

Staffing many-server queues with impatient customers: constraint satisfaction in call centers

Avishai Mandelbaum

Faculty of Industrial Engineering & Management,
Technion, Haifa 32000, Israel,
avim@tx.technion.ac.il

Sergey Zeltyn

IBM Research Lab, Haifa 31905, Israel,
sergeyz@il.ibm.com

August 15, 2008

Abstract

Motivated by call center practice, we study asymptotically optimal staffing of many-server queues with abandonment. A call center is modelled as an $M/M/n+G$ queue, which is characterized by Poisson arrivals, exponential service times, n servers and Generally distributed patience times of customers. Our asymptotic analysis is performed as the arrival rate, and hence the number of servers n , increase indefinitely.

We consider a constraint satisfaction problem, where one chooses the minimal staffing level n that adheres to a given cost constraint. The cost can incorporate the fraction abandoning, average wait and tail probabilities of wait. Depending on the cost, several operational regimes arise as asymptotically optimal: Efficiency-Driven (ED), Quality and Efficiency Driven (QED) and also a new ED+QED operational regime that enables QED tuning of the ED regime. Numerical experiments demonstrate that, over a wide range of system parameters, our approximations provide useful insight as well as excellent fit to exact optimal solutions. It turns out that the QED regime is preferable either for small-to-moderate call centers or for large call centers with relatively tight performance constraints. The other two regimes are more appropriate for large call centers with loose constraints.

We consider two versions of the constraint satisfaction problem. The first one is constraint satisfaction on a single time-interval, say one hour, which is common in practice. Of special interest is a constraint on the tail probability, in which case our new ED+QED staffing turns out asymptotically optimal. We also address a global constraint problem, say over a full day. Here several time intervals, say 24 hours, are considered, with interval-dependent staffing levels allowed; one seeks to minimize staffing levels, or more generally costs, given overall performance constraint. In this case, there is the added flexibility of trading service levels among time intervals, but we demonstrate that only little gain is associated with this flexibility if one is concerned with the fraction abandoning.

Subject classifications: queues: abandonment, limit theorems, optimization; call centers, staffing, workforce management, Halfin-Whitt (QED), ED+QED.

Area of review: Stochastic Models.

Contents

1	Introduction	1
1.1	Taking abandonment into account: M/M/n+G and Erlang-A	1
1.2	Performance measures and types of constraints	2
1.3	Main contributions	3
1.4	Structure of the paper	4
2	Asymptotic operational regimes	4
2.1	The QED (Quality and Efficiency-Driven) operational regime	5
2.2	The ED (Efficiency-Driven) operational regime	6
2.3	The ED+QED operational regime	6
2.4	Operational regimes and practical recommendations	7
2.5	Global constraint satisfaction	8
3	Related literature	10
4	Constraint satisfaction on a single interval	12
4.1	General formulation of the problem	12
4.2	QED	13
4.3	ED	15
4.4	ED+QED	16
5	Global constraint satisfaction	18
5.1	Global constraints in the QED regime	19
5.2	Global constraints in the ED Regime	21
6	Numerical examples	24
6.1	Constraint satisfaction in the QED regime	24
6.2	Constraint satisfaction in the ED regime	25
6.3	Constraint satisfaction in the ED+QED regime	26
6.4	Comparison between operational regimes	26
7	Possible future research	28

1 Introduction

During the last two decades, one observes an explosive growth in the number of companies that provide services via the telephone, as well as in the variety of telephone services provided. According to some estimates, worldwide expenditure on call centers exceeds \$300 billion [17] and the approximate number of call center agents reaches, for example, 4 million in the USA, 800 thousands in the UK and over 500 thousands in Canada [23].

A central challenge in designing and managing a service operation in general, and a call center in particular, is to achieve a desired balance between *operational efficiency* and *service quality*. Here we consider the staffing aspects of this problem, namely having the right number of agents in place.

“The right number” means, first of all, not too many, thus avoiding overstaffing. That is a crucial consideration since personnel costs (e.g. salaries of operators and spending on training) typically constitute about 70% of a call center’s expenditure.

“The right number”, however, also means not too few, thus avoiding understaffing and consequent poor service quality. Indeed, understaffing would imply excessive customers’ wait in tele-queues which is unpleasant in itself and, moreover, is likely to lead to abandonment of frustrated customers. (According to a Purdue University study [11], 63% of the customers name a negative call center experience as their main reason for stopping transactions with a company.)

One could consider the following two approaches to the quality/efficiency tradeoff. The first one is widely used in practice. A manager specifies performance constraint(s) and then assigns the least staffing level that satisfies these constraints, over a pre-determined time interval. In the second approach, one assigns revenues to service completions and costs to delay factors such as wait and abandonment, as well as to staffing. The goal is then to identify the staffing level that maximizes profit. Borst, Mandelbaum and Reiman [6] pursued both approaches in the context of classical queues *without* abandonment (M/M/n, or Erlang-C). In the present paper we focus on the constraint satisfaction problem, with customers’ abandonment taken into account. A subsequent paper [27] will deal with the cost or revenue optimization problem.

1.1 Taking abandonment into account: M/M/n+G and Erlang-A

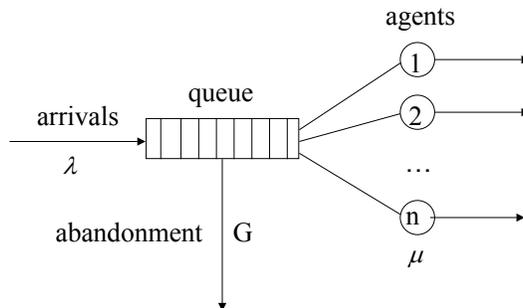
The M/M/n (Erlang-C) model was introduced by Erlang [12], the founder of queueing theory. It has been prevalent in call center applications for many years, being the mathematical engine of Workforce Management (WFM). Erlang-C assumes Poisson arrivals at a constant rate λ , exponentially distributed service times with rate μ , and n independent statistically-identical agents. Erlang-C implicitly assumes infinite patience of customers: all of them are willing to wait indefinitely until they get service (which implies system instability for $\frac{\lambda}{n\mu} \geq 1$).

However, an increasing number of present call centers incorporate customers’ abandonment in their staffing/scheduling software and performance goals, and rightly so: abandonment from tele-queues have a major impact on call center operations, which has lead to a growing body of research

on queues with impatient customers. (See Gans et al. [15] and Zeltyn and Mandelbaum [43] for surveys and references.) It is the goal of the present paper to contribute to this research, by developing theory that both supports staffing practice and enhances our understanding of it.

