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

Khudyakov Polina, 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

June 1, 2009

Abstract

A call center is a popular term for a service operation that caters to customers' needs via the telephone. A call center typically consists 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 accomodates 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]**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 for his help, interest and valuable hints in data analysis.

Contents

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 effective management of operating costs; 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 the assistance of human agents. The IVR technology helps call centers to keep costs from rising (and sometimes to reduce costs), while hopefully improving service levels, revenue and hence profits.

A typical call center spends about 63% of its operational costs [22] on salaries. However, it would be over simplistic to reduce costs by decreasing the number of agents, because sometimes small change in the number of agents can effect dramatically the level of service. Thus, a main goal of a call center manager is to establish an appropriate tradeoff between cost and service level. Here queueing models come to the rescue, by yielding performance-analysis tools that support this tradeoff. Our paper 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, as surveyed by Gans, Koole and Mandelbaum [6]. A mathematical characterization of the QED regime for the GI/M/S queue was established by Halfin and Whitt [11] as having a non-trivial limit (within (0,1)) of the fraction of delayed customers, with S increasing indefinitely. The latter characterization was also established for GI/D/S [13], M/M/S with exponential patience [7] and with general patience [20].

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 [12]; see also Whitt [27], and then Massey and Wallace [19] for the analysis of finite buffers. The phenomenon of abandonment in a call center with multiple servers was analyzed by Garnet, Mandelbaum and Reiman [7] (Erlang-A model (M/M/S+M)) and Mandelbaum and Zeltyn [20] (M/M/S+G).

Erlang's characterization of the QED regime was in terms of the *square-root staffing principle* (50) (sometimes called "safety-staffing principle"). 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 in [19] for the M/M/S/N queue, in [7] for M/M/S+M, and others, 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 [23] 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 [23]. Our goal, in Section 3, is to find approximations for frequently used performance measures, which support decision-making for call center managers and help in analysis of the staffing problem. In Section 4, by fixing or taking to limits various parameters of our model, we specialize it to some classical queueing models. This provides rederivations and strengthening of their existing asymptotic analysis, notably the M/M/S/N queue in the QED regime. Then, in Section 5, we equip customers with finite patience. This gives rise to models with abandonment, for which we derive QED approximations as well. In Section 6, we validate our approximations against data from a real call center, thus establishing their applicability. We conclude, in Section 7, with calculating the (approximately) optimal number of trunk lines and number of servers, subject to constraints on system performance.

2 A model for a call center with an IVR

As mentioned already, a call center (see Figure 1) typically consists of telephone trunk lines, a switching machine known as the Automatic Call Distributor (ACD), an interactive voice response (IVR) unit, and agents to handle the incoming calls.

We consider the following model of a call center, as depicted in Figure 2: 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. If 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 finishing the IVR process, 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 abandonment in our model. (Abandonment will be incorporated in Section 5). 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 the call finds the system full, i.e. all N trunk lines are busy, it is lost.

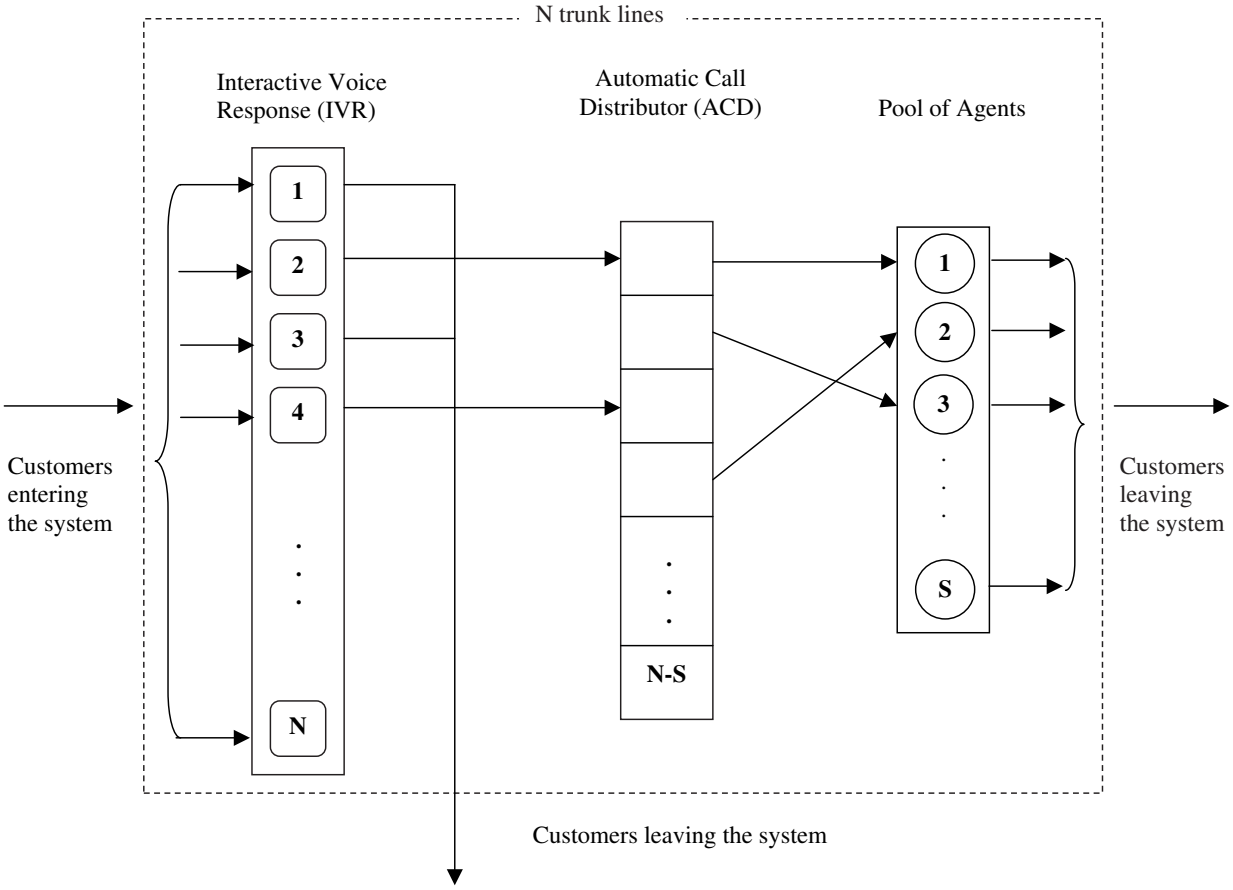

Figure 1: **Schematic model of a call center corresponding to a call center with an IVR, S agents and N trunk lines.**

We now view our model as a system with two multi-server queues connected in series (Figure 2). 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 in the system: at the IVR plus waiting to be served plus being served by the 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 $t \geq 0$. 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 $\{(i, j) \mid\ i + j \leq N,\ \ i, j \geq 0\}$.

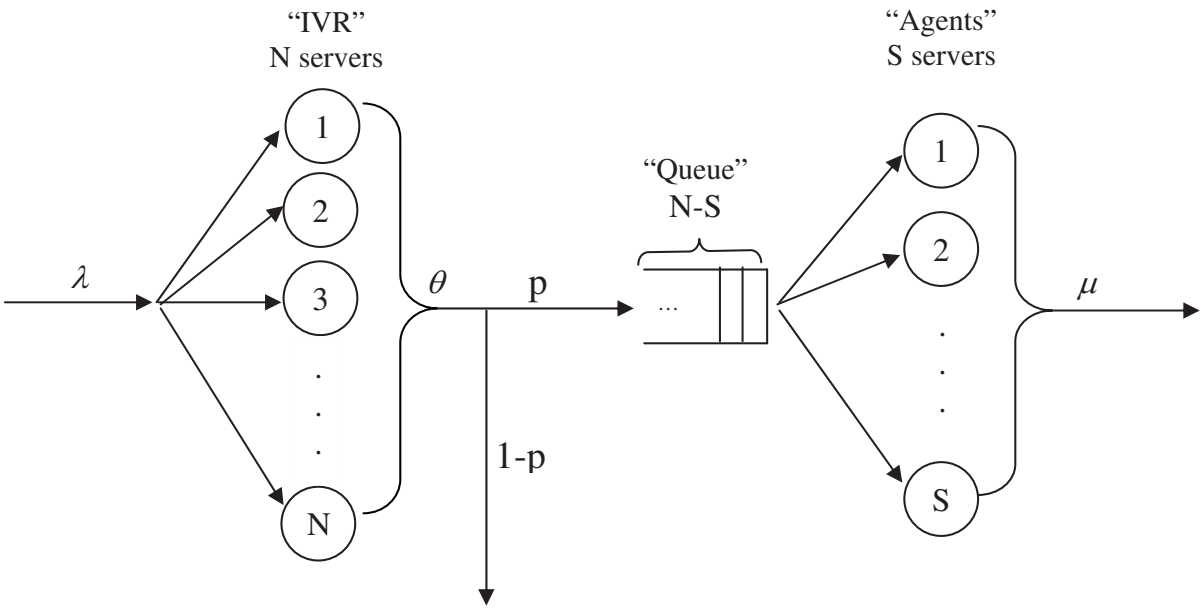

Figure 2: **Schematic model of a queueing system with an IVR, S agents and N trunk lines.**

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 numbers 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 3).

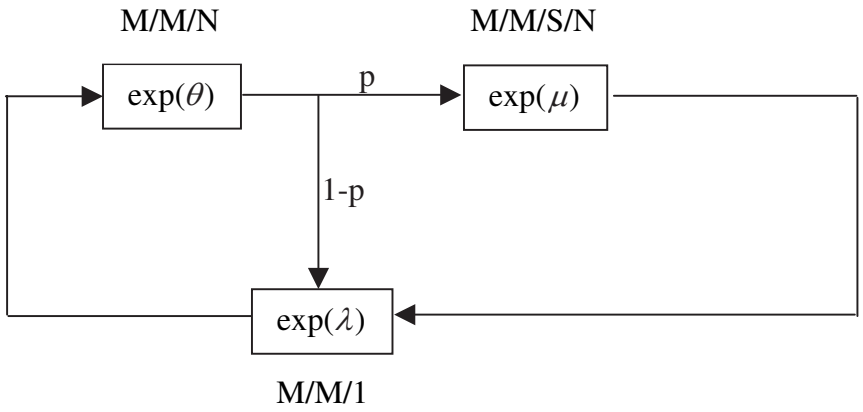

Figure 3: **Schematic model of a corresponding closed Jackson network.**

By normalization, we deduce the stationary probabilities $\pi(i, j)$ of having i calls at the IVR and j calls at the agents' station, which can be written in a normalized product form as follows:

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

where

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

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 finish the IVR process until they start service by an agent. We say that the system is in state $(\hat{k}, j)$, $0 \le j \le \hat{k} \le N$, when it contains exactly k calls, and j is the number of calls in the agents' station (waiting 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 found by Srinivasan, Talim and Wang in [23]; these are given by:

$$P(W \ge t) \triangleq 1 - \sum_{\hat{k}=S+1}^{N} \sum_{j=S}^{\hat{k}-1} \chi(\hat{k}, j) \sum_{l=0}^{j-S} \frac{(\mu S t)^l e^{-\mu S t}}{l!}. \tag{3}$$

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

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

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

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

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

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

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

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" [9]. 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

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

Another measure for the service level of a call center is the probability that an arriving call finds all trunk lines busy. It was found in [23] and has the following form:

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

In many cases, it is also interesting to know the *expected queue length* $E[L_q]$, which can be derived via Little's formula:

$$E[L_q] = \lambda_{eff} E[W_q] = p\lambda(1 - P(block)) \frac{E[W]}{P(W > 0)}. \tag{10}$$

The operating costs in call centers are mainly driven by the costs of the agents. Therefore, the *utilization* of the agents is often used as an operational measure to approximate economic (efficiency) performance. The expected utilization of the agents, say ρ_{eff}, is the ratio between the effective arrival rate λ_{eff} to the station of agents and the maximal service rate:

$$\rho_{eff} = \frac{\lambda_{eff}}{S\mu} = \frac{\lambda p(1 - P(block))}{S\mu}. \tag{11}$$

3 Asymptotic analysis in the QED regime

The ultimate goal of this section is to derive rules of thumb for solving the staffing and trunking problems for a call center with an IVR. This will be done analogously to Halfin and Whitt [11] and Massey and Wallace [19].

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 be overloaded, we assume that the number of agents S and the number of trunk lines N tend to infinity as well.

To motivate our asymptotic regime, we consider the model of a call center with an IVR as an extension of the M/M/S/N queue. This latter model was investigated by Massey and Wallace [19], who derived approximations of the performance measures of the M/M/S/N queue, when λ, S and N

tend to ∞ simultaneously and under the following assumptions:

$$\begin{aligned} &(i)\ \ N-S=\eta\sqrt{\frac{\lambda}{\mu}}+o\left(\sqrt{\lambda}\right),\ \ 0<\eta<\infty;\\ &(ii)\ \ S=\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}+o\left(\sqrt{\lambda}\right),\ \ 0<\beta<\infty; \end{aligned} \tag{12}$$

(In [19], β was assumed positive because of the use of the M/M/S queue in the analysis. We shall dispose of this assumption momentarily). For our call center with an IVR, following Assumptions (12), we need $S+\eta\sqrt{\frac{\lambda p}{\mu}}+o\left(\sqrt{\lambda}\right)$ trunk lines for the queue in the agents' station and $\frac{\lambda}{\theta}+\eta_2\sqrt{\frac{\lambda}{\theta}}+o\left(\sqrt{\lambda}\right)$ trunk lines for the IVR service. We can thus formulate the following conditions for our system. We let λ, S and N tend to ∞ simultaneously so that:

$$\begin{aligned} &(i)\ \ N-S=\eta_1\sqrt{\frac{\lambda p}{\mu}}+\frac{\lambda}{\theta}+\eta_2\sqrt{\frac{\lambda}{\theta}}+o\left(\sqrt{\lambda}\right),\ \ -\infty<\eta_1,\eta_2<\infty;\\ &(ii)\ \ S=\frac{\lambda p}{\mu}+\beta\sqrt{\frac{\lambda p}{\mu}}+o\left(\sqrt{\lambda}\right),\ \ -\infty<\beta<\infty; \end{aligned} \tag{13}$$

Note that we have three parameters η_1, η_2 and β, but one can reduce the number of parameters to two. Indeed, (13) is equivalent to

$$\begin{aligned} &(i)\ \ N-S=\eta\sqrt{\frac{\lambda}{\theta}}+\frac{\lambda}{\theta}+o\left(\sqrt{\lambda}\right),\ -\infty<\eta<\infty;\\ &(ii)\ \ S=\frac{\lambda p}{\mu}+\beta\sqrt{\frac{\lambda p}{\mu}}+o\left(\sqrt{\lambda}\right),\ \ -\infty<\beta<\infty; \end{aligned} \tag{14}$$

where $\eta=\eta_1\sqrt{\frac{p\theta}{\mu}}+\eta_2$.

In fact, as λ, S and N tend to ∞ simultaneous conditions (13) or (14) have also the following equivalent form

$$\begin{aligned} &(i)\ \ \lim_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta,\ \ -\infty<\eta<\infty;\\ &(ii)\ \ \lim_{\lambda\to\infty}\sqrt{S}(1-\frac{\lambda p}{\mu S})=\beta,\ \ -\infty<\beta<\infty; \end{aligned} \tag{15}$$

Conditions (15) constitute a square-root safety-staffing principle, which recommends the number of agents 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 (see [6] for details). Analogously, the number of lines, by rule (15), is the sum of the number of agents and offered load in the IVR with an addition of "safety" $\eta\sqrt{\frac{\lambda}{\theta}}$.

In order to avoid technical problems in calculation, it turns out convenient to distinguish two cases:

1) $\beta\neq 0$;

2) $\beta = 0$.

To this end, it will be convenient to divide (15) into two separate conditions:

$$QED: \begin{cases} (i) & \lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ (ii) & \lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty \ \ \beta \neq 0, \end{cases} \tag{16}$$

and

$$QED_0: \begin{cases} (i) & \lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ (ii) & \lim\limits_{\lambda\to\infty} \sqrt{S}(1-\frac{\lambda p}{\mu S}) = 0, \quad (\beta = 0). \end{cases} \tag{17}$$

3.2 QED approximations ($\beta \neq 0$)

First, assume that $\beta \neq 0$. In this case we prove the following theorem in Section 3.3.

