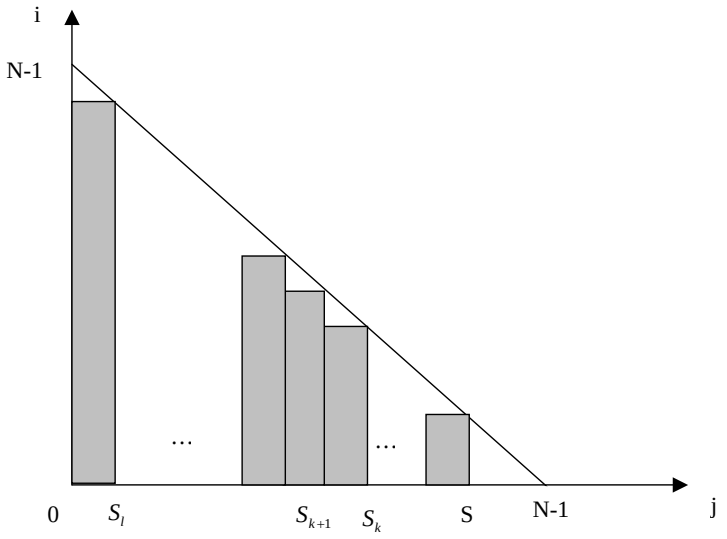

Figure 6: **Area of the summation of the variable $A_1(\lambda)$.**

Applying the Central Limit Theorem and making use of the relations

$$\lim_{\lambda\to\infty}\frac{S_k-\frac{\lambda p}{\mu}}{\sqrt{\frac{\lambda p}{\mu}}}=\beta-k\varepsilon,\quad \lim_{\lambda\to\infty}\frac{N-S_k-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}},\quad k=0,1,...,l, \tag{32}$$

one obtains

$$\lim_{\lambda\to\infty}P(S_{k+1}\le Z_\lambda<S_k)=\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon),\quad k=0,1,...l-1, \tag{33}$$

$$\lim_{\lambda\to\infty}P(0\le Z_\lambda<S_l)=\Phi(\beta-l\varepsilon), \tag{34}$$

$$\lim_{\lambda\to\infty}P(X_\lambda<N-S_k)=\Phi(\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}}),\quad k=0,1,...l. \tag{35}$$

It follows from (30) and (33), (34), (35) that

$$\begin{aligned}\liminf_{\lambda\to\infty}A(\lambda)\ge&\sum_{k=0}^{l-1}\Phi(\eta+k\varepsilon\sqrt{p\theta/\mu})[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]\\&+\Phi(\beta-l\varepsilon)\Phi(\eta+l\varepsilon\sqrt{p\theta/\mu}).\end{aligned} \tag{36}$$

It is easy to see that (36) is the lower Riemann-Stieltjes sum for the integral

$$-\int_0^\infty\Phi\left(\eta+s\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(\beta-s)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \tag{37}$$

corresponding to the partition $\{\beta-k\varepsilon\}_{k=0}^{l}$ of the semi-axis $(-\infty,\beta)$.

Similarly, we obtain the upper Riemann-Stieltjes sum for the integral (37):

$$\limsup_{\lambda\to\infty}A(\lambda)\le\sum_{k=0}^{l-1}\Phi\left(\eta+(k+1)\varepsilon\sqrt{\frac{p\theta}{\mu}}\right)[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]+\Phi(\beta-l\varepsilon). \tag{38}$$

When $\varepsilon\to 0$, the estimates (36), (38) lead to the following equality

$$\lim_{\lambda\to\infty}A(\lambda)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt. \tag{39}$$

A.2 Approximating $B(\lambda)$

We distinguish the two cases $\beta\neq 0$ and $\beta=0$.

A.2.1 Approximating $B(\lambda)$ when $\beta \neq 0$

In this case, we consider $B(\lambda)$ as the difference $B(\lambda) = B_1(\lambda) - B_2(\lambda)$ of two building blocks:

$$B_1(\lambda) = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{\lambda p}{\mu}\right)^S \frac{1}{1-\frac{\lambda p}{S\mu}} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i,$$

and

$$B_2(\lambda) = \frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{\lambda p}{\mu}\right)^S \frac{1}{1-\rho}\rho^{N-S} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i, \qquad \left(\rho = \frac{\lambda p}{S\mu}\right).$$

In view of Stirling's formula, $S! \approx \sqrt{2S\pi}S^S e^{-S}$, one obtains for $B_1(\lambda)$:

$$B_1(\lambda) = \frac{e^{S-\lambda\frac{p}{\mu}}}{\sqrt{2S\pi}S^S}\left(\frac{\lambda p}{\mu}\right)^S \frac{\sqrt{S}}{\beta} \sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} + o(1). \tag{40}$$