In our research, $M/M/n$ is replaced by $M/M/n+G$, assigning to each inbound call a generally distributed *patience time* τ with a common distribution G (Figure 1) and mean $\bar{\tau} \leq \infty$. An arriving customer encounters an *offered waiting time* V , defined as the time that this customer would have to wait given that her patience is infinite. If the offered wait exceeds the customer’s patience time, the call abandons, otherwise the customer eventually gets service. In both cases, the actual waiting time W is equal to $\min(V, \tau)$. Throughout the paper, we assume that service is rendered according to the First-Come-First-Served (FCFS) discipline.

Figure 1: **Schematic representation of the $M/M/n+G$ queue**



$M/M/n+G$ generalizes the $M/M/n+M$ (Erlang-A) model, first introduced in Palm [29], which has exponentially distributed patience times. Erlang-A is the most applicable (computationally tractable) model with abandonment (see [14] for a free software). And indeed, based on our experience, it is increasingly becoming the model of choice in support of WFM.

Remark 1.1 (Measuring waiting in the presence of abandonment) In applications, it is natural and convenient to measure delay in units of the average service time. Without abandonment, such an approach is also supported theoretically: fixing utilization and the number of servers, the average delay in $M/M/n$ is proportional to the average service time $E[S] = 1/\mu$. In contrast, in the presence of abandonment, delay is naturally measured relative to the average patience time $\bar{\tau}$. Indeed, the average delay does not exceed average patience, hence average delay in units of average patience is a unitless number in $(0, 1)$. Practically, however, average patience is not a quantity that managers internalize on a daily basis. To this end, we propose the following approximate rule-of-thumb, that is based on our call-centers experience: $\bar{\tau} \approx 2 \cdot E[S]$. Thus, in particular, measuring waiting in terms of $E[S]$ or $\bar{\tau}$ is essentially equivalent.

1.2 Performance measures and types of constraints

In order to apply a queueing model, one must first define relevant performance measures and then be able to calculate them. Moreover, since call centers can get very large (up to many thousands of

agents), the implementation of these calculations must be scalable (numerically stable).

In this research, we accommodate the following three performance measures: the fraction of abandoning customers $P\{\text{Ab}\}$, average wait in queue $E[W]$ and the tail probability to exceed a deadline $P\{W > T\}$. (An important special case of the tail probability is the delay probability $P\{W > 0\}$.)

The choice of the probability to abandon is natural in models with customers' impatience. In addition to the reasons explained above, abandonment statistics constitutes the only commonly available measurement that is customer-subjective: those who abandon reveal that the service offered is not worth its wait.

Average wait is also useful and taken into account in practice. Finally, the probabilities to exceed deadlines provide us with the distribution of customers' wait. This and similar performance measures are widely used. Indeed, $P\{W \leq T, \text{Sr}\}$, the fraction of *served* customers that wait less than T is often referred to in practice as *service level*.

Once an appropriate performance measure is chosen, a constraint satisfaction problem could be defined. The first approach pursued in this paper is one that is common in the practice of call centers: constraint satisfaction on *a single interval* (Section 4). Specifically, consider a time interval during which the staffing level is to be kept constants (15, 30 or 60 minutes), then specify a service-level constraint (for example, less than 3% of customers abandon) and finally find the minimal staffing level n^* that guarantees this desired service-level in steady state. An alternative approach, treated in Section 5, considers *jointly several time intervals*, for example a whole day, with different staffing levels allowed per interval. The goal now is to minimize the staffing levels or, more generally, staffing costs, given an overall constraint on the performance level. (For example, a constraint on the average wait over a full day of work which is divided into half-hour intervals.) In this case, an optimal solution is expected to compromise the service level at some intervals in order to do better at others.

1.3 Main contributions

As we view it, our main contributions to queuing theory and call center applications are as follows:

- Three operational regimes are studied within the M/M/n+G framework: Quality and Efficiency Driven (QED), Efficiency-Driven (ED) and also a new ED+QED operational regime; staffing in each regime arises as asymptotically optimal for a special case of cost constraints. Asymptotic statements for these regimes are formulated and proved in Theorems 4.1, 4.2 and 4.4, respectively.
- We elaborate on the new ED+QED operational regime: our discussion in Section 2.3 is supported by Theorem 4.3 in Section 4.4, on the asymptotic behavior of its major performance measures.
- Practical recommendations on applications of the three operational regimes and corresponding approximations are provided. Quality of the approximations in various realistic settings is

compared. For each type of setting, at least one operational regime gives rise to highly accurate approximations. Our practical recommendations are summarized in Section 2.4 and, then, substantiated and elaborated in Section 6.

- Motivated by practice, we study a global constraint problem, where several time intervals with interval-dependent staffing levels are considered jointly. Theorems 5.1 and Theorems 5.2 treat two special cases that give rise to global QED and ED staffing, respectively. In the case of QED staffing, a numerical experiment in Section 2.5 shows that staffing flexibility per interval provides only little gain with respect to the single-interval approach. In addition, Theorem 3.1 from the Online Appendix [28] demonstrates that, under some circumstances, the optimal solution combines QED staffing at some intervals with “no staffing” at the other intervals (very small or zero number of servers).
- Propositions 5.1 and 5.2 provide a detailed characterization of the asymptotic solution in the ED case, if the patience hazard rate is monotone, both for Decreasing Hazard Rate (DHR) and Increasing Hazard Rate (IHR).
- We extend the framework developed in Borst, Mandelbaum and Reiman [6] to cover abandonment and cost constraints. Our asymptotic analysis differs from [6], and it requires an extension of the Laplace method used in [43]. In this new framework, many additional interesting questions can be addressed, including optimal staffing with respect to multiple service constraints (see Example 2.1 in Section 2.5).

1.4 Structure of the paper

Sections 2.1-2.3 provide an introduction to the three operational regimes. Section 2.4 compares between approximations, based on these regimes, while providing practical recommendations for their use, and Section 2.5 introduces global constraints. Section 3 surveys some related literature. Sections 4 and 5 present our theoretical results for single-interval and global constraint satisfaction, respectively. Section 6 has several numerical examples. We conclude in Section 7 with some possible directions worthy of further research. Finally, the Online Appendix [28] includes additional theoretical material on global constraints, the proofs of local and global constraint results and an extensive numerical study.

2 Asymptotic operational regimes

How to solve the constraint satisfaction problem on a single interval? A straightforward approach is to apply exact formulae for performance measures of the M/M/n+G queue, developed in Baccelli and Hebuterne [2], Brandt and Brandt [8, 9] and Zeltyn and Mandelbaum [43]. However, this approach has several drawbacks. These formulae for performance measures are relatively complicated, involving double integration of the patience distribution. They provide no intuition and give rise to numerical

problems for large n (number of servers). In addition, the calculations require the whole patience distribution but its estimation is typically a very complicated, sometimes impossible, task (see Brown et al. [10]).