Theorem 3.1. *[QED] Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions* (16)*, where μ, p, θ are fixed. Then the asymptotic behavior of our system, in Figure 2, is captured by the following performance measures:*

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

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

- *the probability of blocking:*

$$\lim_{\lambda\to\infty} \sqrt{S} P(block) = \frac{\nu + \beta\xi_2}{\gamma + \xi_1 - \xi_2},$$

- *the expectation of waiting time:*

$$\lim_{\lambda\to\infty} \sqrt{S} E[W] = \frac{\frac{1}{\mu\beta}\left(\xi_1 + (\beta^2 c^2 - \beta\eta c - 1)\xi_2\right)}{\gamma + \xi_1 - \xi_2},$$

- *the conditional density function of the waiting time, evaluated at $t/\sqrt{S}$:*

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \beta\mu e^{-\beta\mu t} \frac{\Phi(\eta - t\sqrt{p\mu\theta})}{\Phi(\eta) - e^{\eta_2}\Phi(\eta_1)};$$

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

$$\xi_1 = \frac{\varphi(\beta)\Phi(\eta)}{\beta}, \qquad \xi_2 = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\{\frac{\eta_1^2}{2}\}\Phi(\eta_1)}{\beta},$$

$$\gamma = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \qquad \nu = \frac{1}{\sqrt{1+c}}\varphi\left(\frac{\eta c + \beta}{\sqrt{1+c^2}}\right)\Phi\left(\frac{\beta c - \eta}{\sqrt{1+c^2}}\right),$$

$$\eta_1 = \eta - \beta/c, \qquad \eta_2 = \frac{1}{2}c^2\beta^2 - \eta\beta c, \qquad c = \sqrt{\frac{\mu}{p\theta}}. \tag{18}$$

3.3 Exact stationary performance

Our asymptotic analysis is based on representing the performance measures (3), (6), (7) and (9) in terms of building blocks. The asymptotic behavior of these blocks then determines that of the measures. Specifically, according to (3),[2] (6), (7), and (9), the operational characteristics for our model in Figure 2 can be represented as follows:

$$P(W>0)=\left(1+\frac{\gamma(\lambda)}{\xi_1(\lambda)-\xi_2(\lambda)}\right)^{-1}, \tag{19}$$

$$P(block)=\frac{\nu_1(\lambda)+\nu_2(\lambda)}{\gamma(\lambda)+\xi_1(\lambda)-\xi_2(\lambda)}, \tag{20}$$

$$E[W]=\frac{1}{\mu S}\frac{\zeta(\lambda)+\xi_1(\lambda)-\xi_2(\lambda)}{\gamma(\lambda)+\xi_1(\lambda)-\xi_2(\lambda)}, \tag{21}$$

$$f_{W|W>0}\,(t)=\frac{S(1-\rho)^2\mu e^{-S(1-\rho)\mu t}}{e^{\frac{\lambda p}{\mu}}\left(\frac{\lambda}{\theta}\right)^{-i}(\xi_1-\xi_2)}\cdot\delta(\lambda). \tag{22}$$

where

$$\xi_1(\lambda)=\frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{\lambda p}{\mu}\right)^S\frac{1}{1-\frac{\lambda p}{S\mu}}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i; \tag{23}$$

$$\xi_2(\lambda)=\frac{e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}}{S!}\left(\frac{\lambda p}{\mu}\right)^S\frac{1}{1-\rho}\rho^{N-S}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i,\qquad\left(\rho=\frac{\lambda p}{S\mu}\right), \tag{24}$$

$$\gamma(\lambda)=\sum_{i+j\le N-1,j<S}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}; \tag{25}$$

$$\nu_1(\lambda)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{j=0}^{S}\frac{1}{(N-j)!}\frac{1}{j!}\left(\frac{\lambda}{\theta}\right)^{N-j}\left(\frac{\lambda p}{\mu}\right)^j, \tag{26}$$

$$\nu_2(\lambda)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{j=S+1}^{N}\frac{1}{S!S^{j-S}}\left(\frac{\lambda p}{\mu}\right)^j\frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j}, \tag{27}$$

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

$$\delta(\lambda)=\sum_{k=0}^{N-S-1}\frac{\left[\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})\right]^k}{k!}e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})} \tag{29}$$

Theorem 3.1 is now a consequence of the following lemma:

[2]Equation (3) provides the density function of waiting time, but in order to derive its approximate formula we reduced it to (22) by applying Laplace transform (see [14] for details)

Lemma 3.1. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions* (16)*, where μ, p, θ are fixed. Then*

$$\lim_{\lambda\to\infty} \xi_1(\lambda) = \frac{\varphi(\beta)\Phi(\eta)}{\beta}, \tag{30}$$

$$\lim_{\lambda\to\infty} \xi_2(\lambda) = \frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\frac{\eta_1^2}{2}\Phi(\eta_1)}{\beta}, \tag{31}$$

$$\lim_{\lambda\to\infty} \gamma(\lambda) = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)c\right)\varphi(t)dt, \tag{32}$$

$$\lim_{\lambda\to\infty} \sqrt{S}\,\nu_1(\lambda) = \frac{1}{\sqrt{1+\frac{\mu}{p\theta}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}}+\beta}{\sqrt{1+\frac{\mu}{p\theta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{p\theta}{\mu}}-\eta}{\sqrt{1+\frac{\mu}{p\theta}}}\right) \tag{33}$$

$$\lim_{\lambda\to\infty} \sqrt{S}\,\nu_2(\lambda) = \varphi(\sqrt{\eta^2+\beta^2})e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1), \tag{34}$$

$$\lim_{\lambda\to\infty} \sqrt{S}\,\zeta(\lambda) = \frac{\varphi(\beta)\Phi(\eta)}{\beta^2} + \left(\beta^2\frac{\mu}{p\theta} - \eta\beta\sqrt{\frac{\mu}{p\theta}} - 1\right)\frac{\varphi(\sqrt{\eta^2+\beta^2})\exp\frac{\eta_1^2}{2}\Phi(\eta_1)}{\beta^2}, \tag{35}$$

$$\lim_{\lambda\to\infty} \delta(\lambda) = \Phi(\eta - t\sqrt{p\mu\theta}), \tag{36}$$

where $\eta_1 = \eta - c\beta$ and $c = \sqrt{\frac{\mu}{p\theta}}$, as defined in (18)

The proof of Lemma 3.1 is given in Part A.1 of the Appendix. Substituting the Lemma into equations (19)-(123) yields Theorem 3.1.

3.4 QED_0 approximations ($\beta = 0$)

In the case when $\beta = 0$, and in analogy to Theorem 3.1, we develop approximations for performance measures in the following theorem.

Theorem 3.2. *[QED_0] Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED_0 conditions* (17)*, where μ, p, θ are fixed. Then the asymptotic behavior of our system, in Figure 2, is captured by the following performance measures:*

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

$$\lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\gamma}{\xi}\right)^{-1},$$

- *the probability of blocking:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \frac{\nu}{\gamma+\xi},$$

- *the expectation of waiting time:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W] = \frac{1}{2\mu}\frac{\zeta}{\gamma+\xi},$$

- *the conditional density function of the waiting time, evaluated at $t/\sqrt{S}$:*

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{S}}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)=\frac{\mu\Phi(\eta-t\sqrt{p\mu\theta})}{\eta c\Phi(\eta)+c\varphi(\eta)}.$$

Here

$$\xi=\sqrt{\frac{1}{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right),\qquad \gamma=\int_{-\infty}^{0}\Phi\left(\eta-t\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt,$$

$$\nu=\frac{1}{\sqrt{1+c^2}}\varphi\left(\frac{\eta}{\sqrt{1+c^{-2}}}\right)\Phi\left(\frac{-\eta}{\sqrt{1+c^2}}\right)+\frac{1}{\sqrt{2\pi}}\Phi(\eta),$$

$$\zeta=\frac{1}{\sqrt{2\pi}}\left(\eta^2c^2\Phi(\eta)+\eta c^2\varphi(\eta)\right),$$

and $c=\sqrt{\frac{\mu}{p\theta}}$.

In the case when $\beta=0$, the operational characteristics in steady state have the following form:

$$P(W>0)=\left(1+\frac{\gamma(\lambda)}{\xi(\lambda)}\right)^{-1}, \tag{37}$$

$$P(block)=\frac{\nu_1(\lambda)+\nu_2(\lambda)}{\gamma(\lambda)+\xi(\lambda)}, \tag{38}$$

$$E[W]=\frac{1}{\mu S}\frac{\zeta(\lambda)+\xi(\lambda)}{\gamma(\lambda)+\xi(\lambda)}, \tag{39}$$

$$f_{W|W>0}(t)=\frac{\mu Se^{-\frac{\lambda p}{\mu}}}{\frac{1}{S!}\xi(\lambda)}\delta(\lambda), \tag{40}$$

where

$$\xi(\lambda)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{i=0}^{N-S-1}\sum_{j=S}^{N-i-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i\frac{1}{S!S^{j-S}}\left(\frac{\lambda p}{\mu}\right)^j. \tag{41}$$

$$\zeta(\lambda)=e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})}\sum_{m=0}^{N-S-1}\frac{1}{m!}\left(\frac{\lambda}{\theta}\right)^m\sum_{j=S}^{N-m-1}\frac{1}{S!S^{j-S}}\left(\frac{\lambda p}{\mu}\right)^j(j-S). \tag{42}$$

and $\gamma(\lambda)$, $\nu_1(\lambda)$, $\nu_2(\lambda)$ and $\delta(\lambda)$ are the same as in (25), (26) and (27), respectively. The representations in (37)-(40) differ from their analogues (19)-(22), in order to facilitate the analysis of the case $\beta=0$.

Analogously to the case when $\beta\neq 0$, for proving of Theorem 3.2 we rely on the following auxiliary lemma.

Lemma 3.2. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED_0 conditions* (17)*, where μ, p, θ are fixed. Then*

$$\lim_{\lambda\to\infty}\xi(\lambda)=\sqrt{\frac{1}{2\pi}}\frac{1}{c}\left(\eta\Phi(\eta)+\varphi(\eta)\right), \tag{43}$$

$$\lim_{\lambda\to\infty}\sqrt{S}\,\zeta(\lambda)=\frac{1}{\sqrt{2\pi}}\left(\eta^2\frac{\mu}{p\theta}\Phi(\eta)+\eta\frac{\mu}{p\theta}\varphi(\eta)\right). \tag{44}$$

where $c=\sqrt{\frac{p\theta}{\mu}}$*, as in* (18).

The proof of Lemma 3.2 is given in Part A.2 of the Appendix. Then combining equations (37)-(40) and the results of Lemma 3.2 yields Theorem 3.2.

4 Special cases

In this section, some special cases of our model are presented. In all these cases our model, under specific assumptions, becomes a well-analyzed model, such as Erlang-B, Erlang-C and others. The goal is to chart the boundary of our model and to show that this model, and the results obtained for it, coincide or extend well-known results for corresponding models.

4.1 The M/PH/S/S loss system

When the number of agents is equal to the number of trunk lines ($N = S$), the call center model can be presented as a special M/G/S/S loss system. There is no waiting in this model. The service time G has the Phase Type distribution in the following figure:

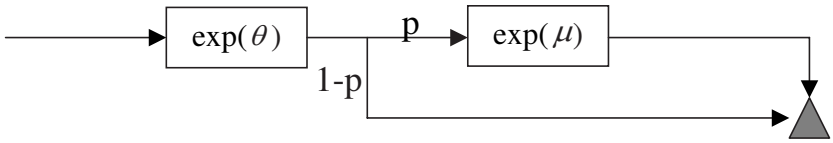

Figure 4: Schematic model of the Phase Type distribution that corresponds to the service time in a call center with an IVR, when the number of agents is equal to the number of trunk lines ($N = S$).

It is easy to see that when $p = 0$, i.e. no one wishes to be served by an agent, our system is the well-known M/M/N/N queue (Erlang-B model).

Similarly, when the service time of the agents goes to 0 (μ goes to infinity), the system is equivalent to the M/M/S/S loss system with exponential service time with rate θ. Only the IVR phase is taken into account. We have precisely the same picture if the service time in the IVR goes to 0 (θ goes to infinity). We still have the M/M/S/S loss system, but now the service time is exponential with the rate μ. In each case, by letting $\mu \to \infty$, or $\theta \to \infty$, the approximation for the loss probability agrees with the well-known asymptotic for the Erlang-B formula (Wolff [28]).

4.2 The M/M/S/N queue

Massey and Wallace [19] found 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 (12) prevails.

The condition $\eta > 0$ in (12) 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 [19] is their using the M/M/S queue for

finding the operational characteristics for M/M/S/N. Thus, in this section, we find approximations for the probability to wait and the probability to find the system busy for M/M/S/N, when $-\infty < \beta < \infty$. We also find approximations for the expected waiting time and the density of the waiting time, and show that, with some specific parameters, our system in Figure 2 can be represented as an M/M/S/N queue.

4.2.1 Operational characteristics for M/M/S/N

Theorem 4.1. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:* [3]

$$\begin{aligned} &(i) \quad \lim_{\lambda\to\infty} \frac{N-S}{\sqrt{\frac{\lambda}{\mu}}} = \eta, \quad 0 < \eta < \infty; \\ &(ii) \quad \lim_{\lambda\to\infty} \frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} = \beta, \quad -\infty < \beta < \infty; \end{aligned} \tag{45}$$

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 to wait:*

$$\lim_{\lambda\to\infty} P(W>0) = \begin{cases} \left(1 + \dfrac{\beta\Phi(\beta)}{\varphi(\beta)(1-e^{-\eta\beta})}\right)^{-1}, & \beta \neq 0, \\ \left(1 + \dfrac{\sqrt{\pi}}{\eta\sqrt{2}}\right)^{-1}, & \beta = 0; \end{cases} \tag{46}$$

- *the probability to find the system full:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \begin{cases} \dfrac{\beta\varphi(\beta)e^{-\eta\beta}}{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}, & \beta \neq 0; \\ \dfrac{1}{\sqrt{\dfrac{\pi}{2}}+\eta}, & \beta = 0; \end{cases} \tag{47}$$

- *the expectation of waiting time:*

[3] Note that these conditions can be also rewritten in the following form:

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

$$\lim_{\lambda\to\infty}\sqrt{S}E[W]=\begin{cases}\dfrac{\varphi(\beta)}{\mu}\left[\dfrac{1-e^{-\eta\beta}}{\beta}-\eta e^{-\eta\beta}\right], & \beta\neq 0;\\ \dfrac{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}{\beta\Phi(\beta)+\varphi(\beta)(1-e^{-\eta\beta})}, & \\ \dfrac{\eta^2}{2\mu(\eta+\sqrt{\frac{\pi}{2}})}, & \beta=0.\end{cases} \tag{48}$$

- *the density function of waiting time:*

$$\lim_{\lambda\to\infty}\frac{1}{\sqrt{S}}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)=\begin{cases}\dfrac{\mu\beta e^{-\mu\beta t}}{(1-e^{-\eta\beta})}, & \mu t<\eta, \quad \beta\neq 0;\\ \dfrac{\mu}{\eta}, & \mu t<\eta, \quad \beta=0;\\ 0, & \mu t\geq\eta.\end{cases} \tag{49}$$

The proof of this Theorem is provided in Part A.3 of the Appendix.

4.2.2 M/M/S/N as a special case of a call center with an IVR

Suppose that the IVR processing time is negligible. We capture this by letting $\theta/\lambda\to\infty$ (in particular, $\theta\to\infty$ since $\lambda\to\infty$). We also need to assume that all the customers wish to be served by an agent, i.e. $p=1$. In this case the system with an IVR can be presented as an M/M/S/N queue. By substitution of appropriative parameters into our approximation, it is easy to derive the corresponding approximations for M/M/S/N.

4.3 The M/M/S queue (Erlang-C)

Note that the M/M/S queue is an extreme case of M/M/S/N, which is obtained when $N\to\infty$, i.e. there are infinitely many places in the queue. Thus, to our previous assumption, that the IVR processing time is negligible when compared to the talk time of the agents, i.e. $\theta\to\infty$, we add that the number of trunk lines N tends to infinity, i.e. $\eta\to\infty$. It is easy to show that when $\beta>0$, that is

$$S=\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}+o(\sqrt{\lambda}), \tag{50}$$

the following holds:

- the approximation of the probability to wait has the form:

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}P(W>0)=\left(1+\frac{\Phi(\beta)\beta}{\varphi(\beta)}\right)^{-1}.$$

- the approximation of the conditional expectation of the waiting time is:

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}E[W|W>0]=\frac{1}{\mu\beta}.$$

• the approximation of the conditional density function of the waiting time is easy to obtain from (49), by letting $\eta \to \infty$:

$$\lim_{\eta\to\infty}\lim_{\lambda\to\infty}\sqrt{S}f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right)=\beta\mu e^{-\beta\mu t}.$$

These results coincide with the approximations of Halfin and Whitt [11] for M/M/S.

4.4 The M/M/S/∞/N queue

The $M/M/S/\infty/N$ system was presented and analyzed by de Véricourt and Jennings in [25]. This system is a particular case of our model, when $\lambda \to \infty$ more quickly then S ($\lambda = o(S)$). The M/M/S/∞/N system it one where there are exactly N customers in the system, which means that it is impossible to leave the system after service in the IVR, i.e. $p = 1$, and a customer that leaves the system after agent's service is instantaneously replaced by a new customer (or, alternatively, returns to the IVR).