The sum in (40) can be rewritten as $P\left(X_\lambda \leq N-S-1\right)$, where $X_\lambda \stackrel{d}{=} Pois(\frac{\lambda}{\theta})$ is a random variable with the Poisson distribution with parameter $\frac{\lambda}{\theta}$, thus $E[X_\lambda] = \frac{\lambda}{\theta}$, $Var[X_\lambda] = \frac{\lambda}{\theta}$. If $\lambda \to \infty$, then $\frac{\lambda}{\theta} \to \infty$ (θ-fixed). Note that

$$P\left(X_\lambda \leq N-S-1\right) = P\left(\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \leq \frac{N-S-1-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}\right). \tag{41}$$

Thus, when $\lambda \to \infty$, by the Central Limit Theorem (Normal approximation to Poisson) we have

$$\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \Rightarrow N(0,1) \tag{42}$$

and due to Assumption *(ii)* in (12) we get

$$\sum_{i=0}^{N-S-1} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \longrightarrow P\left(N(0,1) \leq \eta\right) = \Phi(\eta), \quad when \;\; \lambda \to \infty, \tag{43}$$

where $N(0,1)$ is a standard normal random variable with distribution function Φ. Now, with the use of (40) - (43) we get

$$\lim_{\lambda\to\infty} B_1(\lambda) = \frac{\varphi(\beta)\Phi(\eta)}{\beta}. \tag{44}$$

Similarly, one establishes the limit

$$\lim_{\lambda\to\infty} B_2(\lambda) = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\{\frac{1}{2}(\eta-\beta/c)^2\}\Phi(\eta-\beta/c)}{\beta}. \tag{45}$$

A.2.2 Approximating $B(\lambda)$ when $\beta = 0$

Using Stirling's approximation and Assumptions (12) we have

$$B(\lambda) = \rho^S \frac{e^{S-\frac{\lambda p}{\mu}}}{\sqrt{2\pi S}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} \cdot \frac{1-\rho^{N-S-i}}{1-\rho} + o(1), \tag{46}$$

where $\rho = \frac{\lambda p}{S\mu}$. Under Condition (12)*(ii)*, $\beta = 0$ is possible when $\rho = 1$ or $\rho \to 1$. When $\rho \to 1$, $\rho \neq 1$, we use $\left(1-\rho^k\right)/(1-\rho) = k + o(1-\rho)$ that implies the following:

$$B(\lambda) = \frac{e^{S-\frac{\lambda p}{\mu}}}{\sqrt{2\pi S}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} (N-S-i) + o(1).$$

When $\rho = 1$, the sum $\sum_{j=0}^{N-S-i-1} \rho^j$ in (28) is equal to $N-S-i$ and this leads to the same expression for $B(\lambda)$. Simple calculations show that

$$B(\lambda) = \frac{1}{\sqrt{2\pi S}} \left((N-S-\frac{\lambda}{\theta}) \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} + e^{-\frac{\lambda}{\theta}} \frac{(\frac{\lambda}{\theta})^{N-S}}{(N-S-1)!} \right) + o(1).$$

It follows from Stirling's formula, Taylor's expansion of $\ln(\lambda/\theta)$ and Conditions (12) that

$$\frac{(\frac{\lambda}{\theta})^{N-S} e^{-\frac{\lambda}{\theta}}}{(N-S-1)!} = \sqrt{\frac{\lambda}{\theta}} \varphi(\eta) + o(1). \tag{47}$$

Using (12)*(ii)*, (43) and (47), we get that when $\beta = 0$,

$$\lim_{\lambda\to\infty} B(\lambda) = \lim_{\lambda\to\infty} \frac{1}{\sqrt{2\pi S}} \sqrt{\frac{\lambda}{\theta}} \left(\eta\Phi(\eta) + \varphi(\eta)\right) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{\mu}{p\theta}} \left(\eta\Phi(\eta) + \varphi(\eta)\right). \tag{48}$$

Combining (38), (44), (45) and (48), we have thus proved formula (13) in Theorem 3.1.