In this research, an alternative approach is pursued. Depending on the structure of the cost function, several operational (staffing) regimes arise as asymptotically optimal. Each regime corresponds to a different approximate solution of the constraint satisfaction problem. These approximations are theoretically validated for large systems but they also provide excellent fit for moderate and even small ones. The final outcome are regime-specific staffing rules that are highly useful for call centers management.

The operational regimes are described in terms of the *offered load* parameter R , which is defined as

$$R = \frac{\lambda}{\mu} = \lambda \cdot E[S];$$

here λ is the arrival rate and μ is the service rate, or the reciprocal of the average service time $E[S]$. (As customary in industry, we shall measure R in units of Erlangs.) The quantity R represents the amount of work, measured in time-units of service, that arrives to the system per unit of time. It is significant to the staffing problem since R and its neighborhood provide nominal staffing levels, deviations from which could result in extreme performance: staffing “high above” R would result in a very high quality of service, and staffing “far below” R would result in a very high utilization of servers.

An important goal of our paper is to have these last statements quantified. To this end, we now present three operational regimes that arise in our research as asymptotically optimal. We continue with comments on the quality of approximations based on these regimes and provide recommendations for their use. Finally, Section 2.5 gives a brief introduction to global constraint satisfaction.

2.1 The QED (Quality and Efficiency-Driven) operational regime

The QED regime corresponds to the least staffing level that adheres to the constraint $P\{W > 0\} \leq \alpha$, on the delay probability, given that α is neither too close to 0 nor to 1. It is characterized by the so-called *Square-Root Staffing Rule*:

$$n_{QED} = R + \beta\sqrt{R} + o(\sqrt{R}), \quad -\infty < \beta < \infty, \quad (2.1)$$

where β is a Quality-of-Service (QoS) parameter – the larger it is, the better is the operational service-level. (Throughout the paper, the notation $o(f(R))$ stands for $o(f(R))/f(R) \rightarrow 0$, as $R \rightarrow \infty$.) The QED regime enables one to combine high levels of efficiency (agents’ utilization over 90%) and service quality (short delays, scarce abandonment), given sufficient scale.

If we fix the service rate μ and patience distribution G and let λ and hence n converge to infinity according to (2.1), the delay probability $P\{W > 0\}$ converges to a constant strictly between 0 and 1. Also, the probability to abandon and average wait vanish at rate $\frac{1}{\sqrt{n}}$. (See Garnett et al. [16] and Zeltyn and Mandelbaum [43].)

In our paper we shall define the QED approximation for the optimal staffing level by

$$n_{QED}^* = \lceil R + \beta^* \sqrt{R} \rceil, \quad (2.2)$$

where the optimal QoS parameter β^* depends on performance goals and is calculated via equation (4.15) from Section 4.2. Numerical examples in Section 6 and the Online Appendix [28] demonstrate that these approximations turn out to be very accurate even for small to moderate-size systems (few 10's of agents).

2.2 The ED (Efficiency-Driven) operational regime

The ED regime corresponds to the least staffing level that adheres to a constraint on the fraction abandoning, or on the average wait, given that the constraints are relatively loose; for example, the former is to be above 10%, and the latter in the order of average service time. The ED regime is characterized by the staffing rule

$$n_{ED} = (1 - \gamma) \cdot R + o(R), \quad \gamma > 0, \quad (2.3)$$

which implies understaffing with respect to the offered load. In this case, as shown in [43], with n and λ increasing indefinitely, virtually all customers wait ($P\{W > 0\}$ converges to 1), the probability to abandon always converges to γ and average wait converges to a constant that depends on the patience distribution. As a rule, ED approximations require relatively large n (more than 100) in order to provide a satisfactory fit; see the numerical experiments in [43, 44].

Our ED approximation for the optimal staffing level is defined by

$$n_{ED}^* = \lceil (1 - \gamma^*) \cdot R \rceil, \quad (2.4)$$

where the value of γ^* is established via equation (4.20) from Theorem 4.2.

2.3 The ED+QED operational regime

The following new ED+QED regime corresponds to the least staffing level that adheres to the constraint $P\{W > T\} \leq \alpha$, on the tail probability of delay, given that T is in the order of a mean service time (or mean patience, see Remark 1.1) and α is neither too close to 0 nor to 1.

Why is a new staffing regime needed? Assume that we vary the number of servers according to the ED staffing rule (2.3), holding other system parameters constant. It follows from [43] that there exists an ED parameter γ^* such that the tail probability $P\{W > T\}$ converges to zero for $\gamma < \gamma^*$ and to $1 - G(T)$ for $\gamma > \gamma^*$. The ED approximation is thus too “crude” for the constraint satisfaction problem $P\{W > T\} \leq \alpha$, $0 < \alpha < 1 - G(T)$. However, it turns out that QED fine-tuning around the ED staffing level $(1 - \gamma^*)R$, taking into account α , provides one with the staffing level that asymptotically adheres to $P\{W > T\} \sim \alpha$.

Formally, the ED+QED regime is characterized by the staffing rule:

$$n_{ED+QED} = (1 - \gamma) \cdot R + \delta \sqrt{R} + o(\sqrt{R}), \quad \gamma > 0. \quad (2.5)$$

See Section 4.4 and Theorems 4.3 and 4.4 for rigorous results on the ED+QED regime and further clarifications. The approximate ED+QED staffing formula is

$$n_{ED+QED}^* = \lceil (1 - \gamma^*) \cdot R + \delta^* \sqrt{R} \rceil, \quad (2.6)$$

with values of γ^* and δ^* derived from Theorem 4.4.

Remark 2.1 (ED+QED regime in Erlang-A) Assume that the patience times are exponential with rate θ , namely mean $1/\theta$. (According to our experience, typical mean patience is around 5-10 minutes.) Fix the deadline to be $c\%$ of the patience mean: $T = c/\theta$, for some $0 < c < 1$. Then Theorem 4.4 implies that the ED parameter is equal to

$$\gamma = G(T) = 1 - e^{-\theta T} = 1 - e^{-c} \approx c,$$

where the last approximation is reasonable for c around 0.1. In this case, according to the ED approximation

$$P\{\text{Ab}\} \approx \gamma \approx c.$$

This argument provides us with a useful rule-of-thumb for Erlang-A: a deadline T of around 10% of the average patience (namely T between 30 seconds and 1 minute) corresponds to approximately 10% abandonment.