This model in [25], which is actually the classical machine breakdown model, was used for describing a hospital internal ward.

In order to reduce our model to $M/M/S/\infty/N$, let us define the states of our system when there are exactly N customers in the system and no possibility to leave it after the IVR. The states will take the following form: $(N-j, j)$, where $0 \le j \le N-S$. Thus, the stationary probabilities are the following:

$$\pi(N-j,j)=\begin{cases}\pi_0\dfrac{1}{(N-j)!}\left(\dfrac{\lambda}{\theta}\right)^{N-j}\dfrac{1}{j!}\left(\dfrac{\lambda}{\mu}\right)^{j}, & j\le S;\\ \pi_0\dfrac{1}{(N-j)!}\left(\dfrac{\lambda}{\theta}\right)^{N-j}\dfrac{1}{S!S^{j-S}}\left(\dfrac{\lambda}{\mu}\right)^{j} & j\ge S;\\ 0 & \text{otherwise,}\end{cases} \tag{51}$$

where

$$\pi_0=\left(\sum_{j=S}^{N}\frac{1}{N!}\left(\frac{\lambda}{\theta}\right)^{N-j}\frac{1}{S!S^{j-S}}\left(\frac{\lambda}{\mu}\right)^{j}+\sum_{j=0}^{S}\frac{1}{(N-j)!}\left(\frac{\lambda}{\theta}\right)^{N-j}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^{j}\right)^{-1}. \tag{52}$$

We can also write the stationary probabilities in the equivalent form:

$$\pi(N-j,j)=\begin{cases}\widetilde{\pi}_0\begin{pmatrix}N\\ j\end{pmatrix}\left(\dfrac{1}{\theta}\right)^{N-j}\left(\dfrac{1}{\mu}\right)^{j}, & j\le S;\\ \widetilde{\pi}_0\begin{pmatrix}N\\ j\end{pmatrix}\left(\dfrac{1}{\theta}\right)^{N-j}\dfrac{j!}{S!S^{j-S}}\left(\dfrac{1}{\mu}\right)^{j} & j\ge S;\\ 0 & \text{otherwise,}\end{cases} \tag{53}$$

where

$$\widetilde{\pi}_0=\left(\left(\frac{1}{\theta}+\frac{1}{\mu}\right)^{N-j}+\sum_{j=S+1}^{N}\frac{N!}{(N-j)!}\left(\frac{1}{S!S^{j-S}}-\frac{1}{j!}\right)\left(\frac{1}{\theta}\right)^{N-j}\left(\frac{1}{\mu}\right)^{j}\right)^{-1}. \tag{54}$$

The equations (53) and (54) have the same form as the stationary probabilities in [25], and thus all the exact results for $M/M/S/\infty/N$ are contained in this particular case of our model.

Note, however, that our asymptotic analysis does not cover that in [25] since, in our case, $\lambda \to \infty$ together with the parameters S and N. In contrast, to obtain the limit in [25] from our results, one must first take $\lambda \to \infty$; this results in a closed system with N customers and S servers which is now approximated by increasing N and S via QED scaling.

5 Adding abandonment

5.1 Model description

In this section, we add the feature of customers' patience, which could lead to their abandonment from the queue prior to service. The modelling assumptions are the same as those in Section 2 (see figure 3), which are characterized by the parameters (λ, N, θ, p, S, μ). In addition, if a call waits in the queue, it may leave the system after an exponentially distributed time with rate δ (impatience), or it answered by an agent, whatever happens first.

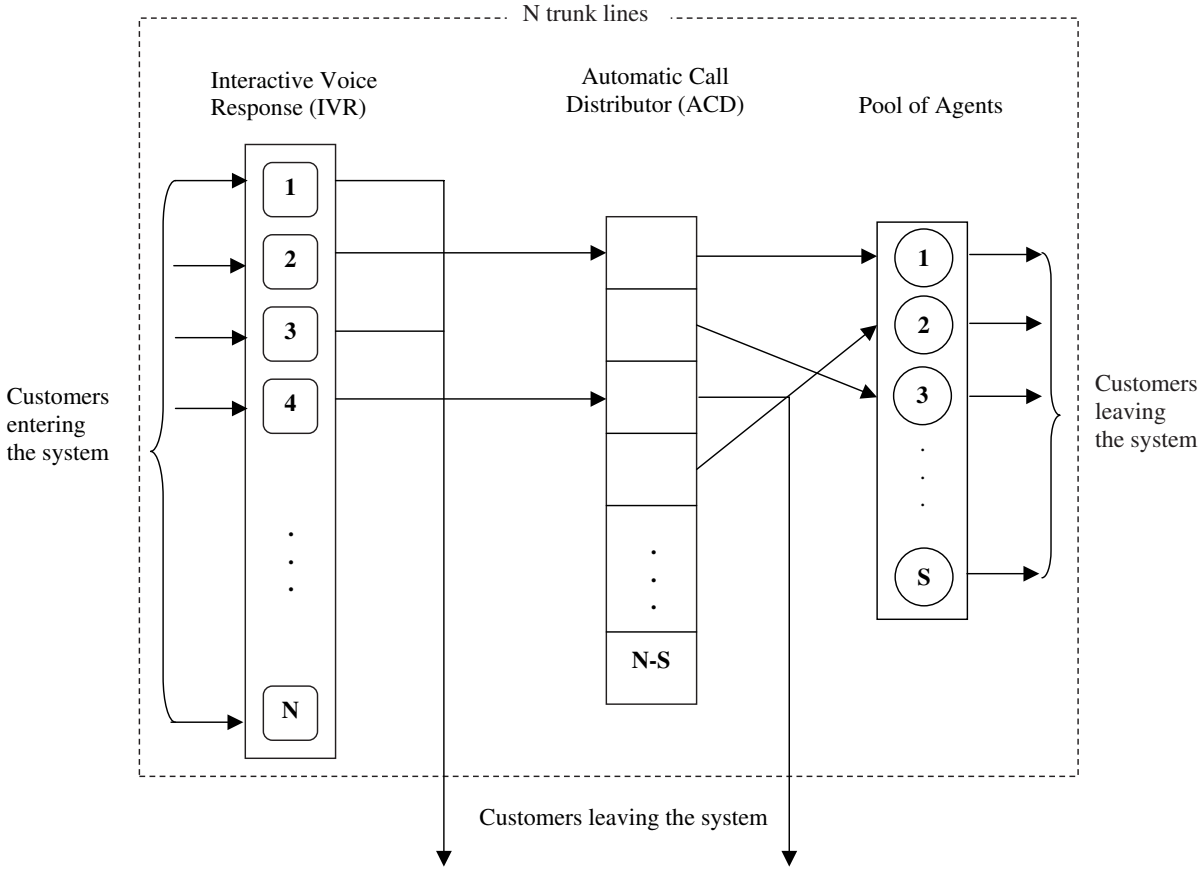

Figure 5: **Schematic model of a call center with an IVR, S agents, N trunk lines and customers' abandonment.**

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 3 is replaced by M/M/S/N+M, with impatience that is distributed $exp(\delta)$, as in Figure 6.

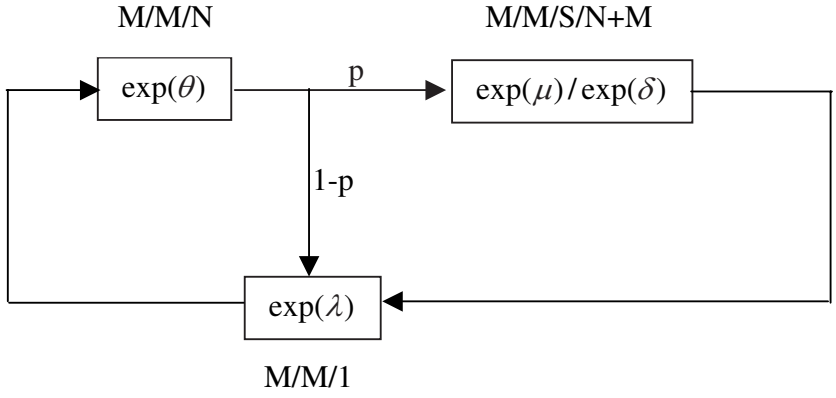

Figure 6: **Schematic model of a call center with an interactive voice response, S agents and N trunk lines.**

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 as follows:

$$\pi(i,j) = \begin{cases} \pi_0 \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j, & j \le S,\ 0 \le i+j \le N; \\ \pi_0 \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!}\left(\frac{\lambda p}{\mu}\right)^S \frac{(\lambda p)^{j-S}}{\prod_{k=1}^{j}(S\mu + k\delta)} & j \ge S,\ 0 \le i+j \le N; \\ 0 & \text{otherwise,} \end{cases} \tag{55}$$

where

$$\pi_0 = \left(\sum_{i=0}^{N-S} \sum_{j=S+1}^{N-i} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{S!}\left(\frac{\lambda p}{\mu}\right)^S \frac{(\lambda p)^{j-S}}{\prod_{k=1}^{j}(S\mu + k\delta)} + \sum_{i+j \le N, j<S} \frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i \frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j \right)^{-1}. \tag{56}$$

5.2 A call center with abandonment and an IVR

As previously, our goal is to find 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 a special treatment, because we do not have to divide by β as previously (see Theorem 3.1). Therefore, we consider only one version of our asymptotic analysis domain, which we refer to as the following QED condition:

$$QED: \begin{cases} (i) & \lim\limits_{\lambda\to\infty} \frac{N-S-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} = \eta, \quad -\infty < \eta < \infty; \\ (ii) & \lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda p}{\mu S}) = \beta, \quad -\infty < \beta < \infty. \end{cases} \tag{57}$$

Analogously to the calculations in Section 3.2, we now introduce the approximations for performance measures with abandonments.

Theorem 5.1. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the QED conditions (57), 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 will wait after the IVR process:*

$$\lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\gamma}{\xi_1 - \xi_2}\right)^{-1}, \tag{58}$$

- *the probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(Ab|W > 0) = \frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta) - \beta \int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}{\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}, \tag{59}$$

- *the expectation of waiting time, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W|W > 0] = \frac{1}{\delta}\frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta) - \beta \int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}{\int\limits_{\beta\sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta\sqrt{\frac{\mu}{\delta}} - t)\sqrt{\frac{p\theta}{\mu}})\varphi(t)dt}, \tag{60}$$

- *the probability of blocking:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(block) = \frac{\nu + \xi_2\varphi\left(\frac{\eta + \beta\sqrt{\frac{p\mu\theta}{\delta}}}{\sqrt{1 + \frac{p\theta}{\delta}}}\right)/(1 - \Phi(\beta\sqrt{\frac{\mu}{\delta}}))}{\gamma + \xi_1 - \xi_2}; \tag{61}$$

in the above,

$$\xi_1 = \sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})} \int\limits_{-\infty}^{\eta} \Phi\left((\eta - t)\sqrt{\frac{\delta}{p\theta}} + \beta\sqrt{\frac{\mu}{\delta}}\right)\varphi(t)dt, \quad \xi_2 = \sqrt{\frac{\mu}{\delta}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\Phi(\beta\sqrt{\frac{\mu}{\delta}})\Phi(\eta),$$

$$\gamma = \int\limits_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \quad \nu = \frac{1}{\sqrt{1 + \sqrt{\frac{\mu}{p\theta}}}}\varphi\left(\frac{\eta\sqrt{\frac{\mu}{p\theta}} + \beta}{\sqrt{1 + \frac{\mu}{p\theta}}}\right)\Phi\left(\frac{\beta\sqrt{\frac{\mu}{p\theta}} - \eta}{\sqrt{1 + \frac{\mu}{p\theta}}}\right).$$

5.3 The M/M/S/N+M queue

We now present approximations for performance characteristics of the M/M/S/N+M queue. As in [21], we apply the approximations under the following conditions:

$$QED: \begin{cases} (i) & S = R + \beta\sqrt{R} + o(\sqrt{R}), \quad -\infty < \beta < \infty; \\ (ii) & N = S + \eta\sqrt{S} + o(\sqrt{S}), \quad \eta \geq 0, \end{cases} \tag{62}$$

where $R = \frac{\lambda}{\mu}$. The results are formalized in the following theorem.

Theorem 5.2. *Let the variables λ, S and N tend to ∞ simultaneously and satisfy the following conditions:*

$$\begin{cases} (i) & \lim\limits_{\lambda\to\infty} \sqrt{S}(1 - \frac{\lambda}{\mu S}) = \beta, \quad -\infty < \beta < \infty; \\ (ii) & \lim\limits_{\lambda\to\infty} \frac{N-S}{\sqrt{S}} = \eta, \quad \eta \geq 0, \end{cases}$$

where μ is fixed. [4] *Then the asymptotic behavior of the system is described in terms of the following performance measures:*

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

$$\lim_{\lambda\to\infty} P(W > 0) = \left(1 + \frac{\sqrt{\frac{\mu}{\delta}}\Phi(\beta)\varphi(\beta\sqrt{\frac{\mu}{\delta}})}{\varphi(\beta)\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]} \right)^{-1}, \tag{63}$$

- *the probability of abandonment, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}P(Ab|W > 0) = \frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}}) - \beta\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]}{\Phi(\eta\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})}, \tag{64}$$

- *the expectation of the waiting time, given waiting:*

$$\lim_{\lambda\to\infty} \sqrt{S}E[W|W > 0] = \frac{1}{\delta}\frac{\sqrt{\frac{\delta}{\mu}}\varphi(\beta\sqrt{\frac{\mu}{\delta}}) - \beta\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]}{\Phi(\eta\sqrt{\frac{\delta}{\mu}} + \beta\sqrt{\frac{\mu}{\delta}}) - \Phi(\beta\sqrt{\frac{\mu}{\delta}})}, \tag{65}$$

[4] When $\eta = 0$, the M/M/S/N+M queue is equivalent to the M/M/S/S loss system. In this case $P(Ab|W > 0)$ and $E[W]$ are equal to 0 and their approximations are not relevant.

- *the probability of blocking:*

$$\lim_{\lambda\to\infty}\sqrt{S}P(block)=\frac{\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\varphi(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}})}{\Phi(\beta)+\sqrt{\frac{\delta}{\mu}}\frac{\varphi(\beta)}{\varphi(\beta\sqrt{\frac{\mu}{\delta}})}\left[\Phi(\eta\sqrt{\frac{\delta}{\mu}}+\beta\sqrt{\frac{\mu}{\delta}})-\Phi(\beta\sqrt{\frac{\mu}{\delta}})\right]}. \tag{66}$$

The proofs of Theorems 5.1 and 5.2 are analogous to those of Theorems 3.1 and 3.2, and are carried out by the using of formulae (3)-(7), where the stationary probabilities are defined by (55) and (56).

6 Accuracy of the approximations

We now compare the exact formulae with their approximations, via graphs that include both. First, the real values were calculated by a program written in Visual Basic, and the approximations' values were calculated in Maple. Next, all this data was processed in Excel and the graphs were created.

6.1 Approximations of the model with an IVR and abandonment

In Khudyakov [14], Chapter 5, we demonstrated the accuracy of our approximations for a model *without* abandonment. These approximations turn out extremely accurate, over a very wide range of parameters (S already from 10 and above, $N \geq 50$). Here, we have chosen to present approximations that accommodate abandonments. The numerical analysis is heavier due to the increased prevalence of integral-approximations. For example, the approximation of $P(W > 0)$ involves an integral in both γ and ξ (as opposed to only γ previously, in the model without abandonment). In addition, for calculations of exact values we are restricted to smaller N's ($N \leq 80$ here, as opposed to $N \leq 170$ before).

To test our approximations, we compare performance measures of a model with an IVR and abandonment that corresponds to a mid-sized call center that has the arrival rate λ of 30 customers per minute. The number of agents S is in the domain where the traffic intensity $\rho = \frac{\lambda p}{\mu S}$ is about 1 (namely, the number of agents is between 20 and 40, i.e. $S \approx 30 \pm 2 \cdot \sqrt{30}$). For simplicity, we let $p = \mu = \theta = \delta = 1$. The number of trunk lines is 80.

For each value of the number of agent S, we calculate the parameters η and β by using (57).

Figures 7 and 8 depict the comparison of the exact calculated probability to wait and the conditional probability to abandon with their approximations. The approximations are clearly close to the exact values.