References

[1] Borst S., Mandelbaum A. and Reiman M.: Dimensioning Large Call Centers. Operations Research **52**(1), 17-34 (2004). 2, 11

[2] Brandt A., Brandt M., Spahl G. and Weber D.: Modelling and Optimization of Call Distribution Systems. Elsevier Science B.V. (1997). 2, 4

[3] Brown L., Gans N., Mandelbaum A., Sakov A., Zeltyn S., Zhao L. and Haipeng S.: Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. Journal of the American Statistical Association **100**, 36-50 (2005). 15

[4] Chen H. and Yao D.D.: Fundamentals of Queueing Networks. Springer-Verlag, New York (2001). 5

[5] Erlang A.K.: On the Rational Determination of the Number of Circuits. In "The Life and Works of A.K. Erlang." E. Brochmeyer, H.L. Halstrom and A. Jensen, eds. Copenhagen: The Copenhagen Telephone Company. 4.1.1, 4.2.1 (1948). 2

[6] Gans N., Koole G. and Mandelbaum A.: Telephone Call Centers: Tutorial, Review, and Research Prospects. Invited review paper by Manufacturing and Service Operations Management (MSOM) **5**(2), 79-141 (2003). 2

[7] Garnett O., Mandelbaum A. and Reiman M.: Designing a Call Center with Impatient Customers. Manufacturing and Service Operations Management (MSOM) **4**(3), 208-227 (2002). 2, 11

[8] Gilson K.A. and Khandelwal D.K.: Getting More from Call Centers. The McKinsey Quarterly, Web exclusive (2005). Available at `http://www.mckinseyquarterly.com/article_page.aspx` 1

[9] Donin O., Feigin P., Ishay E., Khudyakov P., Mandelbaum A., Nadjarov E., Trofimov V. and Zeltyn S.: DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 4.1: The Call Center of "US Bank" (2006). Available at `http://ie.technion.ac.il/Labs/Serveng/SEE_Documents_List.php` 14

[10] Halfin S. and Whitt W.: Heavy-Traffic Limits for Queues with Many Exponential Servers. Operations Research **29**, 567-587 (1981). 2, 6

[11] Jagerman D.L.: Some Properties of the Erlang Loss Function. Bell Systems Technical Journal **53**(3), 525-551 (1974). 2

[12] Jelenkovic P., Mandelbaum A. and Momcilovic P.: Heavy Traffic Limits for Queues with Many Deterministic Servers. Queueing Systems **47**, 53-69 (2004). 2

[13] Khudyakov P.: Designing a Call Center with an IVR (Interactive Voice Response). M.Sc. Thesis, Technion (2006). Available at `http://iew3.technion.ac.il/serveng/References/references.html` 11, 12

[14] Khudyakov P., Feigin P. and Mandelbaum A.: Internet Supplement to the Paper: Designing a Call Center with an IVR (Interactive Voice Response) (2010). Available at `http://iew3.technion.ac.il/serveng/References/references.html` 4, 5, 7, 9, 10, 11, 12, 13, 15, 17

[15] Liberman P., Trofimov V. and Mandelbaum A.: DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 5.1: Skills-Based-Routing-USBank (2008). Available at `http://iew3.technion.ac.il/serveng/References/` 16

[16] Maman S.: Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments. M.Sc. Thesis, Technion (2009). Available at `http://iew3.technion.ac.il/serveng/References/references.html` 15

[17] Mandelbaum A.: Lecture Notes on QED Queues. Available at `http://iew3.technion.ac.il/serveng/Lectures` 12

[18] Massey A.W. and Wallace B.R.: An Optimal Design of the M/M/C/K Queue for Call Centers. To appear in Queueing Systems (2006). 2, 6, 8

[19] Mandelbaum A. and Zeltyn S.: Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n+G Queue. Queueing Systems **51**, 361-402 (2005). Available at `http://iew3.technion.ac.il/serveng/References/references.html` 2, 10

[20] Pang R., Talreja R. and Whitt W.: Martingale Proofs of Many-Server Heavy-Traffic Limits for Markovian Queues. Probability Surveys **4**, 193-267 (2007). Available at `http://www.columbia.edu/~ww2040/PangTalrejaWhitt2007pub.pdf` 11

[21] Stolletz R.: Performance Analysis and Optimization of Inbound Call Centers. Springer-Verlag, Berlin Heidelberg (2003). 2

[22] Srinivasan R., Talim J. and Wang J.: Performance Analysis of a Call Center with Interacting Voice Response Units. TOP **12**, 91-110, (2004). 2, 4, 5

[23] Trofimov V., Feigin P., Mandelbaum A., Ishay E. and Nadjarov E.: DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 1: Model Description and Introduction to User Interface (2006). Available at `http://iew3.technion.ac.il/serveng/References/references.html` 14

[24] Weerasinghe A. and Mandelbaum A.: Abandonment vs. Blocking in Many-Server Queues: Asymptotic Optimality in the QED Regime. Working paper (2008). 13

[25] Whitt W.: Understanding the Efficiency of Multi-Server Service Systems. Management Science **38**, 708-723 (1992). 2