Remark 2.2 (Quality-Driven regime) Zeltyn and Mandelbaum [43] also analyzed the QD (Quality-Driven) operational regime, with staffing levels

$$n_{QD} = \lceil (1 + \gamma) \cdot R + o(\sqrt{R}) \rceil, \quad \gamma > 0.$$

In this regime, the main performance characteristics, namely $P\{\text{Ab}\}$, $E[W]$ and $P\{W > 0\}$, converge to zero exponentially fast, as λ , $n \rightarrow \infty$. A similar regime was discussed also in Borst et al. [6] for Erlang-C. The QD regime can become relevant for extreme constraints on service level, say $P\{W > 0\} \leq 2\%$ (essentially no one waits), which are appropriate, for example, in amply-staffed emergency operations. In this paper, we are interested in queues with non-negligible wait and abandonment, hence the QD regime will not be considered.

2.4 Operational regimes and practical recommendations

In Sections 2.1-2.3 we introduced our three operational regimes. Note that, given a specific constraint satisfaction problem, one can fit to it several different approximations, based on these regimes. A natural question then arises as to the existence of a single operational regime that is preferable over the two others, at least for all constraint satisfaction problems of practical interest. For Erlang-C, Borst et al. [6] discovered that the QED regime qualifies: it is extremely robust and can be applied over a very wide range of system parameters so as to render the ED and QD regimes almost practically useless. In Section 6 and the Online Appendix [28] we analyze this question for M/M/n+G. Here, it turns out that, in contrast to Erlang-C, there is no single best operational regime.

In Table 1, we summarize our guidelines for the use of the three operational regimes, as providing approximations for the following operational performance measures: $P\{Ab\}$, $E[W]$, $P\{W > T\}$, $T > 0$. We observe that in most special cases considered in Table 1, the QED approximations are preferable over the alternatives. However, and in contrast to Erlang-C [6], QED staffing turns out far from optimal for loose constraints, if one controls either *average wait* (in the case of non-exponential patience) or the *tail probability of wait* in moderate-to-large call centers. (See Example 6.2 in Section 6.4.) Note that, in both cases, the optimal staffing level is significantly smaller than the offered load.

Table 1: **Summary of practical recommendations.**

Constraint	P{Ab}		E[W]		P{W > T}	
	Tight 1-10%	Loose $\geq 10\%$	Tight $\leq 10\%\bar{\tau}$	Loose $\geq 10\%\bar{\tau}$	Tight $0 \leq T \leq 10\%\bar{\tau}$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\bar{\tau}$ $5\% \leq \alpha \leq 50\%$
Offered Load						
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \stackrel{d}{=} \exp$	QED	ED+QED

Comments on Table 1.

- In Table 1 we measure waiting time in units of the mean patience $\bar{\tau}$. This is natural since $\bar{\tau}$ is a tight upper bound for $E[W]$, hence one measures delay relative to $\bar{\tau}$.
- As we mentioned in Section 1.2, the constraint on the delay probability, $P\{W > 0\} \leq \alpha$, is relevant in many applications, including call centers. We treat it as a special case of a tight tail-probability constraint ($T = 0$). According to numerical experiments in the Online Appendix [28], QED approximations for $P\{W > 0\}$ are excellent for a wide range of α (10-90%).

2.5 Global constraint satisfaction

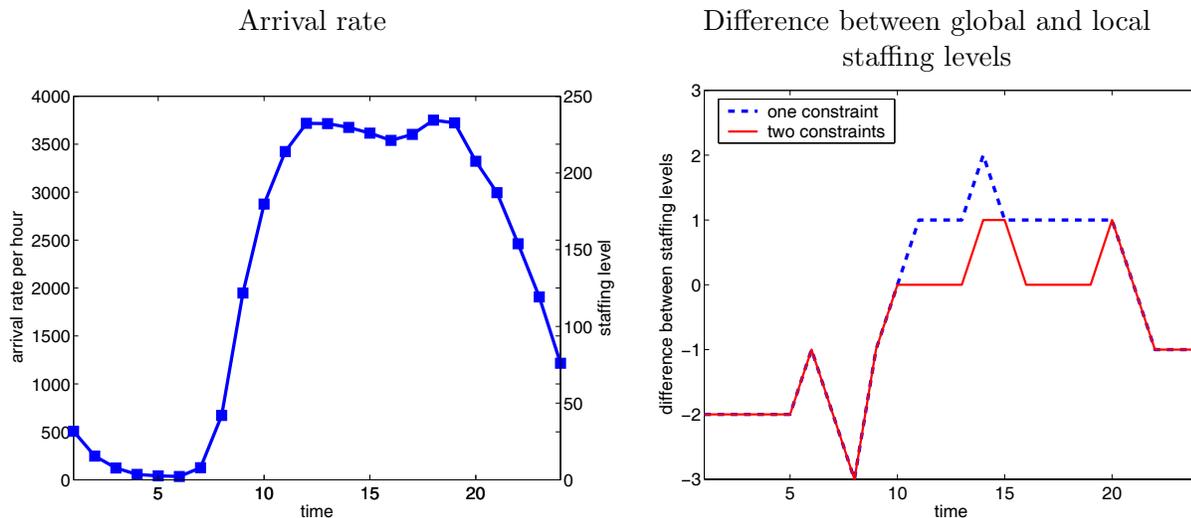
In Sections 2.1-2.4 we described our approach for constraint satisfaction on a single time-interval. In Section 5 we shall study approximate solutions of some global constraint problems where one seeks to minimize staffing costs over several time intervals, with differing staffing levels allowed per interval. If the staffing costs are equal at all intervals, the problem reduces to minimizing the total staffing level. We then expect a smaller overall number of servers relative to interval-by-interval optimization, due to the added flexibility of trading service levels among intervals. The following example checks the impact of this flexibility in a realistic call-center setting.

Example 2.1 (Comparison between global and local constraint satisfaction) Consider a week-day arrival pattern to the call center of some Israeli telecom company (left plot in Figure 2). This pattern is rather typical of call centers; see for example Brown et al. [10] and Green et al. [20]. Assume constant arrival rates during each one-hour interval. The average service time is equal to 218

seconds. Mean patience is taken to be 6 minutes (a reasonable number, according to our experience; see [10], for example). Assume that both service and patience times are exponential.

Consider the constraint $P\{Ab\} \leq 1\%$. We compare QED staffing (2.2) that seeks to sustain this service level over each one-hour interval, against QED staffing that adheres to a global daily (24 hours) service level. The latter staffing levels are derived via Theorem 5.1.