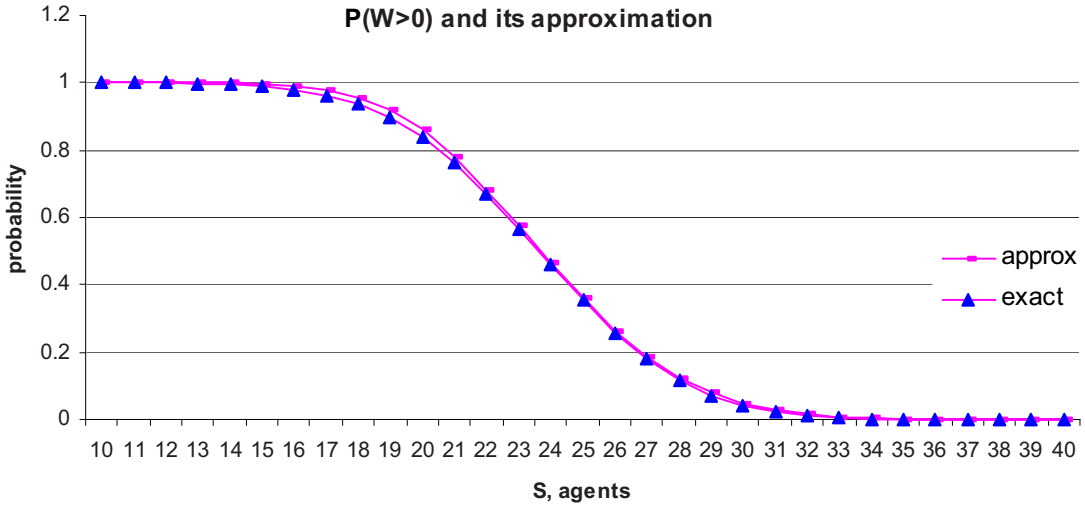

Figure 7: Comparison of the exact calculated probability to wait and its approximation (58) for a mid-sized call center with arrival rate 30 and 80 trunk lines.

Note, that

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

Thus, it is natural that the approximation for $E[W]$ will also be close to the exact calculated average of the waiting time.

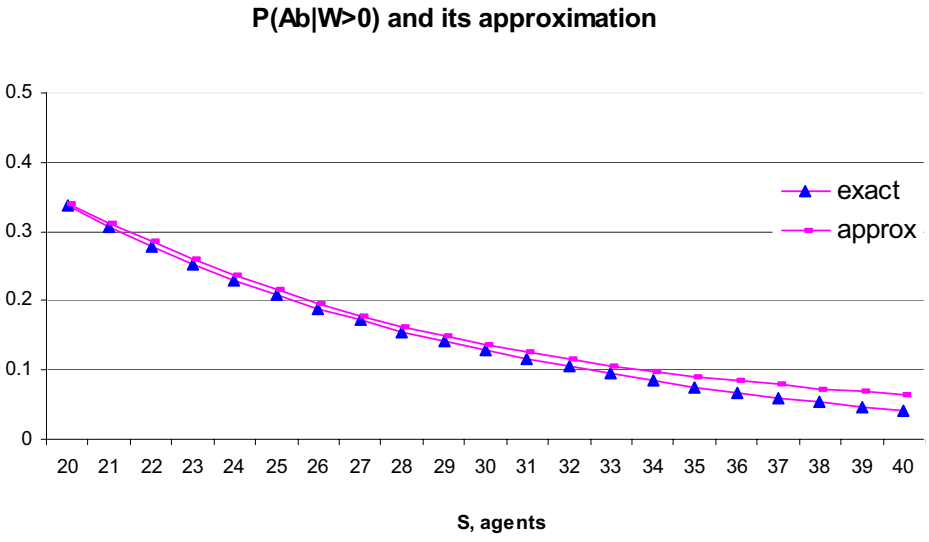

Figure 8: Comparison of the exact calculated probability of abandonment, given waiting, and its approximation (59), for a mid-sized call center with arrival rate 30 and 80 trunk lines.

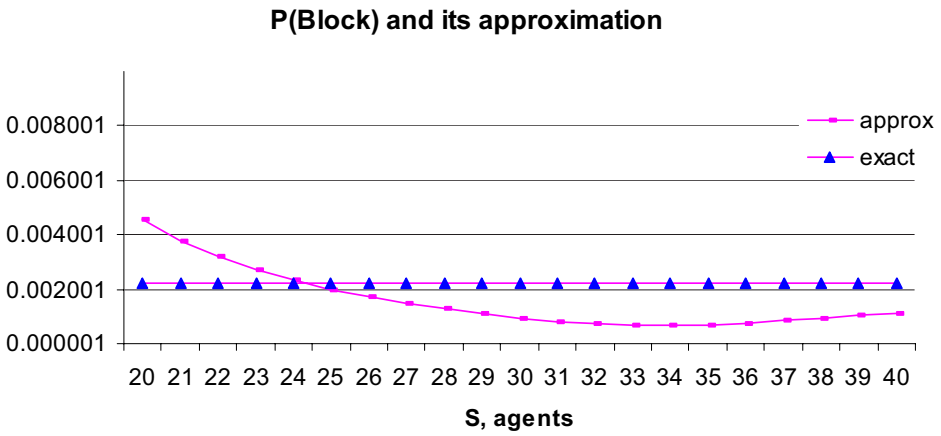

Figure 9: Comparison of the exact calculated probability to find all trunks busy and its approximation (61), for a mid-sized call center with arrival rate 30 and 80 trunk lines.

Figure 9 shows that the approximation of the probability to find the system busy is, relatively speak not as accurate as previous measures. On the other hand, this figure has a very high resolution and the differences between the exact and approximate probabilities are less than 0.002. One can thus argue that our approximation for the probability to find all trunks busy also works well.

6.2 Problematic domain

It is important to note that sometime the approximations are not so accurate. As an example, let us consider a mid-sized call center with the same parameters as previously, but let the number of trunk lines be equals to 50 (instead of 80).

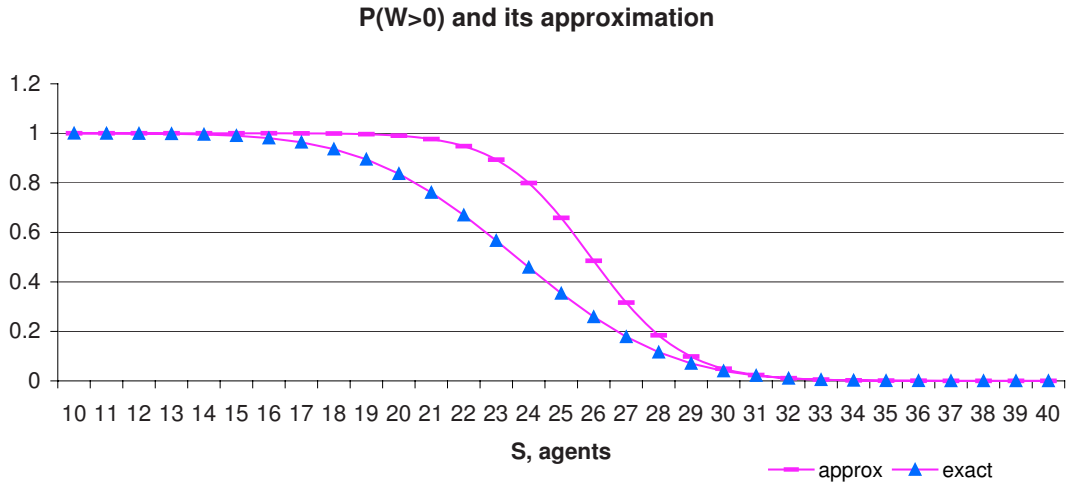

Figure 10: Comparison of the exact calculated probability to find all trunks busy and its approximation, for a mid-sized call center with arrival rate 30 and 50 trunk lines.

Figure 10 shows that the deviations between the exact and approximate probability can reach 0.4, or about 100% error. This happens in the area that corresponds to heavy traffic (QED) regime, but with a small number of trunk lines. The reason for such an inaccuracy in the approximations is the following: the approximations formulae consist of $\Phi(\cdot)$, the standard normal distribution, and $\varphi(\cdot)$, the standard normal density functions, which depend on the parameters η and β. In the considered area, these parameters are negative and large in value ($\eta < 0$ and $\beta < 0$). Consequently, the functions $\Phi(\cdot)$ and $\varphi(\cdot)$ give rise to very small values, which influence numerical accuracy. In practice, however, modern technology allows one to operate as many trunk lines as needed. Therefore, the considered ranges of S and N ($10 \leq S \leq 40$, $N = 50$) are not natural for prevalent operations of a call center.

6.3 Approximations for the M/M/S/N+M queue

Examining the approximations for performance measures of the M/M/S/N+M queue, we model a mid-sized call center, in which the arrival rate λ is 100 customers per minute. The number of agents S is in the domain where the traffic intensity $\rho = \frac{\lambda p}{\mu S}$ is about 1 (namely, the number of agents is between 80 and 120). As before, we let $p = \mu = \theta = \delta = 1$. The number of trunk lines is mostly 150, but when we check the probability of blocking we take the number of trunk lines to be 120, this in order to avoid very small values. (Note that here we are able to cover values of N and S that are larger than those considered in section 6.1. The formulae here are simpler and hence enable their numerical analysis.)

As previously, for each value of the number of agent S, we calculate the parameters η and β from the conditions in Theorem 5.2.

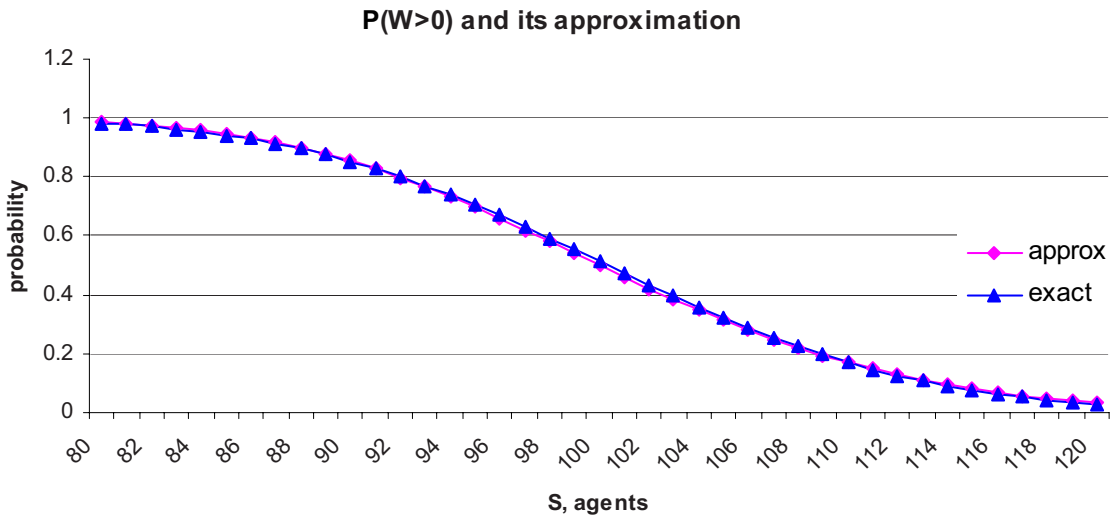

Figure 11: Comparison of the exact calculated probability to wait and its approximation, for a mid-sized call center with arrival rate 100 and 150 trunk lines.

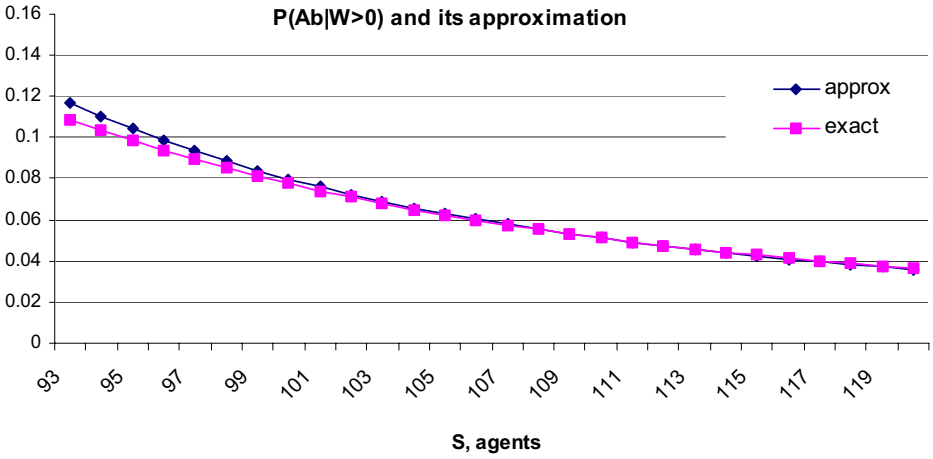

Figure 12: Comparison of the exact calculated probability of abandonment, given waiting, and its approximation, for a mid-sized call center with arrival rate 100 and 150 trunk lines.

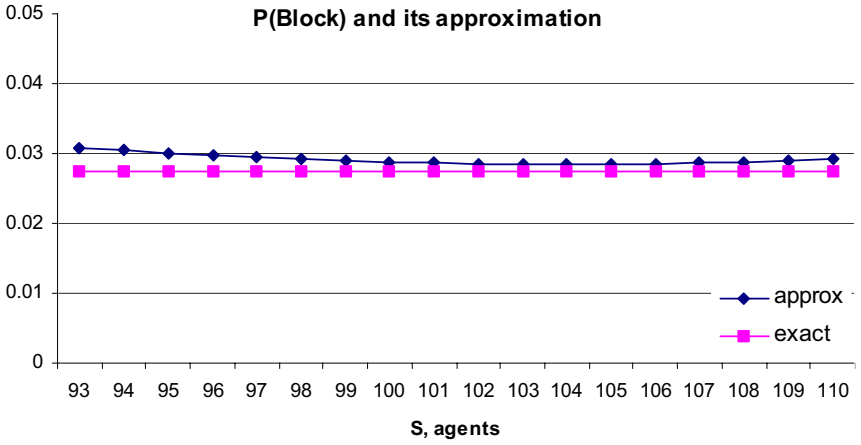

Figure 13: Comparison of the exact calculated probability to find the system busy and its approximation, for a mid-sized call center with arrival rate 100 and 120 trunk lines.

7 Rules of thumb

We derived approximations for performance measures in the QED regime (Quality and Efficiency Driven), as characterized by conditions (16) or (17). The detailed comparison in [14], between exact vs. approximated performance, shows that the approximations often work perfectly, even *outside* the QED regime. In this section, we attempt to chart the boundary of this "outside", thriving to summarize our findings through practical rules-of-thumb (expressed via the offered load $R = \frac{\lambda p}{\mu}$). These rules of thumb were derived via extensive numerical analysis (using Maple) of our analytical

results; for an elaboration, readers are referred to [14].

7.1 Operational regimes.

As customary, one distinguishes three types of staffing regimes:

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

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

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

We shall use the characterization of the operational regimes, as formulated in [18] and presented in Table 1, in order to specify numerical ranges for the parameters β and η, in the $M/M/S/N$ queue and in the model with an IVR with and without abandonment. Specifying β corresponds to determining a staffing level, and specifying η to determining the number of trunk lines.

	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

Table 1: Rules-of thumb for operational regimes.

In Table 1, AST stands for Average Service Time.

7.2 System parameters

The performance measures of a call center with an IVR, *without* abandonment, depends on β, η, $\frac{p\theta}{\mu}$ and S; in particular, large values of $\frac{p\theta}{\mu}$ and S improve performance (See [14] 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.

One must thus take into account 5 system's parameters. In order to reduce the dimension of this problem, we fix some parameter, at values that correspond to a realistic call center, based on our experience (see [24]):

IVR service time equals, on average, 1 minute;

Agents' service time equals, on average, 3 minutes;

Customers' patience, on average, takes values between 3 and 10 minutes;

Fraction of customers requesting agents' service, in addition to the IVR, equals 30%;

Offered load equals 200 Erlangs (200 minutes per minute).

Our goal is to identify the parameter values for η (determines the number of trunk-lines) and β (determines the number of agents) that ensure QED performance as described in Table 1, while simultaneously estimating the value of the probability of blocking in each case (which does not appear in Table 1).

7.3 QED regime in the M/M/S/N and M/M/S/N+M queues

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. Table 2 shows that, when $\eta > 3$, the $M/M/S/N$ queue behaves as the $M/M/S$ queue (negligible blocking).

$S \approx R + \beta\sqrt{R}$ $N \approx S + \eta\sqrt{S}$	**M/M/S/N**	**M/M/S/N+M**
$0.5 \le \eta < 1.5$	$-1.5 < \beta < 0.5$	$-1.6 < \beta < 0.4$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.02, & \beta \ge 0; \end{cases}$	$\begin{cases} < -\beta/\sqrt{S}, & \beta < 0, \\ < 0.05, & \beta \ge 0; \end{cases}$
$1.5 \le \eta < 3$	$-0.5 < \beta < 0.8$	$-0.8 < \beta < 0.6$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.01, & \beta \ge 0; \end{cases}$	$\begin{cases} < 0.02, & \beta < 0 \\ \approx 0, & \beta \ge 0; \end{cases}$
$\eta > 3$	$\beta > 0$	$-0.5 < \beta < 0.8$
P(block)	≈ 0	≈ 0

Table 2: Rules-of thumb for the QED regime in $M/M/S/N$ and $M/M/S/N + M$.

The values presented in Table 2, for the $M/M/S/N + M$ queue, and in Table 3 for the system with an IVR, were calculated under the assumption that average customer patience equals 3 minutes (same as the average service time). As already noted, in practice this value can get much larger, but performance is rather insensitive to average patience until values of about 15 minutes. For average patience above 15 minutes, performance gets similar to the corresponding model without abandonment.

7.4 QED regime for a call center with an IVR with and without abandonment

In the case of the system with an IVR, there are no mathematical restrictions for η to be non-negative, but we propose $\eta \ge 0$ because otherwise ($\eta < 0$), the probability of blocking is higher then 0.1. We

believe that a call center can not 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: as noted in Table 2, $\eta > 3$ suffices.

Table 3 shows that, sometimes, one can reduce the number of trunk lines in order to improve service level. For instance, starting with $\eta > 3$ and the number of agents corresponding to $\beta = -0.8$ (ED performance), we can achieve QED performance by reducing the number of trunk lines via $\eta = 2$; in that way, we loose on waiting time and abandonment while the probability of blocking is still less than 0.01. Moreover, modern technology enables a message that replaces a busy-signal, with a suggestion to leave one's telephone number, in order to be called back latter; alternatively, a blocked call can be routed to an outsorsing alternative. Thus, we are not necessary loosing these "blocked" customers. See [15] and [26] for an analysis where the asymptotically optimal number of trunk-lines is determined.

$S \approx R + \beta\sqrt{R}$ $N \approx S + \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}$	**IVR without abandonment**	**IVR with abandonment**
$0 \le \eta < 1$	$-1.2 < \beta < 0.2$	$-1.6 < \beta < 0$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.04, & \beta \ge 0; \end{cases}$	< 0.08
$1 \le \eta < 2$	$-0.7 < \beta < 0.5$	$-1.2 < \beta < 0.4$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.03, & \beta \ge 0; \end{cases}$	< 0.04
$2 \le \eta < 3$	$-0.3 < \beta < 0.7$	$-0.8 < \beta < 0.6$
P(block)	$\begin{cases} -\beta/\sqrt{S}, & \beta < 0, \\ < 0.02, & \beta \ge 0; \end{cases}$	< 0.01
$\eta > 3$	$\beta > 0$	$-0.6 < \beta < 0.8$
P(block)	≈ 0	≈ 0

Table 3: Rules-of thumb for the QED regime in a call center with an IVR with and without abandonment.

According to Table 3, when $\eta > 3$, the system with an IVR behaves as one with an infinite number of trunk lines.

7.5 QD and ED regimes

For the QD and ED regimes (see Table 1), the number of agents can be specified via $0.1 \le \gamma \le 0.25$. In the case of QD, the number of agents is over-staffing; limiting the number of trunk lines will cause unreasonable levels of agents' idleness, hence $\eta \ge 3$ makes sense. In the case of ED, the number of agents is under-staffing, and we are interested in reducing the system's offered load. Therefore, we propose to take $\eta = 2$. This choice yields a probability of blocking to be approximately $\gamma/2$ (based on

numerical experience).

7.6 Conclusions

Our rules of thumb demonstrate that for providing services 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 Section 7.4, 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 signal. Hence, such a solution must result from a tradeoff between the probability of blocking and the probability to abandon.

8 Model validation with real data

The approximations that have been developed 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 analyze approximations of a real call center by models with and without an IVR, starting with the model without (M/M/S/N+M) in order to later evaluate the value of adding an IVR. This evaluation is the goal of our empirical study, which is based on analyzing real data from a large call center. (The size of our call center, around 600-700 agents, forces one to use our approximations, as opposed to exact calculations which are numerically prohibitive.)

8.1 Data description

The data for the current analysis come from a call center of a large U.S. bank - it will be referred to as the US Bank Call Center in the sequel. The full database archives all the calls handled by the call center over the period of 30 months from March 2001 until September 2003[5]. The call center consists of four different contact centers (nodes), which are connected using high technology switches so that, in effect, they can be considered as 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 choose 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 one of the four call centers and 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, the customer waits in the queue; otherwise, he or she is 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 an agent with an appropriate skill from any of the four nodes. After service by an agent, customers may 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.

A schematic model of our US Bank Call Center is presented in Figure 14:

[5]The data is available at `http://seeserver.iem.technion.ac.il/see-terminal/`.

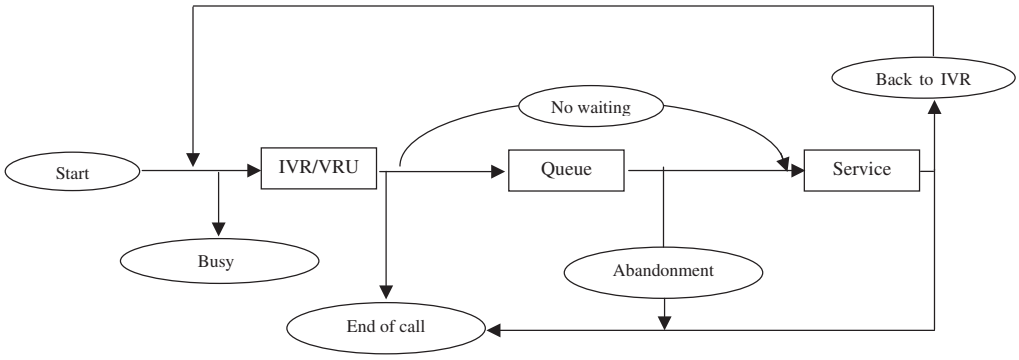

Figure 14: Schematic diagram of the call of a "Retail" customer in our US Bank call center.

8.2 Fitting the theoretical model to a real system

Figure 14 describes the flow of a call through our call center. It differs somewhat from the models described in Sections 2 and 5. 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. The possibility that queued customers abandon (hang up) without being answered is not acknowledge in our model from Section 2. We thus compare only the models from section 5 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. Figures 15 and 16, produced using the SEEStat program [24], display the distribution of service time in the IVR and agents' service time, respectively.

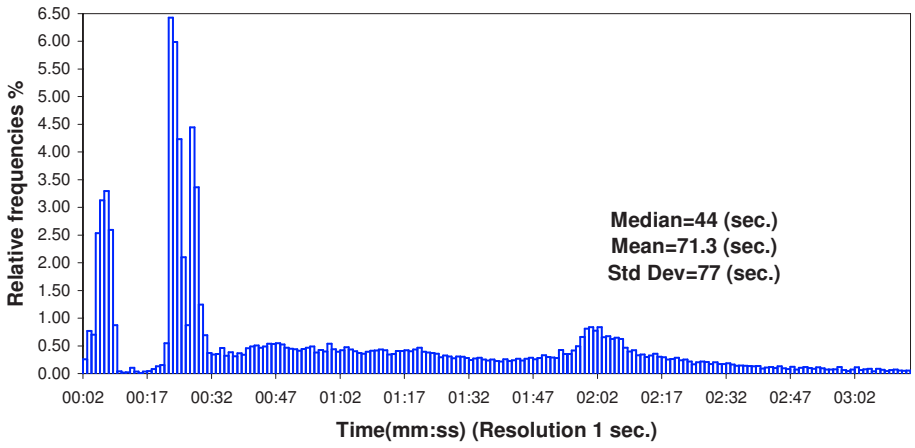

Figure 15: Histogram of the IVR service time for "Retail" customers

Figure 15 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; and the third peak can be related to the most common service in the IVR.

The distribution of the IVR service time is thus not exponential (see also [10]). A similar conclusion applies to agents' service time, as presented in Figure 16. Indeed, service time turns out to be log-normal (up to a probability mass near the origin) for about to 93% of calls; the other 7% calls 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]).

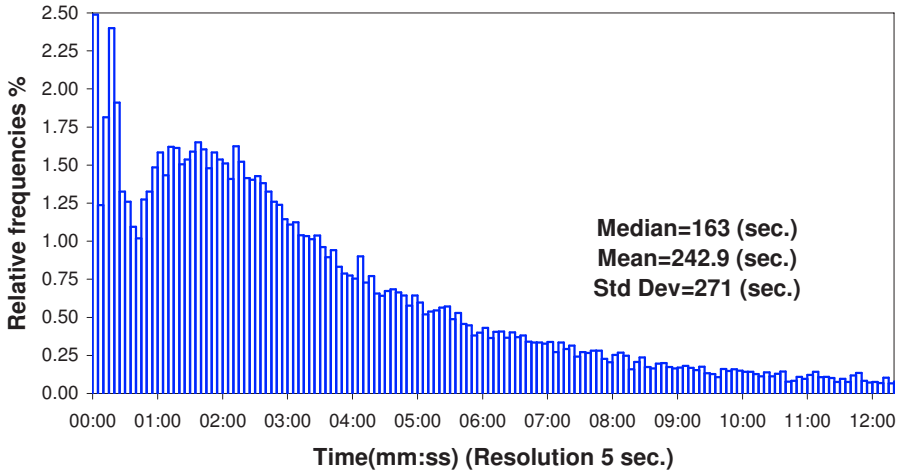

Figure 16: Histogram of Agents service time for "Retail" customers

Similarly to non-Markovian (non-exponentially distributed) service times, 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 [17]). 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 the model assumptions do not prevail in practice, notably 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.

8.3 Comparison of real and approximated performance measures

For our calculations, the following parameters must be estimated:

- λ - average arrival rate;
- θ - average rate of service in the IVR;

- μ - average rate of service by an agent;
- p - probability that a customer requests service by an agent;
- δ - average rate of customers' (im)patience;
- S - number of agents;
- N - number of trunk lines.

We consider the Retail service time distribution for 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.

The first four parameters were calculated for each 30 minute interval as follows:

$$\widehat{\lambda} = \text{number of calls arriving to the system (30 min)}$$

$$\widehat{\theta} = \frac{30 \times 60}{\text{average IVR service time (sec)}}$$

$$\widehat{\mu} = \frac{30 \times 60}{\text{average agent service time (sec)}}$$

$$\widehat{p} = \frac{\text{number of calls seeking agent service}}{\widehat{\lambda}}$$

It should be noted that, strictly speaking, we are not calculating the actual average arrival rate because we see only the calls which did not find all trunks busy; practically, the fraction of customers that found all trunks busy is very small and hence the difference between the real and approximated (calculated by our way) arrival rate is not significant.

The average rate of customers' patience was calculated via the relation

$$\delta = \frac{P(Ab|W > 0)}{E[W|W > 0]}, \tag{67}$$

which applies for the M/M/S/N+M queue (see [20] for details). Note that (67) assumes a linear relation between $P(Ab|W > 0)$ and $E[W|W > 0]$. The following Figure 17 demonstrates that this assumption is not unreasonable for our call center.

The estimation of the average rate of the customers' patience is thus the following:

$$\widehat{\delta} = \frac{\text{proportion of abandoned calls}}{\text{the average of the waiting time (sec)}} \times 30 \times 60, \tag{68}$$

where both numerator and denominator are calculated for customers with positive queueing time. Calculation of the estimation of the average rate of the customers' patience for our data gave varying behavior of this parameter, for example at 14:30 its value is 5, at 15:00 it equals to 1, and at 15:30

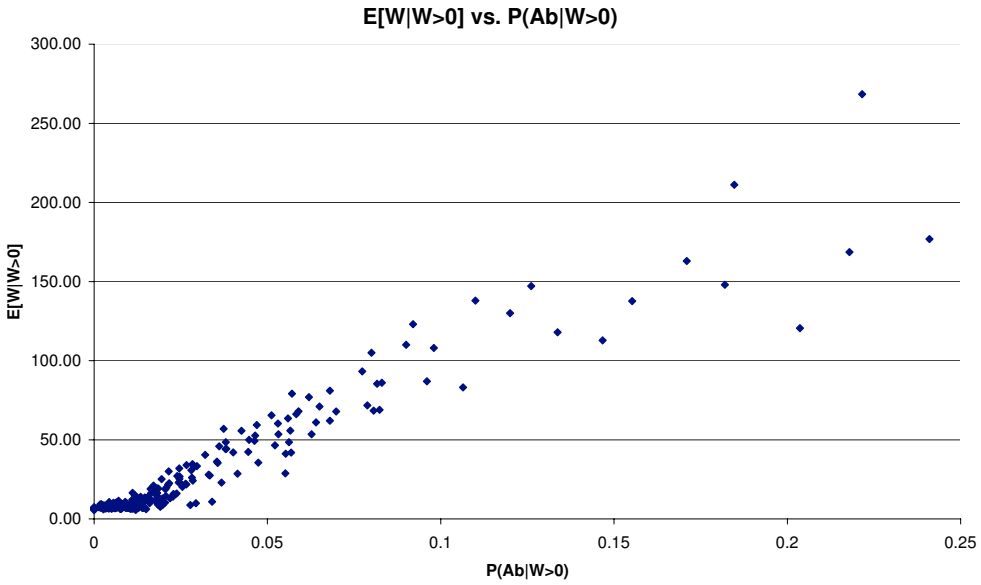

Figure 17: Relationship between the average waiting time given waiting, $E[W|W > 0]$, and the proportion of abandoning calls given waiting, $P(Ab|W > 0)$, for 30 minutes intervals over 20 days.

it equals to 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 use our approximations, we must assign an appropriate value for N, the number of trunk lines. 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 is as follows[6]:

$$\hat{N} = \frac{\text{total duration of all calls that arrived to the system}}{30 \cdot 60} .$$

The calculation of the number of agents is also problematic, because the agents who serve retail customers may also serve other types of customers, and vice versa: if all Retail agents are busy, the other agent types may serve Retail customers (see [16] for details). Thus, it is practically impossible to determine their exact number and that is why we use an averaged value, as follows:

$$\hat{S} = \frac{\text{total agent service time}}{30 \cdot 60}$$

The figure below shows the comparison of the approximate value for the probability to wait, calculated with the help of the above estimated parameters, against the exact proportion of waiting customers, as estimated from real data.

[6]Note that for the system with an IVR, $\hat{N}$ depends on the total duration of calls in the IVR, agents queue and service. For the system without IVR, it depends only on the total duration of calls in the agents queue and service.

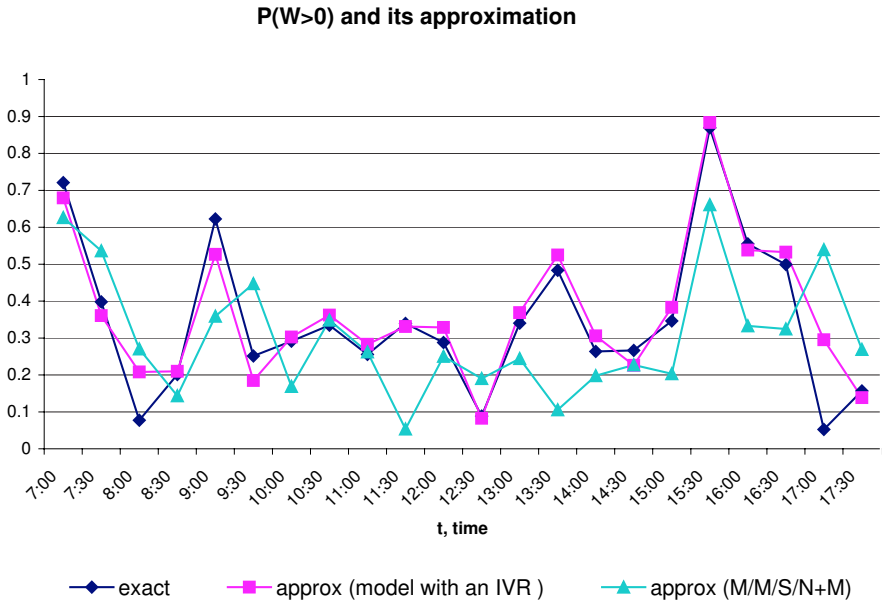

Figure 18: Comparison of approximated and real probabilities to wait.

In Figure 18 there are three curves. The blue one (line with diamonds) shows the proportion of customers that are waiting in the queue before agent service. This proportion is calculated for each half-hour period. The lilac one (squares) shows the approximation from the model with an IVR to the probability of waiting, which is calculated for each half-hour period. The last one (triangles) corresponds to the approximation of the probability to wait from the M/M/S/N+M queue model.

Considering the accuracy of the approximations, one observes that it is satisfactory, especially for the model with an IVR. The approximated values for this model, in many intervals, are very close to the exact proportion. In some intervals the difference is about 10%, which can be attributed to the non-perfect correspondence between the model and the real call center. An additional explanation is in the estimation of the parameters, such as N and S, which we estimate in a very crude way. The approximation from the M/M/S/N+M queue works less well and sometimes it 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, as well as due to ignoring the IVR influence.

In the figures below, we compare the real and approximate conditional probability for customer to abandon the system and the conditional average waiting time, given waiting:

Figures 19 and 20 show almost the same behavior as in Figure 18. Sometimes we see a larger deviation and a possible explanation is the sensitivity of our measures under heavy traffic, i.e. a little change of parameter values can dramatically change the performance measures.