Figure 2: **Comparison between local and global constraints.**



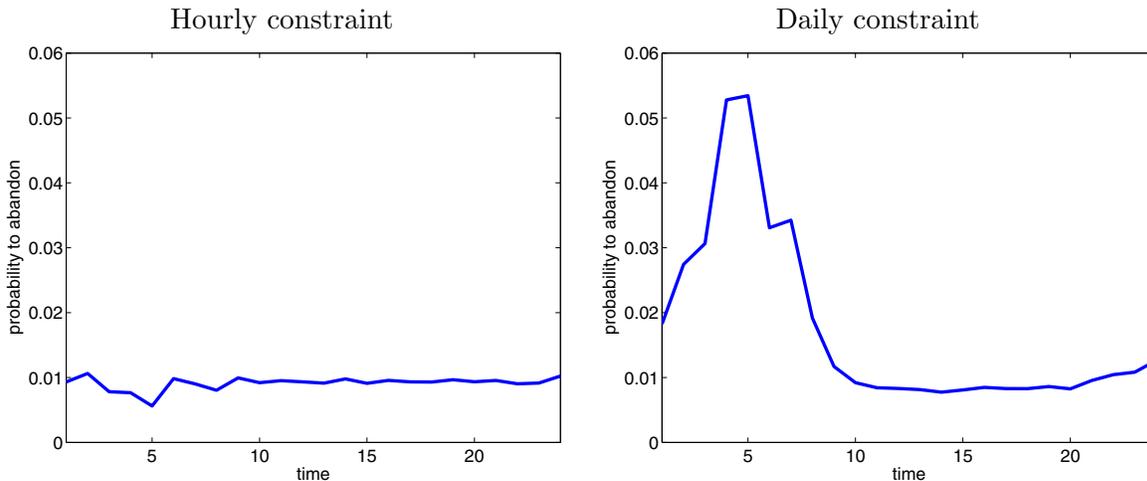
Overall, the two staffing levels are rather close. (Their pattern is also indistinguishable from the pattern of the arrival rate in Figure 2, hence it has been omitted.) The “one constraint” curve in the right plot of Figure 2 presents their difference: global staffing assigns more servers to heavily loaded intervals and less to lightly loaded. Overall, using global constraint saves only 9 work-hours (out of more than 3,000 hours over the day). Now assume that, in addition, a second constraint on the delay probability $P\{W > 0\} \leq 30\%$ should be satisfied. In the global constraint case, the same staffing as for a single global constraint $P\{Ab\} \leq 1\%$ is sufficient. However, if two local constraints should be satisfied at each interval, additional servers should be added at some intervals and the difference between the two schedules is 17 work-hours. (See the “two constraint” curve in the right plot of Figure 2.) Hence, even in the case of two constraints (which is common practice in call centers) global staffing does not add much with respect to single-interval staffing.

Figure 3 demonstrates the dynamics of $P\{Ab\}$ over the day for hourly and daily $P\{Ab\} \leq 1\%$ constraints. ($P\{Ab\}$ is calculated via exact M/M/n+G formulae [43].) As expected, we observe a stable pattern in the first plot. In the second plot, however, a deterioration of the service level at night takes place (which, indeed, conforms to our personal experience with the way that many 24-hour call centers are run). Since minor savings in the number-of-agents probably do not justify instability of the service level, the staffing levels that arise from the hourly constraint seem preferable.

Remark 2.3 (Global vs. single-interval constraint satisfaction) The above example seems important to us for the following reason. Global constraint satisfaction is a standard approach to Service

Level Agreements in call centers and other service organizations. In these organizations, performance constraints are usually enforced over long time-intervals (days, weeks or even months) and not over the basic staffing intervals with constant staffing level (half-hour or an hour). However, it is much simpler, both mathematically and computationally, to calculate optimal staffing and scheduling separately for each basic staffing interval. Example 2.1 demonstrates that if a tight constraint on the probability to abandon is considered, the latter method suffices to provide staffing that is very close to the optimal staffing for global constraints. Note that other types of constraints can lead to a different type of results. For example, if a single constraint on the delay probability is considered, Theorem 3.1 from the Online Supplement implies that the global solution (combination of “no staffing” and QED staffing at different intervals) can result in significant workforce savings with respect to the local QED optimal staffing.

Figure 3: **Dynamics of the probability to abandon.**



3 Related literature

For a comprehensive summary on queues with impatient customers, operational regimes and dimensioning, readers are referred to Sections 4.1 and 4.2 in Gans et al. [15]. Here we summarize research that is most relevant to ours.

Queues with impatient customers. The seminal work on queueing systems with impatient customers is Palm [29], where he introduced the basic Erlang-A ($M/M/n+M$) queue with exponential patience times. See Mandelbaum and Zeltyn [26] for a recent summary of this model. Erlang-A was generalized to $M/M/n+G$ (General patience) by Baccelli and Hebuterne [2], Brandt and Brandt [8, 9] and Zeltyn and Mandelbaum [43]. In the present research we adopt the theoretical approach of [2] and [43] to the $M/M/n+G$ model.

If the service distribution is not exponential, as is often the case in practice (see Brown et al. [10]), exact theoretical solutions are not available and one has to resort to approximations and simulation.

In addition to ED approximations, as discussed below, one should mention the papers of Whitt [40, 41] that develop and validate an approximation for the M/G/n+G model with generally distributed i.i.d. service times. Finally, Boxma and de Waal [7] addressed the problem of cost optimization in the M/G/n+G queue via simulation and interpolation between M/M/n+D and Erlang-A.

Remark 3.1 (Types of approximations) Below we survey several types of approximations to the queueing systems that arise in our research. One distinguishes between two main types of approximations: *steady-state* (asymptotic expressions for steady-state performance measures like $P\{\text{Ab}\}$ or $E[W]$) and *process-limit* (asymptotics for stochastic processes such as the queue-length process). In our paper, we are mainly interested in steady-state approximations. However, many papers referred to below present also process-limit approximations.

QED operational regime. As mentioned above, the square-root staffing rule (2.1) was first introduced by Erlang [12]. He reports that it had in fact been in use at the Copenhagen Telephone Company since 1913. A formal analysis for the Erlang-C queue appeared only in 1981, in the seminal paper of Halfin and Whitt [21]. In this paper, the authors establish an important relation: as λ increases indefinitely, sustaining the QED operational regime (2.1) with fixed $\beta > 0$ is equivalent to the delay probability converging to a fixed level α , $0 < \alpha < 1$. Whitt [36] surveys QED approximations for various classical queues without abandonment.

Garnett, Mandelbaum and Reiman [16] studied the QED regime for Erlang-A with exponential abandonment, establishing results that are analogous to [21] and complemented also by the ED and QD regimes. Zeltyn and Mandelbaum [43] presented a comprehensive treatment of the QED, ED and QD regimes in steady-state for the M/M/n+G queue.

ED operational regime. ED approximations are cruder than the QED approximations, hence they enable the analysis of very general models. For example, Whitt [37] presents a general fluid model (the ED approximation) for the G/G/n+G queue with general distributions of arrivals, services and patience times and Whitt [42] presents a multi-class fluid model that takes skills-based routing into account.

Another important family of models that can be treated in the ED regime are those with uncertainty about the arrival-rate. (Whitt [38] showed that Erlang-A and other queues with abandonment are sensitive to changes in the arrival rate.) Recent papers of Whitt [39] and Harrison and Zeevi [22] study ED approximations for such models and develop asymptotic rules for optimal staffing. In addition, Bassamboo, Harrison and Zeevi [4, 5] provide asymptotic methods of routing and admission control.

ED+QED operational regime. The only reference to this regime that is known to us is Baron and Milner [3]. That work is motivated by Service Level Agreements that arise in outsourcing contracts. It includes an approximation for the tail probability of wait in Erlang-A. This approximation

is a special case of our approximation for M/M/n+G, which is covered in Theorem 4.4.

It is worth mentioning that the QED staffing rule that arose in models of membership (subscriber) services is (only formally) similar to our ED+QED staffing rule (2.5); see Randhawa and Kumar [30, 31] and de Véricourt and Jennings [34].

Dimensioning Erlang-C: cost optimization and constraint satisfaction. Borst, Mandelbaum and Reiman [6] developed a mathematical framework for the problem of optimal staffing in the Erlang-C queue. The main focus of the paper is on cost optimization with convex staffing costs and general increasing waiting costs. Depending on the relative importance of these costs, [6] identifies the QED, ED and QD regimes as asymptotically optimal. It is shown that the QED regime balances and, in fact, unifies the other two regimes. In the case of linear costs, a relation between the waiting/staffing costs ratio and the QoS parameter β from (2.1) is established. In addition, the constraint satisfaction problem is also analyzed, with the QED, ED and QD regimes arising as well.

Global constraint satisfaction. In [6], the problem of optimal staffing was studied on a single interval in steady-state, as conventionally assumed in the literature and practice. Koole and van der Sluis [24] analyze a shift scheduling problem with overall, say daily, service-level objective. They prove a useful property, called multimodularity, which, if prevailing, significantly facilitates the search for the exact optimal staffing levels.

SIPP staffing. Both the single- and multiple-interval approaches can be viewed as the SIPP (Stationary Independent Period by Period) method for staffing; see Green, Kolesar and Soares [18]. Advantages, drawbacks and possible modifications of SIPP were studied in Green, Kolesar and Soares [18, 19] and Green, Kolesar and Whitt [20]. Overall, SIPP works well if the arrival rate is slow-varying with respect to the durations of services. Otherwise, time-dependent models should be used; see [18, 19, 20] and Feldman et al. [13].

4 Constraint satisfaction on a single interval

4.1 General formulation of the problem

We analyze the M/M/n+G queue with a fixed service rate μ and patience distribution G . We determine *staffing level* (number of agents) n_λ , $\lambda \geq 0$, as a function of the arrival rate λ , focusing on large λ (formally $\lambda \rightarrow \infty$).

Define a *performance cost* function $U(n, \lambda)$ as a weighted sum of the three performance measures that were introduced in Section 1.2:

$$U(n, \lambda) = C_{ab}(\lambda) \cdot P_{n,\lambda}\{\text{Ab}\} + C_w(\lambda) \cdot E_{n,\lambda}[W] + C_b(\lambda) \cdot P_{n,\lambda}\{W > d_\lambda\}. \quad (4.1)$$

(The function $U(n, \lambda)$ is to be interpreted as the performance cost per arrival.) The coefficients $C_{ab}(\lambda)$, $C_w(\lambda)$ and $C_b(\lambda)$ in (4.1) are abandonment cost, waiting cost and deadline cost, respectively.

Note that all these coefficients and also the deadline d_λ can depend on the arrival rate λ . We introduce such dependence since scaling with λ , as will be demonstrated, gives rise to very accurate approximations that are applicable to practical problems with unscaled cost coefficients. See Remark 4.3 that illustrates the relation between the unscaled and scaled problems.

Define the *optimal staffing level* by

$$n_\lambda^* = \arg \min_n \{U(n, \lambda) \leq M\}, \quad (4.2)$$

where M is a cost constraint per arrival.

Since we study different types of asymptotic solutions to (4.2), the following two additional definitions turn out natural and useful. The staffing level n_λ is called *asymptotically feasible* if

$$\limsup_\lambda U(n_\lambda, \lambda) \leq M. \quad (4.3)$$

In addition, n_λ is *asymptotically optimal* if

$$|n_\lambda^* - n_\lambda| = o(f(\lambda)) = o(f(R)), \quad (4.4)$$

where a specific function for $f(\cdot)$ will be defined separately and naturally for every special case. (For example, it can be equal to λ , $\sqrt{\lambda}$ etc.) In words, asymptotic feasibility guarantees that the asymptotic performance cost does not exceed the cost constraint and asymptotic optimality implies that the staffing level is close to the optimal one.

For general $C_{ab}(\cdot)$, $C_w(\cdot)$ or $C_b(\cdot)$, the problem (4.2) is rather complicated and un insightful. Instead, we explore several basic important special cases that give rise to the three operational regimes introduced in Section 2.

Remark 4.1 As already mentioned, in this research, the staffing level n depends on the arrival rate λ and, consequently, performance measures under consideration depend on λ as well. For simplicity of notation, in the rest of the paper (except some proofs in the Online Appendix [28]) we shall omit indices that correspond to λ and n .

4.2 QED

Here we explore types of constraints that give rise to the QED regime, discussed in Section 2.1. Recall that this regime allows us to combine efficiency (high servers' utilization) and service quality.

Assume that the patience density at the origin exists and is positive: $G'(0) \triangleq g_0 > 0$. Define the following functions:

$$P_w(\beta) \triangleq \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h_\phi(\hat{\beta})}{h_\phi(-\beta)} \right]^{-1}, \quad -\infty < \beta < \infty, \quad (4.5)$$

$$P_a(\beta) \triangleq \sqrt{g_0} \cdot (h_\phi(\hat{\beta}) - \hat{\beta}), \quad -\infty < \beta < \infty, \quad (4.6)$$

$$W(\beta, t) \triangleq \frac{\bar{\Phi}(\hat{\beta} + \sqrt{g_0} \cdot t)}{\bar{\Phi}(\hat{\beta})}, \quad -\infty < \beta < \infty, \quad t \geq 0, \quad (4.7)$$

where

$$\hat{\beta} \triangleq \beta \sqrt{\frac{\mu}{g_0}}, \quad (4.8)$$

and

$$h_\phi(x) = \frac{\phi(x)}{1 - \Phi(x)} = \frac{\phi(x)}{\bar{\Phi}(x)}$$

is the hazard rate of the standard normal distribution. ($\Phi(x)$ is the standard normal cumulative distribution function, $\bar{\Phi}(x) = 1 - \Phi(x)$ is the survival function and $\phi(x) = \Phi'(x)$ is the density.)