In summary, both models considered above provide useful approximations to reality. Visual inspection reveals that the model with an IVR does it much better than the M/M/S/N+M queue.

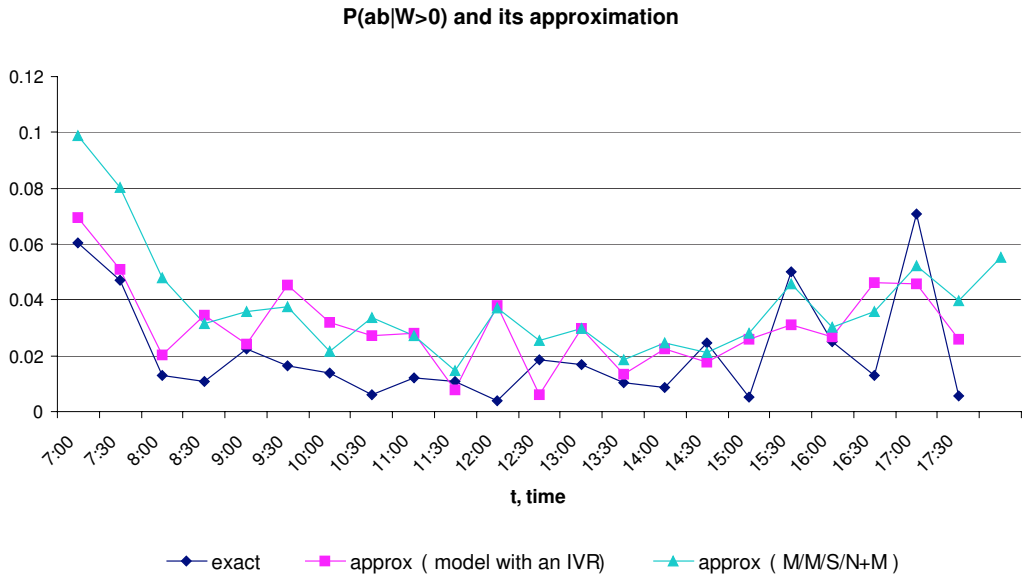

Figure 19: Comparison of the approximate and real conditional probability to abandon $P(ab|W > 0)$.

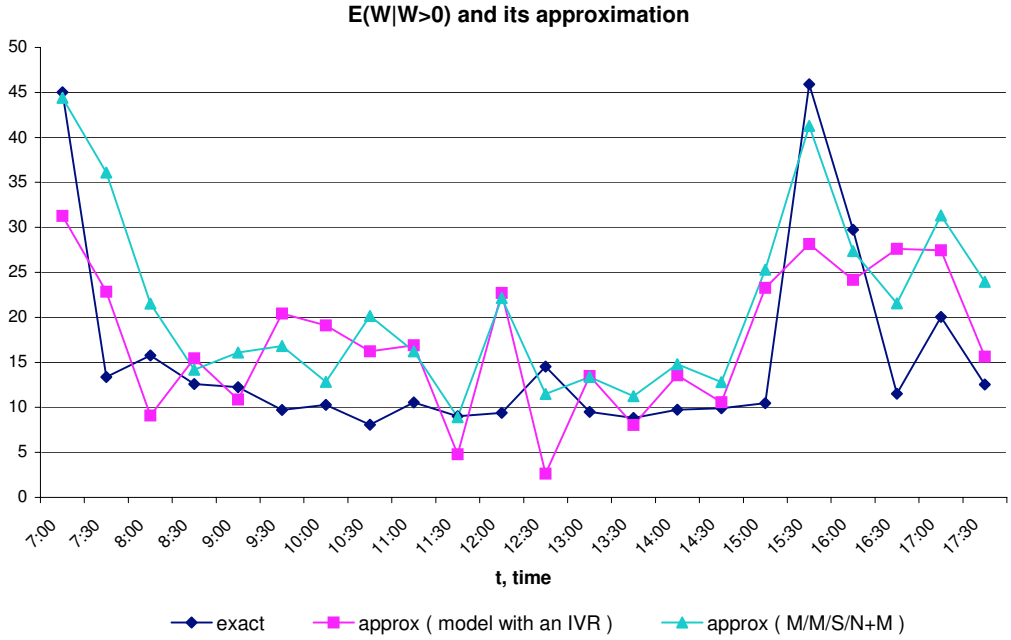

Figure 20: Comparison of the approximate and real conditional average waiting time $E(W|W > 0)$, in seconds.

9 Adding functionality to the IVR as a way to reduce operating costs

In this section, we use our approximations to analyze the tradeoff between cost and service level. As noted in [22], staffing costs (salary, training, etc.) account for over 65% of the operating costs of a typical call center. Hence, the main way to attempt cost reduction is to reduce the number of agents. A common way to attempt cost reduction without sacrificing service level is to extend IVR's capabilities. Adding functionality to the IVR will decrease the probability p to be served by an agent. Indeed, operations which previously only an agent could perform are nowadays carried out routinely by the "self-service" customer (via the IVR). Therefore, the number of customers wishing to be served by agents will decrease and, as a result, the number of agents S could decrease as well.

As an example, consider a call center with the following parameters:

(i) average arrival rate $\lambda = 1000$,

(ii) average service rate in the IVR θ equals 1,

(iii) average agent's service rate μ equals 1.

The above parametrization corresponds to the following state of affairs:

(1) the distributions of the IVR and agents service time are close to each other. (Our model could accommodate any other relation as well).

(2) time is measured in units of average service time (IVR time).

Suppose that there are performance constraints as follows:

$$P(W > 0) < 0.4, \qquad P(block) < 0.02. \tag{69}$$

We search for the optimal pair (S, N), where N is the number of trunk lines and S is the number of agents, $0 \leq S \leq N$. Optimality is in the sense that this pair (S, N) minimizes costs, subject the desired level of service.

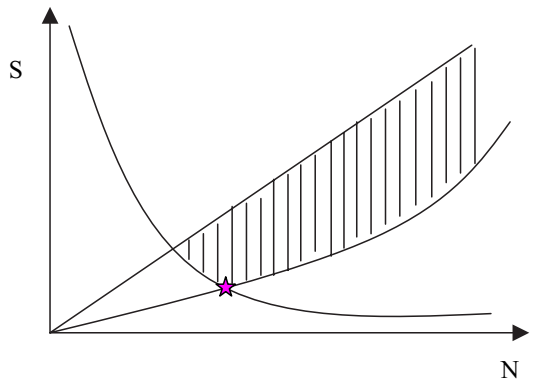

Figure 21: The feasible domain of pairs (S, N) that satisfy (69), with the "optimal" solution of this problem.

The area of admissible values of S and N for problem (69) is the shaded area in Figure 21. The number of solutions is infinite and we do not have an explicit cost function, but taking into account the fact that the hourly cost of agents is the main component (about 63% of total costs [22]), we safely assume that staffing cost is dominating trunk costs. This suggests that the optimal point is when S is the least one feasible, as indicated in the figure.

The algorithm for solving problem (69) was described in [14]. In that algorithm we used the exact formulae of performance measures. This was easy, because we considered a relatively small call center (with an arrival rate $\lambda = 20$). In the current example, and in all other examples in the present section, we consider large call centers (the arrival rate λ equals 1000). The calculation of exact performance measures in such a large call center, at the least, takes a long time and requires a complicated programming process; sometimes, the exact calculation is, in fact, numerically impossible. Thus, we use our approximations for the performance measures in finding the optimal solution.

Let us vary p - the probability to be served by an agent, over the range from 0 to 1, and for each value of p find the optimal solution (S, N).

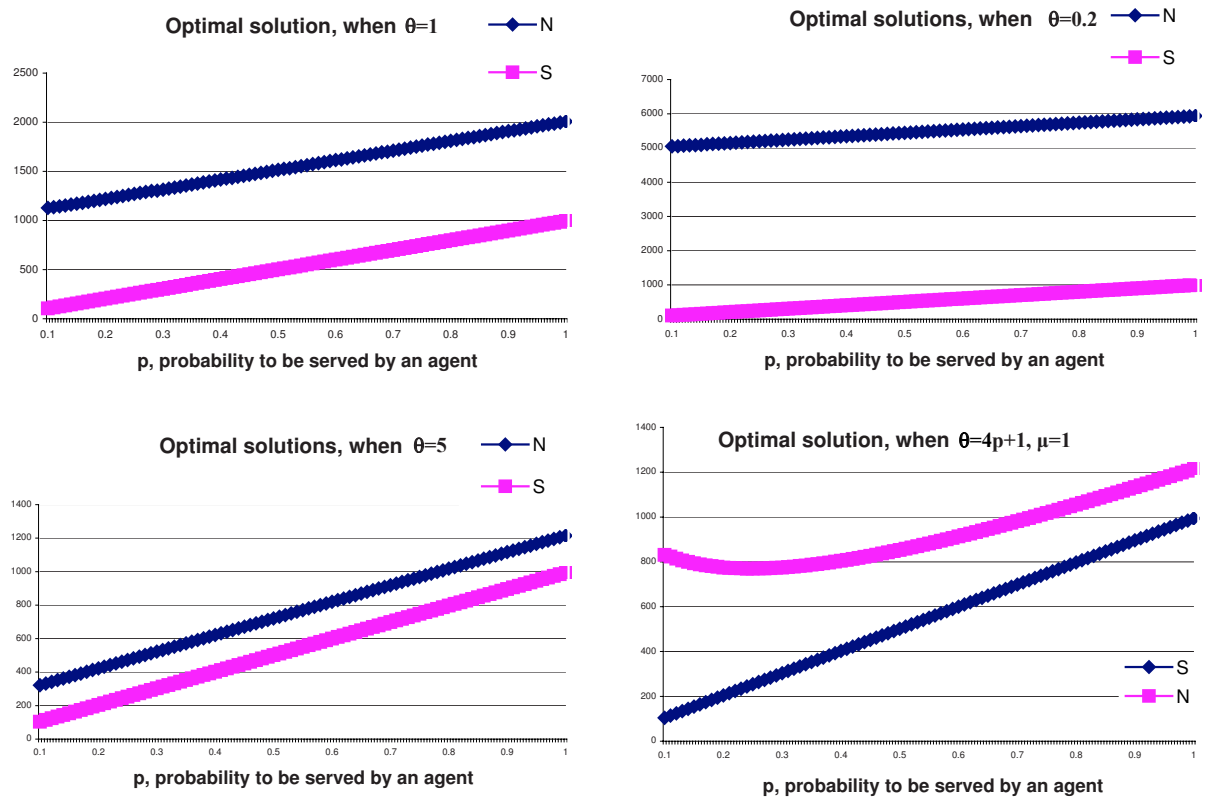

Figure 22: The optimal pairs (S, N) for a call center with an IVR, when the arrival rate equals 1000, the agent's and the IVR's service rates equals 1 and p varies from 0 to 1 .

The domain of (S, N) values is determined by (16) and (17). From these conditions the relationship between S and p is $S \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}}$. The values of β are small and hence, the influence of $\beta\sqrt{\frac{\lambda p}{\mu}}$

is negligible, relative to $\frac{\lambda p}{\mu}$. Using this fact, we can perform a rough approximation of the optimal number of agents. For example, if $p = 0.5$ we can predict that the optimal number of agents S will be equal to $500 = 1000 \cdot 0.5/1$. We can see in Figure 22 that this is almost true, but because of the small resolution one cannot infer the exact value. Note that such linear approximation is too rough and the optimal number of agents can be actually equal to 550, i.e. the error is 10%.

The relationship between N and p is similar to that between S and p, plus a term that does not depend on p:

$$N \approx \frac{\lambda p}{\mu} + \beta\sqrt{\frac{\lambda p}{\mu}} + \frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}.$$

Because of this extra term, the line for N is parallel to that for S, and the difference is equal to $\frac{\lambda}{\theta} + \eta\sqrt{\frac{\lambda}{\theta}}$. Moreover, we can roughly say that the difference is equal to $\frac{\lambda}{\theta}$, i.e. 1000 in our example, because $\eta\sqrt{\frac{\lambda}{\theta}}$ is negligible compared with $\frac{\lambda}{\theta}$. The exception is only in the case when θ also depends on p. In this case, the changes of N are not linear.

Experiment 1: Varying the IVR service duration. We now consider what happens when we change the parameter θ. Figure 22 shows how the optimal pair changes when the average service rate in the IVR is equal to 0.2, 1 and 5, i.e. the average service time in the IVR varies from 0.2 to 5 times more than what it was before. When $\theta = 0.2$, the optimal number of agents is exactly the same as when $\theta = 1$, and only the number of trunk lines N changes. Actually, this is not surprising, because we saw in Figure 22 that the optimal value of S does not depend on θ. When $\theta = 0.2$, the values of N are about 5000 trunk lines more than when $\theta = 1$, and this happens because now the difference between N and S is about $\frac{\lambda}{\theta} = \frac{1000}{0.2} = 5000$. When $\theta = 5$, the average service time in the IVR is 5 times less than in the case with $\theta = 1$. We can hypothesize that the optimal number of agents will be as follows: $N_{opt}(\theta = 5) \approx \frac{1}{5}(N_{opt}(\theta = 1) - S_{opt}(\theta = 1)) + S_{opt}(\theta = 5)$. Our intuition is that the optimal number of agents will not change and therefore the optimal number of trunk lines will be about $300 = (1100 - 100)/5 + 100$. Figure 22 supports this intuition.

Experiment 2: More opting for agents implies shorter IVR services. It is reasonable to assume that, along with changing the probability to be served by an agent, the service time in the IVR is changing as well. Unfortunately, we do not know how those changes occur. Figure 22 also presents what happens when the service rate in the IVR is a function of p. Intuitively this function must be an increasing function, because when the number of customers wishing to be served by an agent increases the time that these customers spend in the IVR is decreasing, therefore the service rate in the IVR is increasing. For simplicity, suppose that this function is linear such that when nobody wishes to be served by an agent ($p = 0$), the average service rate in the IVR equals 1, and when everyone wishes to be served by an agent ($p = 1$), the average service rate equals 5. Thus, this function has the following form:

$$\theta(p) = 4 \cdot p + 1. \tag{70}$$

According to the previous analysis, one would guess that changes in the service time in the IVR will not influence the optimal agent's number. Now let us look at the optimal solution (S, N) to the problem (69), but in the case when θ is given by (70). Figure 22 shows that our intuition was correct. The optimal number of agents did not change. This fact is very important, because we can see once again that adding functions to an IVR is a good way to reduce costs of a call center. The optimal

N values are not linear anymore it is a line similar to $N \approx \frac{\lambda}{\theta} + \frac{\lambda p}{\mu}$. This is a rough approximation, because we do not take into account the terms $\eta\sqrt{\frac{\lambda}{\theta}}$ and $\beta\sqrt{\frac{\lambda p}{\mu}}$, since these values are negligible and do not influence the form of the line N.

Experiment 3: More IVR functionality implies shorter or longer services by agents. Another property that can be manipulated with the addition of functions to the IVR is the service time at the agents' station. Indeed, if the IVR has more functionality, then the agents need not do some of these functions. This will decrease the average agent's service time and, as a result, will lead to a decrease in the optimal number of agents. But agent's service time might also increase as a result of additional functions to an IVR. Customers might have questions about the IVR usage, since it is more complicated now. Moreover, the customers who will be served by an agent, after adding more functions to the IVR, are expected to have more complicated requests which take longer to be satisfied. Thus, the relationship between the probability p to be served by an agent and the rate μ of an agent's service is not easy to predict. We thus consider two scenarios of changing the rate of the agent's service:

$$(I) \qquad \mu = 1 + 2 \cdot p \cdot (1 - p) \tag{71}$$

and

$$(II) \qquad \mu = 1 - 2 \cdot p \cdot (1 - p). \tag{72}$$

Let us now compare the optimal number of agents and trunk lines in these the two scenarios.

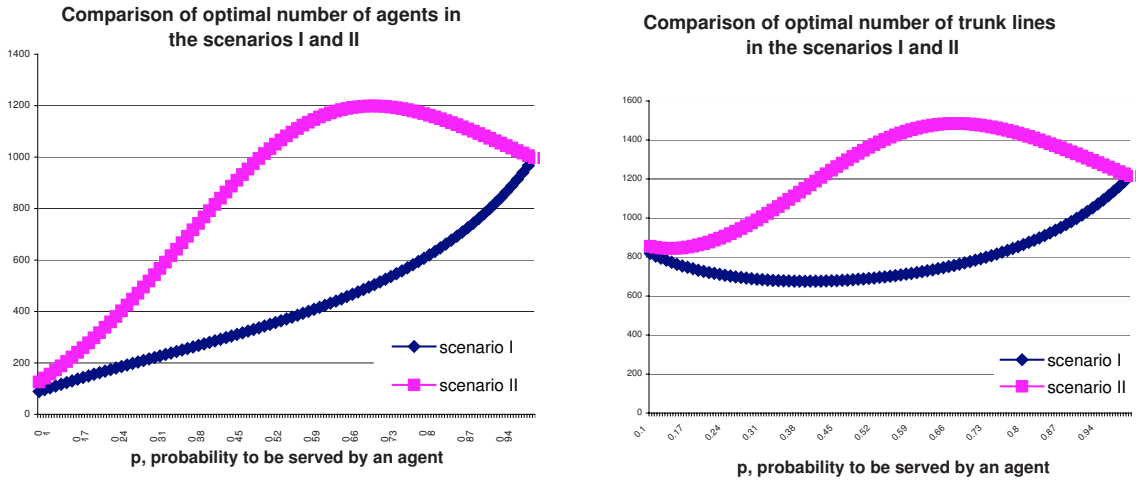

Figure 23: A comparison of the optimal number of agents and trunk lines for a call center with an IVR, when the arrival rate equals 1000, the IVR's service rate depends on p, p changes from 0 to 1, and the agent's service rate μ in scenario (I) equals $1 + 2p(1 - p)$ and in scenario (II) $1 - 2p(1 - p)$.