Theorem 4.1 from Zeltyn and Mandelbaum [43] implies that under the QED scaling, as defined by the square-root staffing rule (2.1) with $\lambda \rightarrow \infty$ (or $R = \lambda/\mu \rightarrow \infty$), the following approximations prevail:

$$\mathbb{P}\{W > 0\} \sim P_w(\beta), \quad (4.9)$$

$$\mathbb{P}\{\text{Ab}\} \sim \frac{1}{\sqrt{\lambda}} \cdot P_a(\beta) P_w(\beta), \quad (4.10)$$

$$\mathbb{E}[W] \sim \frac{1}{\sqrt{\lambda}} \cdot \frac{1}{g_0} \cdot P_a(\beta) P_w(\beta), \quad (4.11)$$

$$\mathbb{P}\left\{W > \frac{t}{\sqrt{\lambda}}\right\} \sim W(\beta, t) \cdot P_w(\beta), \quad t \geq 0. \quad (4.12)$$

(Here and later $f \sim g$ stands for $\lim_{\lambda \rightarrow \infty} f(\lambda)/g(\lambda) = 1$.)

Theorem 4.1 (QED) Assume that the cost function (4.1) is given by

$$U(n, \lambda) = C_{ab} \cdot \sqrt{\lambda} \cdot \mathbb{P}\{\text{Ab}\} + C_w \cdot \sqrt{\lambda} \cdot \mathbb{E}[W] + C_b \cdot \mathbb{P}\left\{W > \frac{t}{\sqrt{\lambda}}\right\}, \quad (4.13)$$

where the constants C_{ab}, C_w, C_b, t are non-negative. Assume that either $C_{ab} > 0$ or $C_w > 0$, or the cost constraint $M < C_b$.

a. The optimal staffing level (4.2) satisfies

$$n^* = R + \beta^* \sqrt{R} + o(\sqrt{R}), \quad (4.14)$$

where β^* is the unique solution of the following equation with respect to β :

$$\left\{ \left(C_{ab} + \frac{C_w}{g_0} \right) \cdot P_a(\beta) + C_b \cdot W(\beta, t) \right\} \cdot P_w(\beta) = M. \quad (4.15)$$

b. Introduce the staffing level

$$n_{QED}^* = \left\lceil R + \beta^* \sqrt{R} \right\rceil. \quad (4.16)$$

Then the staffing level n_{QED}^* is asymptotically feasible (4.3) and asymptotically optimal (4.4) in the sense that

$$|n_{QED}^* - n^*| = o(\sqrt{R}). \quad \blacksquare$$

See Section 2.1 in the Online Appendix [28] for the proof of Theorem 4.1.

Remark 4.2 (On the solution of the equation (4.15)) As it will be shown in the proof of Theorem 4.1, all functions in (4.15) are monotone. Therefore, its solution is tantamount to the calculation of the inverse of the left-hand side.

Remark 4.3 (On $\sqrt{\lambda}$ scaling of the probability to abandon and waiting time) In Theorem 4.1 we scale some parameters of the problem by $\sqrt{\lambda}$. In call center practice, on the other hand, one works with unscaled constraints. The translation between the two types of the constraints is straightforward. Assume, for example, that the service-level constraint is given by $P\{\text{Ab}\} \leq \epsilon$. According to (4.10), we should first solve

$$P_a(\beta)P_w(\beta) = \epsilon\sqrt{\lambda}$$

with respect to β and then apply the staffing level (4.16) with the solution β^* .

4.3 ED

Here we study the ED regime defined in (2.3). It is characterized by significant understaffing that gives rise to very high servers' utilization and only moderate service level.

Assume that the patience distribution function G is strictly increasing for all x such that $0 < G(x) < 1$. Define

$$H(x) = E[\min(\tau, x)] = \int_0^x \bar{G}(u)du, \quad x \geq 0, \quad (4.17)$$

where $\bar{G}(\cdot) = 1 - G(\cdot)$ is the survival function of patience.

Theorem 6.1 from Zeltyn and Mandelbaum [43] states that in the ED operational regime, defined by (2.3) and $\lambda \rightarrow \infty$, the probability to abandon converges to γ and the ED approximation for average wait is given by:

$$E[W] \sim H(G^{-1}(\gamma)). \quad (4.18)$$

Theorem 4.2 (ED) Assume that the performance cost function (4.1) is given by

$$U(n, \lambda) = C_{ab} \cdot P\{\text{Ab}\} + C_w \cdot E[W], \quad (4.19)$$

where the constants C_{ab}, C_w are non-negative. Assume that the positive cost constraint M is less than $C_{ab} + C_w \cdot E[\tau]$, where $E[\tau] = \infty$ is allowed. Then

a. The optimal staffing level (4.2) satisfies

$$n^* = (1 - \gamma^*) \cdot R + o(R),$$

where γ^* is the unique solution of the following equation with respect to γ :

$$C_{ab} \cdot \gamma + C_w \cdot H(G^{-1}(\gamma)) = M. \quad (4.20)$$

b. Introduce the staffing level

$$n_{ED}^* = \lceil (1 - \gamma^*) \cdot R \rceil .$$

Then n_{ED}^* is asymptotically feasible (4.3) and asymptotically optimal (4.4) in the sense that

$$|n_{ED}^* - n^*| = o(R) . \quad \blacksquare$$

See Section 2.2 in the Online Appendix [28] for the proof of Theorem 4.2.

Remark 4.4 With each one of the cost coefficients in (4.19) vanishing, we get two important special cases: a constraint on $P\{Ab\}$ and a constraint on $E[W]$. Note that, in the first case, $\gamma = M/C_{ab}$ and the ED staffing n_{ED}^* does not depend on the patience distribution. In contrast, in the second case one can observe a very significant dependence of n_{ED}^* on G ; see the example in Section 6.2 and in Section 3.2 of the Online Appendix [28].