Figure 23 shows that, in the first scenario, the optimal number of agents as well as trunk lines first decrease with the growth of p (the probability to continue service). When $p = 0.5$ these values start to increase with the growth of p but they are always less than they would have been if the growth was

linear. In contrast, in the second scenario, the optimal values are always larger than if the growth was linear.

Thus, we see that adding functions to an IVR can provide an attractive solution for cost reduction. However, this may sometimes bring undesirable changes, for example in the second scenario. Indeed, before adding functions to the IVR, i.e. when p was equal to 1, the optimal agent's number was 503. After the addition of some functions to the IVR, for example, when $p = 0.7$, the optimal number of agents in the second scenario increased to 605. This is almost a 20% increase. As noted in [22], trunk costs constitute 5% of the staffing costs - the latter being 65% of a call center's operational costs. If we assume that these proportions are not changing with adding or reducing agents or trunk lines, then after adding functions to the IVR, the call center's costs increase by about 14%. Such a scenario can happen as a result of an unsuccessful design of the IVR, which underscores the importance in its proper deployment. As indicated in Section 1.1, when implementing an IVR, one must take into consideration not only the call center's interests, but also the wishes, needs and capabilities of customers in order to make the IVR friendly and effective.

A Appendix: Proofs

A.1 Proof of Lemma 3.1

Proof. In view of Stirling's formula, $S! \approx \sqrt{2S\pi}S^S e^{-S}$, one obtains for $\xi_1(\lambda)$:

$$\xi_1(\lambda) = \frac{e^{S-\lambda\frac{p}{\mu}}}{\sqrt{2S\pi}S^S}\left(\frac{\lambda p}{\mu}\right)^S \frac{\sqrt{S}}{\beta}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} + o(1) \tag{73}$$

The last sum 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

$$P\left(X_\lambda \leq N-S-1\right) = P\left(\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \leq \frac{N-S-1-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}\right) \tag{74}$$

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

$$\frac{X_\lambda - \frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}} \Rightarrow N(0,1) \tag{75}$$

and due to assumption *(ii)* in (16) we get [7]

$$\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \longrightarrow P\left(N(0,1) \leq \eta\right) = \Phi(\eta), \quad when \quad \lambda \to \infty, \tag{77}$$

[7] Here we are using the following theorem (from [4], p. 114)

Theorem A.1. *Let $\zeta_n \Rightarrow \zeta$ and assume that F_ζ - the distribution function of ζ, is everywhere continuous. Let also $x_n \to x_\infty$ as $n \to \infty$, where $\{x_n\}$ is a sequence of scalars. Here $x_\infty \in [-\infty, \infty]$. Then*

$$F_{\zeta_n}(x_n) \longrightarrow F_\zeta(x_\infty) \tag{76}$$

where $N(0,1)$ is a standard normal random variable with distribution function Φ. It follows from (73)-(77) that

$$\xi_1(\lambda) = \frac{e^{S(1-\rho)}}{\sqrt{2\pi}\beta}\rho^S\Phi(\eta) + o(1) = \frac{e^{S((1-\rho)+\ln\rho)}}{\sqrt{2\pi}\beta}\Phi(\eta) + o(1), \tag{78}$$

where $\rho = \frac{\lambda p}{S\mu}$. Making use of the Taylor expansion

$$\ln\rho = \ln\left(1-(1-\rho)\right) = -(1-\rho) - \frac{(1-\rho)^2}{2} + o(1-\rho)^2 \quad (\rho \to 1), \tag{79}$$

one obtains from (79) and (16)*(ii)* that

$$\xi_1(\lambda) = \frac{e^{S((1-\rho)-(1-\rho)-\frac{(1-\rho)^2}{2})}}{\sqrt{2\pi}\beta}\Phi(\eta) + o(1) = \frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi}\beta}\Phi(\eta) + o(1). \tag{80}$$

This proves equation (30).

By applying the Stirling's formula and using that $\rho = \frac{\lambda p}{S\mu} \to 1$, as $\lambda \to \infty$ and $S \to \infty$, one obtains,

$$\xi_2(\lambda) = \frac{e^{S-\lambda\frac{p}{\mu}+\frac{\lambda(1-\rho)}{\theta\rho}}}{\sqrt{2S\pi}}\rho^N\frac{\sqrt{S}}{\beta}\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i e^{-\frac{\lambda}{\theta\rho}} + o(1) \tag{81}$$

The last sum can be rewritten as $P\left(Y_\lambda \le N-S-1\right)$ where $Y_\lambda \stackrel{d}{=} Pois(\frac{\lambda}{\theta\rho})$, and $E[Y_\lambda] = \frac{\lambda}{\theta\rho}$, $Var[Y_\lambda] = \frac{\lambda}{\theta\rho}$. Note that

$$P\left(Y_\lambda \le N-S-1\right) = P\left(\frac{X_\lambda - \frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}} \le \frac{N-S-1-\frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}}\right) \tag{82}$$

It follows from (16)*(ii)* that

$$\lim_{\lambda\to\infty}\frac{N-S-\frac{\lambda}{\theta\rho}}{\sqrt{\frac{\lambda}{\theta\rho}}} = \eta - \sqrt{\frac{\mu}{p\theta}}\beta. \tag{83}$$

Taking into account equations (81) and (83), the Central Limit Theorem and Theorem A.1 we have that

$$\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta\rho}\right)^i e^{-\frac{\lambda}{\theta\rho}} \longrightarrow P\left(N(0,1) \le \eta_1\right) = \Phi(\eta_1), \quad when\ \lambda \longrightarrow \infty. \tag{84}$$

It follows from the assumption (16) and (79) that when $\rho \longrightarrow 1$

$$\begin{aligned} S - \frac{\lambda p}{\mu} + \frac{\lambda(1-\rho)}{\theta\rho} + N\ln\rho &= S(1-\rho) + \frac{\lambda}{\rho\theta}(1-\rho) - N(1-\rho) - \frac{N}{2}(1-\rho)^2 + o((1-\rho)^2) \\ &= (\frac{\lambda}{\theta\rho} - \frac{N}{2})(1-\rho)^2 - \eta\sqrt{\frac{\lambda}{\theta}}(1-\rho) + o((1-\rho)^2) \\ &= -\frac{1}{2}(\eta^2+\beta^2) + \frac{1}{2}(\eta - \sqrt{\frac{\mu}{p\theta}}\beta)^2 + o(1) \end{aligned}$$

Therefore,

$$\lim_{\lambda\to\infty}\xi_2 = \frac{e^{-\frac{1}{2}(\eta^2+\beta^2)+\frac{1}{2}\eta_1^2}}{\sqrt{2\pi}\beta}\Phi(\eta_1) = \frac{\varphi(\sqrt{\eta^2+\beta^2})}{\beta}e^{\frac{1}{2}\eta_1^2}\Phi(\eta_1), \tag{85}$$

and this proves equation (31).

Approximating $\gamma(\lambda)$: For this purpose, consider a partition $\{S_j\}_{j=0}^{l}$ of the interval $[0, S]$.

$$S_j = S - j\delta, \quad j = 0, 1, ..., l; \quad S_{l+1} = 0, \tag{86}$$

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 (ii), then l is less then $\frac{S}{\delta}$ 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.

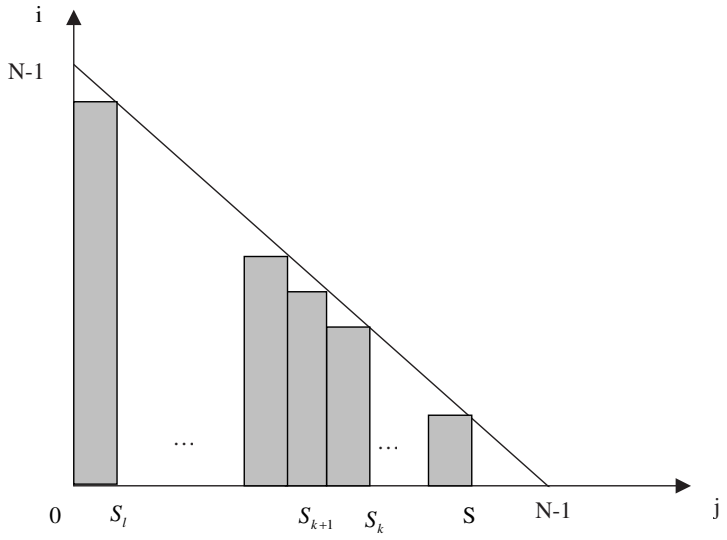

Figure 24: **Area of the summation of the variable γ_1.**

The variable $\gamma(\lambda)$ is given by formula (25), where the summation is taken over the trapezoid. Consider the lower estimate for $\gamma(\lambda)$, given by the following sum, where the summation is over the shaded area in Figure 4.1:

$$\begin{aligned}\gamma(\lambda) \geq \gamma_1(\lambda) &= \sum_{k=0}^{l} \sum_{j=S_{k+1}}^{S_k-1} \frac{1}{j!} \left(\frac{\lambda p}{\mu}\right)^j e^{-\frac{\lambda p}{\mu}} \sum_{i=0}^{N-S_k} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}} \\ &= \sum_{k=0}^{l} P(S_{k+1} \leq Z_\lambda < S_k) P(X_\lambda \leq N - S_k),\end{aligned} \tag{87}$$

where

$$\begin{aligned} Z_\lambda &\stackrel{d}{=} Pois\left(\frac{\lambda p}{\mu}\right), \qquad E[Z_\lambda] = \frac{\lambda p}{\mu}, \qquad Var[Z_\lambda] = \frac{\lambda p}{\mu}; \\ X_\lambda &\stackrel{d}{=} Pois\left(\frac{\lambda}{\theta}\right), \qquad E[X_\lambda] = \frac{\lambda}{\theta}, \qquad Var[X_\lambda] = \frac{\lambda}{\theta}. \end{aligned} \tag{88}$$

Analogously to Lemmas 3.1 and 3.2, applying the Central Limit Theorem and making use of the relations

$$\lim_{\lambda\to\infty}\frac{S_k-\frac{\lambda p}{\mu}}{\sqrt{\frac{\lambda p}{\mu}}}=\beta-k\varepsilon,\qquad k=0,1,...,l, \tag{89}$$

$$\lim_{\lambda\to\infty}\frac{N-S_k-\frac{\lambda}{\theta}}{\sqrt{\frac{\lambda}{\theta}}}=\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}},\qquad k=0,1,...,l, \tag{90}$$

one obtains

$$\lim_{\lambda\to\infty}P(S_{k+1}\le Z_\lambda<S_k)=\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon),\quad k=0,1,...l-1, \tag{91}$$

$$\lim_{\lambda\to\infty}P(0\le Z_\lambda<S_l)=\Phi(\beta-l\varepsilon), \tag{92}$$

$$\lim_{\lambda\to\infty}P(X_\lambda<N-S_k)=\Phi(\eta+k\varepsilon\sqrt{\frac{p\theta}{\mu}}),\quad k=0,1,...l. \tag{93}$$

It follows from (87) and (91), (92), (93) that

$$\begin{aligned}\liminf_{\lambda\to\infty}\gamma(\lambda)\ge&\sum_{k=0}^{l-1}\Phi(\eta+k\varepsilon\sqrt{p\theta/\mu})[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]\\&+\Phi(\beta-l\varepsilon)\Phi(\eta+l\varepsilon\sqrt{p\theta/\mu}).\end{aligned} \tag{94}$$

It is easy to see that (94) is the lower Riemann-Stieltjes sum for the integral

$$-\int_0^\infty\Phi\left(\eta+s\sqrt{\frac{p\theta}{\mu}}\right)d\Phi(\beta-s)=\int_{-\infty}^{\beta}\Phi\left(\eta+(\beta-t)\sqrt{\frac{p\theta}{\mu}}\right)\varphi(t)dt, \tag{95}$$

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

Similarly, let us take the upper estimate for γ as the following sum

$$\begin{aligned}\gamma\le\gamma_2&=\sum_{k=0}^{l}\sum_{j=S_{k+1}}^{S_k-1}\frac{1}{j!}\left(\frac{\lambda p}{\mu}\right)^j e^{-\frac{\lambda p}{\mu}}\sum_{i=0}^{N-S_{k+1}}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}}\\&=\sum_{k=0}^{l}P(S_{k+1}\le Z_\lambda<S_k)P(X_\lambda\le N-S_{k+1}).\end{aligned} \tag{96}$$

The above calculations, applied to the sum (96), give the following asymptotic estimate for γ:

$$\limsup_{\lambda\to\infty}\gamma(\lambda)\le\sum_{k=0}^{l-1}\Phi\left(\eta+(k+1)\varepsilon\sqrt{\frac{p\theta}{\mu}}\right)[\Phi(\beta-k\varepsilon)-\Phi(\beta-(k+1)\varepsilon)]+\Phi(\beta-l\varepsilon), \tag{97}$$

which is the upper Riemann-Stieltjes sum for the integral (95).

When $\varepsilon \to 0$, the estimates (94), (96) lead to the following equality

$$\lim_{\lambda\to\infty} \gamma(\lambda) = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t)\sqrt{\frac{p\theta}{\mu}}\right) \varphi(t)dt. \tag{98}$$

This proves equation (32).
Now, consider $\delta(\lambda)$:

$$\delta(\lambda) = \sum_{k=0}^{N-S-1} \frac{\left[\lambda(\frac{1}{\theta} + \frac{pt}{\sqrt{S}})\right]^k}{k!} e^{-\lambda(\frac{1}{\theta}+\frac{pt}{\sqrt{S}})} = P(X_\lambda < N - S),$$

where

$$X_\lambda \stackrel{d}{=} Pois(\lambda(\frac{1}{\theta} + \frac{pt}{\sqrt{S}})), \qquad E[X_\lambda] = \lambda(\frac{1}{\theta} + \frac{pt}{\sqrt{S}}), \qquad Var[X_\lambda] = \lambda(\frac{1}{\theta} + \frac{pt}{\sqrt{S}}).$$

By the Central Limits Theorem and Theorem A.1, one obtains

$$\lim_{\lambda\to\infty} P(X_\lambda < N - S) = \Phi(\eta - \sqrt{p\mu\theta}t). \tag{99}$$

□

A.2 Proof of Lemma 3.2

Proof. Using Stirling's approximation and assumptions (17) we have

$$\begin{aligned}
\xi(\lambda) &= e^{-\lambda(\frac{1}{\theta}+\frac{p}{\mu})} \frac{e^S}{\sqrt{2\pi S}} \left(\frac{\lambda p}{\mu S}\right)^S \sum_{i=0}^{N-S-1} \frac{1}{i!} \left(\frac{\lambda}{\theta}\right)^i \sum_{j=0}^{N-S-i-1} \left(\frac{\lambda p}{\mu S}\right)^j + o(1) \\
&= \rho^S \frac{e^{S-\frac{\lambda p}{\mu}}}{\sqrt{2\pi S}} \sum_{i=0}^{N-S-1} \frac{e^{-\frac{\lambda}{\theta}} \left(\frac{\lambda}{\theta}\right)^i}{i!} \cdot \frac{1-\rho^{N-S-i}}{1-\rho} + o(1),
\end{aligned} \tag{100}$$

where $\rho = \frac{\lambda p}{S\mu}$. Under condition (17)*(ii)* $\beta = 0$, and this happens when $\rho = 1$ or $\rho \to 1$. When $\rho \to 1$, $\rho \neq 1$ we use the well known approximations

$$\frac{1-\rho^k}{1-\rho} = k + o(1-\rho). \tag{101}$$

This implies that in this case

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

When $\rho = 1$, the sum $\sum_{j=0}^{N-S-i-1} \rho^j$ in (100) is equal to $N-S-i$ and this leads to the same expression for ξ.

Simple calculations show that

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

Due to (77), the first term in (100) can be rewritten as follows:

$$(N-S-\frac{\lambda}{\theta})\sum_{i=0}^{N-S-1}\frac{1}{i!}\left(\frac{\lambda}{\theta}\right)^i e^{-\frac{\lambda}{\theta}}=\eta\sqrt{\frac{\lambda}{\theta}}\Phi(\eta),\qquad(\lambda\to\infty)+o(1).\tag{102}$$

It follows from Stirling's formula, (79) and conditions (17) that in (102)

$$\begin{aligned}\frac{(\frac{\lambda}{\theta})^{N-S}e^{-\frac{\lambda}{\theta}}}{(N-S-1)!} &= \sqrt{\frac{N-S}{2\pi}}e^{(N-S)[1-\frac{\lambda}{\theta(N-S)}+\ln\frac{\lambda}{\theta(N-S)}]}+o(1)\\ &= \sqrt{\frac{\lambda}{\theta}}\varphi(\eta)+o(1).\end{aligned}\tag{103}$$

Using (17)*(ii)*, (102) and (103) we obtain

$$\begin{aligned}\lim_{\lambda\to\infty}\xi(\lambda) &= \lim_{\lambda\to\infty}\frac{1}{\sqrt{2\pi S}}\sqrt{\frac{\lambda}{\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right)\\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\mu}{p\theta}}\left(\eta\Phi(\eta)+\varphi(\eta)\right).\end{aligned}\tag{104}$$

This proves Lemma 3.2. □

A.3 Proof of Theorem 4.1

Proof. Recall that the M/M/S/N queue has the following stationary distribution:

$$\pi_i=\begin{cases}\pi_0\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i, & 0\le i<S;\\ \pi_0\frac{1}{S!S^{i-S}}\left(\frac{\lambda}{\mu}\right)^i, & S\le i\le N;\\ 0, & \text{otherwise.}\end{cases}\tag{105}$$

where

$$\pi_0=\left(\sum_{i=0}^{S-1}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i+\sum_{i=S}^{N}\frac{1}{S!S^{i-S}}\left(\frac{\lambda}{\mu}\right)^i\right)^{-1}.\tag{106}$$

As in the previous analysis, we denote the waiting time by W. Using PASTA, we now find the probability to wait:

$$P(W>0)=\sum_{i=S}^{N}\pi_i=\frac{\sum_{i=S}^{N-1}\frac{1}{S!S^{i-S}}\left(\frac{\lambda}{\mu}\right)^i}{\sum_{i=0}^{S-1}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i+\sum_{i=S}^{N}\frac{1}{S!S^{i-S}}\left(\frac{\lambda}{\mu}\right)^i}\tag{107}$$

and the probability to find the system busy:

$$P(block) = \pi_N. \tag{108}$$

The expectation of waiting time is obtained from Little's formula:

$$E[W] = \frac{L_{queue}}{\lambda_{eff}} = \frac{\sum_{i=S+1}^{N} (i-S)\pi_i}{\lambda(1-P(block))}. \tag{109}$$

The conditional density function of the waiting time for the M/M/S/N queue has the following form:

$$f_{W|W>0}(t) = \begin{cases} \dfrac{\mu S(1-\frac{\lambda}{\mu S})e^{-\mu S(1-\frac{\lambda}{\mu S})t}}{1-\left(\frac{\lambda}{\mu S}\right)^{N-S}} \sum_{k=0}^{N-S-1} \dfrac{e^{-\lambda t}(\lambda t)^k}{k!}, & \rho \neq 1; \\ \dfrac{\mu S}{N-S} \sum_{k=0}^{N-S-1} \dfrac{e^{-\lambda t}(\lambda t)^k}{k!}, & \rho = 1. \end{cases} \tag{110}$$

In the case $\beta > 0$, this formula was found in [19] by Massey and Wallace. When $\beta \leq 0$, it can be obtained with the help of Laplace transform in a way similar to that in Section 6.1. Define

$$\gamma(\lambda) = \sum_{i=0}^{S-1} \frac{e^{-\frac{\lambda}{\mu}}}{i!} \left(\frac{\lambda}{\mu}\right)^i; \tag{111}$$

$$\xi(\lambda) = \sum_{i=S}^{N-1} \frac{e^{-\frac{\lambda}{\mu}}}{S!S^{i-S}} \left(\frac{\lambda}{\mu}\right)^i; \tag{112}$$

$$\delta(\lambda) = \frac{e^{-\frac{\lambda}{\mu}}}{S!S^{N-S}} \left(\frac{\lambda}{\mu}\right)^N; \tag{113}$$

$$\zeta(\lambda) = \frac{1}{\lambda} \sum_{i=S+1}^{N} \frac{e^{-\frac{\lambda}{\mu}}}{S!S^{i-S}} (i-S) \left(\frac{\lambda}{\mu}\right)^i. \tag{114}$$

Thus, we can rewrite the operational characteristics of the M/M/S/N queue as follows:

$$P(W>0) = \frac{\xi(\lambda)}{\gamma(\lambda)+\xi(\lambda)}; \tag{115}$$

$$P(block) = \frac{\delta(\lambda)}{\gamma(\lambda)+\xi(\lambda)}; \tag{116}$$

$$E[W] = \frac{\zeta(\lambda)}{\gamma(\lambda)+\xi(\lambda)}. \tag{117}$$

Note that $\gamma(\lambda)$ can be rewritten as $P(X_\lambda < S)$, where $X_\lambda \stackrel{d}{=} Pois(\frac{\lambda}{\mu})$, and $E[X_\lambda] = \frac{\lambda}{\mu}$, $Var[X_\lambda] = \frac{\lambda}{\mu}$. Then by the Central Limit Theorem, condition (ii) of the Theorem 4.1 and Theorem A.1, one obtains

$$\gamma(\lambda) = P\left(\frac{X_\lambda - \frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}} < \frac{S-\frac{\lambda}{\mu}}{\sqrt{\frac{\lambda}{\mu}}}\right) \to \Phi(\beta). \tag{118}$$

We now derive the approximation for $\xi(\lambda)$. In view of Stirling's formula and Taylor expansion, one obtains

$$\lim_{\lambda\to\infty}\xi(\lambda)=\frac{e^{-\frac{\beta^2}{2}}}{\sqrt{2\pi S}}\sum_{k=0}^{N-S-1}\rho^k. \tag{119}$$

Due to the conditions (i) and (ii) of theorem and (79), the expression ρ^{N-S} can be rewritten in the equivalent form

$$\rho^{N-S}=e^{(N-S)\ln\rho}=e^{(N-S)}(\rho-1)+o(1)=e^{-\eta\beta}+o(1). \tag{120}$$

Therefore, when $\beta\neq 0$,

$$\lim_{\lambda\to\infty}\xi(\lambda)=\frac{\varphi(\beta)}{\beta}(1-e^{-\eta\beta}). \tag{121}$$

Now, let $\beta=0$, i.e. $\rho=1$ or $\rho\to 1$. In this case

$$\lim_{\rho\to 1}\sum_{i=S}^{N-1}\rho^i=N-S.$$

Using the conditions (i) and (ii) of the Theorem, one obtains

$$\lim_{\lambda\to\infty}\xi(\lambda)=\frac{\eta}{\sqrt{2\pi}}. \tag{122}$$

Now, consider $\delta(\lambda)$. By Stirling's formula and relation (120), one gets

$$\lim_{\lambda\to\infty}\sqrt{S}\delta(\lambda)=\varphi(\beta)e^{-\eta\beta}. \tag{123}$$

In order to find an approximation for ζ, let us use the the formula

$$\sum_{k=1}^{M}k\rho^k=\frac{\rho^{M+1}}{\rho-1}\cdot M+\frac{1-\rho^M}{(1-\rho)^2}\cdot\rho,$$

conditions (i) and (ii) of the theorem and relation (120). Thus, one deduces that

$$\frac{\rho^{N-S+1}}{\rho-1}(N-S)+\frac{1-\rho^{N-S}}{(1-\rho)^2}\rho=-\eta\sqrt{\frac{\lambda}{\mu}}\frac{e^{-\eta\beta}\sqrt{S}}{\beta}+\frac{\lambda(1-e^{-\eta\beta})}{\beta^2\mu}+o(1). \tag{124}$$

Taking into account equation (124), we have

$$\lim_{\lambda\to\infty}\sqrt{S}\zeta=\frac{\varphi(\beta)}{\mu\beta}\left[\frac{1-e^{-\eta\beta}}{\beta}-\eta e^{-\eta\beta}\right], \tag{125}$$

when $\beta\neq 0$. The case $\beta=0$ means that $\rho=1$ or $\rho\to 1$. If $\rho=1$ it is easy to see that

$$\frac{1}{\lambda}\sum_{k=1}^{N-S}k\rho^k=\frac{1}{\lambda}\frac{(N-S)(N-S+1)}{2}\approx\frac{\eta^2}{2\mu}. \tag{126}$$

If $\rho\to 1$, by using Taylor's expansion one gets (126) (see [14] p.69 for details). Thus, when $\beta=0$, the approximation for $\sqrt{S}\zeta$ has the form

$$\lim_{\lambda\to\infty}\sqrt{S}\zeta=\frac{\eta^2}{2\mu\sqrt{2\pi}}. \tag{127}$$

Now, consider the conditional density function of the waiting time for the M/M/S/N queue. When $\beta \neq 0$, using condition (i) of the theorem and equation (120), one gets

$$\frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \frac{\mu S(1-\rho)e^{-\mu S(1-\rho)t}}{1-\rho^{N-S}} \sum_{k=0}^{N-S-1} \frac{e^{-\sqrt{\lambda\mu}t}\left(\sqrt{\lambda\mu}t\right)^k}{k!} + o(1).$$

The last sum can be rewritten as $P(X_\lambda < N-S)$, where $X_\lambda \stackrel{d}{=} Pois(\sqrt{\lambda\mu}t)$. From the strong law of large numbers for the Poisson process we have

$$\lim_{\lambda\to\infty} P\left(\frac{X_\lambda}{\sqrt{\frac{\lambda}{\mu}}} = \mu t\right) = 1. \tag{128}$$

Thus,

$$\lim_{\lambda\to\infty} P(X_\lambda < N-S) = \lim_{\lambda\to\infty} P\left(\frac{X_\lambda}{\sqrt{\frac{\lambda}{\mu}}} < \eta\right) = \begin{cases} 1, & \mu t < \eta, \\ 0, & \mu t \geq \eta; \end{cases} \tag{129}$$

and approximation of the density function when $\beta \neq 0$ has the following form:

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \dfrac{\mu\beta e^{-\mu\beta t}}{(1-e^{-\eta\beta})}, & \mu t < \eta, \qquad \beta \neq 0; \\ 0, & \mu t \geq \eta, \qquad \beta \neq 0. \end{cases} \tag{130}$$

When $\beta = 0$ using conditions (i) and (ii) of the theorem and equation (129), one obtains

$$\lim_{\lambda\to\infty} \frac{1}{\sqrt{S}} f_{W|W>0}\left(\frac{t}{\sqrt{S}}\right) = \begin{cases} \dfrac{\mu}{\eta}, & \mu t < \eta, \qquad \beta = 0; \\ 0, & \mu t \geq \eta, \qquad \beta = 0. \end{cases} \tag{131}$$

Combining (118), (121), (122), (123), (125), (127), (130) and (131), we have thus proved Theorem 4.1. □

B Frequently used notation

λ arrival rate

θ IVR service rate

μ agent service rate

S number of agents

N number of trunk lines

ρ offered load per agent in the system for Call Center with IVR $\quad(\rho = \lambda p/(S\mu))$

R offered load (often $R = \lambda/\mu$ in Markovian queues)

$Q_1(t)$ the number of calls at the IVR

$Q_2(t)$ the number of calls at the agents station (getting service and in queue)

$\pi(i,j)$ the stationary probabilities of having i calls at the IVR and j calls at the agents station

$\chi(k,j)$ the probability that the system is in state (k,j), $(0 \leq j < k \leq N)$, when a call (among the $k-j$ customers) is about to finish its IVR service. Here, k is the total number of calls in the system, and j is the number of calls in the agents' station (waiting or served); hence, $k-j$ is the number of calls at the IVR

$\Phi(\cdot)$ the standard normal distribution function

$\varphi(\cdot)$ the standard normal density function

E expectation

P probability measure

W waiting time after the IVR service, for a customer seeking service

$W(t)$ distribution function of the waiting time: $W(t) = P(W \leq t)$

$f_W(t)$ density function of the waiting time

$a_n \approx b_n$ if $a_n/b_n \to 1$, as $n \to \infty$

$\stackrel{d}{=}$ distributed as (for example, $X \stackrel{d}{=} Pois(\lambda)$ mean that X is a random variable that is Poisson distributed with parameter λ)

$a_n = o(b_n)$ if $a_n/b_n \to 0$, as $n \to \infty$

$L_W(x)$ Laplace transform of $f_W(t)$

$L_W^{-1}(t)$ inverse Laplace transform for the function $L_W(x)$

QED Conditions (16)

QED_0 Conditions (17)

References

[1] Borst S., Mandelbaum A. and Reiman M. (2004). Dimensioning large call centers. *Operations Research*, 52(1), 17-34. 3

[2] Brandt A., Brandt M., Spahl G. and Weber D., "Modelling and Optimization of Call Distribution Systems", Elsevier Science B.V., 1997. 4, 6

[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.", 2005, Journal of the American Statistical Association, Vol 100: 36-50. 32

[4] Chernova N., "Theory of Probability", Lecture notes, 2003, available at http://www.nsu.ru/mmf/tvims/chernova/tv/lec/lec.html 41

[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. 3

[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), pp. 79-141, 2003. 3, 4, 9

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

[8] Gilson K.A. and Khandelwal D.K., "Getting more from call centers", The McKinsey Quarterly, Web exclusive, April 2005, available at `http://www.mckinseyquarterly.com/article_page.aspx` 3

[9] Chen H. and Yao D.D., "Fundamentals of Queueing Networks", New. York: Springer-Verlag, 2001. 8

[10] Donin O., Trofimov V., Feigin P., Mandelbaum A., Zeltyn S., Ishay E., Nadjarov E. and Khudyakov P., "DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 4.1: The Call Center of "US Bank"", August 2006, available at `http://iew3.technion.ac.il/serveng/References/references.html` 32

[11] Halfin S. and Whitt W. "Heavy-Traffic Limits for Queues with Many Exponential Servers",Operations Research, 29, pp. 567-587, 1981. 3, 8, 17

[12] Jagerman D.L., "Some properties of the Erlang loss function", Bell Systems Technical Journal, **53:3**, pp. 525-551, 1974. 3

[13] Jelenkovic P., Mandelbaum A. and Momcilovic P., "Heavy Traffic Limits for Queues with Many Deterministic Servers", QUESTA 47, pp. 53-69, 2004. 3

[14] 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` 6, 11, 22, 26, 27, 38, 48

[15] Kocaga Y.L. and Ward A.R., " Dynamic Outsourcing for Call Centers ", First Submitted January 2009, available at `http://www-rcf.usc.edu/~amyward/KW_Dynamic_Outsourcing.pdf` 29

[16] Liberman P., Trofimov V. and Mandelbaum A. "DATA-MOCCA: Data MOdel for Call Center Analysis. Volume 5.1: Skills-Based-Routing-USBank", February 2008, available at `http://iew3.technion.ac.il/serveng/References/references.html` 34

[17] 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 32

[18] Mandelbaum A., "Lecture Notes on QED Queues", available at http://iew3.technion.ac.il/serveng/Lectures/QED_lecture_Introduction_2008S.pdf 27

[19] 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. 3, 4, 8, 9, 14, 47

[20] Mandelbaum A. and Zeltyn S., "Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue", Queueing Systems, Volume 51, pp. 361-402, December 2005, available at
http://iew3.technion.ac.il/serveng/References/references.html 3, 33

[21] Pang R., Talreja R. and Whitt W., "Martingale proofs of many-server heavy-traffic limits for Markovian queues", Probability Surveys, Vol. 4, pp. 193-267, 2007, available at
http://www.columbia.edu/~ww2040/PangTalrejaWhitt2007pub.pdf 21

[22] Stolletz R., "Performance Analysis and Optimization of Inbound Call Centers", Springer-Verlag Berlin Heidelberg, 2003. 3, 37, 38, 41

[23] Srinivasan R., Talim J. and Wang J., "Performance Analysis of a Call Center with Interacting Voice Response Units", TOP, Volume 12, pp. 91-110 June 2004. 4, 7, 8

[24] 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", July 2006, available at
http://iew3.technion.ac.il/serveng/References/references.html 27, 31

[25] de Véricourt F. and Jennings O.B., "Large-Scale Membership Services", Submitted to *Operations Research*, 2006. 17, 18

[26] Weerasinghe A. and Mandelbaum A., "Abandonment vs. Blocking in Many-Server Queues: Asymptotic Optimality in the QED Regime", working paper, 2008. 29

[27] Whitt W., "Understanding the efficiency of multi-server service systems", *Mgmt. Sc.*, 38, 708-723, 1992. 3

[28] Wolff R.W., "Stochastic modeling and the theory of queues", Prentice Hall, 1989. 14