4.4 ED+QED

Now we study the new operational regime, introduced in Section 2.3. This regime combines ED and QED staffing, which arises when one seeks to satisfy a constraint on the tail probability of wait, namely

$$P\{W > T\} \leq \alpha . \quad (4.21)$$

First, we recall from [43] that in the ED regime (2.3), the waiting time converges weakly to $\min(\tau, G^{-1}(\gamma))$, where τ is the patience time and G is the patience distribution. This suggests the following approximation for the tail probability $P\{W > T\}$:

$$P\{W > T\} \approx \begin{cases} \bar{G}(T), & T < G^{-1}(\gamma), \\ 0, & T > G^{-1}(\gamma). \end{cases} \quad (4.22)$$

In (4.22) we assume that γ is fixed and T is varied. But to identify the least staffing level that adheres to (4.21), we view (4.22) as a function of γ :

$$P\{W > T\} \approx \begin{cases} \bar{G}(T), & \gamma > G(T), \\ 0, & \gamma < G(T), \end{cases} \quad (4.23)$$

which is too crude to capture α in (4.21). Hence one must refine (4.23) around $\gamma = G(T)$. To this end, introduce ED+QED staffing with the ED parameter $\gamma = G(T)$ as follows:

$$n = (1 - G(T)) \cdot R + \delta \sqrt{R} + o(\sqrt{R}), \quad -\infty < \delta < \infty . \quad (4.24)$$

The next theorem enables the calculation of δ that corresponds to the target level α of the tail probability. It also presents approximations for other key performance measures in the ED+QED regime. The theorem is formulated in the spirit of the M/M/ n statement in Halfin and Whitt [21].

Theorem 4.3 (ED+QED performance measures) Consider a sequence of M/M/n+G queues indexed by $n = 1, 2, \dots$, with fixed service rate μ and patience distribution G . Let T and α be scalars such that $0 < T < \infty$, $0 < \alpha < \bar{G}(T)$ and the patience density $g(T) = G'(T) > 0$. Then the following four asymptotic statements are equivalent, as $n \rightarrow \infty$ (and hence $\lambda \rightarrow \infty$ and $R \rightarrow \infty$):

1. **Staffing level:** $n = (1 - \gamma)R + \delta\sqrt{R} + o(\sqrt{R})$;
2. **Tail probability:** $P\{W > T\} = \alpha + o(1)$;
3. **Probability to abandon:** $P\{\text{Ab}\} = \gamma - \frac{\delta}{\sqrt{R}} + o\left(\frac{1}{\sqrt{R}}\right)$;
4. **Average wait:** $E[W] = \int_0^T \bar{G}(u)du - \frac{\delta}{\sqrt{R}} \cdot \frac{1}{h_G(T)} + o\left(\frac{1}{\sqrt{R}}\right)$.

Here $h_G(T) = g(T)/\bar{G}(T)$ is the hazard rate of the patience distribution G , $\gamma = G(T)$ and

$$\delta = \bar{\Phi}^{-1}\left(\frac{\alpha}{\bar{G}(T)}\right) \cdot \sqrt{\frac{g(T)}{\mu}}. \quad (4.25)$$

Remark 4.5 Note that (4.25) and Statement 2 of Theorem 4.3 imply the following approximation for the delay probability under the ED+QED staffing level (4.24):

$$P\{W > T\} \sim \bar{G}(T) \cdot \bar{\Phi}\left(\delta\sqrt{\frac{\mu}{g(T)}}\right).$$

Remark 4.6 If the constraint parameter $\alpha \geq \bar{G}(T)$, then for any staffing level n

$$P\{W > T\} \leq \bar{G}(T) \leq \alpha, \quad (4.26)$$

since the waiting time W does not exceed the patience time τ . Hence, $\alpha > \bar{G}(T)$ cannot be attained as a limit in Part 2, and $\alpha = \bar{G}(T)$ can be attained even if $n = 0$, namely service is not provided at all.

Now we can formulate the constraint satisfaction result for the ED+QED regime.

Theorem 4.4 (ED+QED constraint satisfaction) Assume that the cost function (4.1) is given by

$$U(n, \lambda) = C_b \cdot P\{W > T\}, \quad C_b > 0, T > 0, \quad (4.27)$$

and that the patience density at T is positive: $g(T) > 0$. The optimization problem (4.2) then takes the form

$$n^* = \arg \min_n \{P\{W > T\} \leq \alpha\},$$

where $\alpha \triangleq M/C_b$. Assume that $\alpha < \bar{G}(T)$. Then

a. The optimal staffing level (4.2) satisfies

$$n^* = \bar{G}(T) \cdot R + \delta^* \sqrt{R} + o(\sqrt{R}), \quad (4.28)$$

where

$$\delta^* = \bar{\Phi}^{-1} \left(\frac{\alpha}{\bar{G}(T)} \right) \cdot \sqrt{\frac{g(T)}{\mu}}. \quad (4.29)$$

b. Introduce the staffing level

$$n_{ED+QED}^* = \left\lceil \bar{G}(T) \cdot R + \delta^* \sqrt{R} \right\rceil.$$

Then n_{ED+QED}^* is asymptotically feasible (4.3) and asymptotically optimal (4.4) in the sense that

$$|n_{ED+QED}^* - n^*| = o(\sqrt{R}). \quad \blacksquare$$

See Section 2.3 in the Online Appendix [28] for the proofs of Theorems 4.3 and 4.4.

Remark 4.7 In continuation to Remark 4.6, if the constraint parameter $\alpha \geq \bar{G}(T)$, then the optimal staffing $n^* = 0$. ■

Remark 4.8 Note that the ED coefficient $\bar{G}(T)$ in (4.28) depends on the patience distribution and does not depend on the constraint α . The QED coefficient δ^* provides fine tuning for varying values of the constraint. ■

Finally, recall that Table 1 in Section 2.4 provides guidelines to the practitioners as to which operating regime to use, depending, say, on the service-level agreement and the time scales involved. See also Section 6 for numerical examples on the three regimes in the single-interval case.

5 Global constraint satisfaction

Here we study the global constraint satisfaction problem that was introduced in Section 2.5. We make the same assumptions as in Sections 4.1 and 4.2: the service rate μ and patience distribution G are fixed. However, now we consider a set of K time intervals that constitute a day of work. Arrivals to each interval are governed by a Poisson process with rates $r_i \lambda$, $1 \leq i \leq K$, $\sum_{i=1}^K r_i = 1$. Here we assume that these intervals are of the same length. (This assumption could be relaxed at the cost of more complicated notation.) Then r_i can be interpreted as the fractions of the daily arrival rate during interval i . The staffing costs at the intervals are given by c_i , $1 \leq i \leq K$.

The vector of staffing levels is determined according to the overall arrival rate and is denoted by

$$\bar{n}(\lambda) \triangleq (n_1(\lambda), \dots, n_K(\lambda)),$$

where $n_i(\lambda)$ is the staffing level during interval i . More precisely, we shall let $\lambda \rightarrow \infty$ while maintaining the fractions (r_1, \dots, r_k) fixed.

Introduce the performance cost function

$$U(\bar{n}, \lambda) = \sum_{i=1}^K r_i U_i(n_i, \lambda),$$

where each $U_i(n_i, \lambda)$ is calculated as in (4.1). Note that the performance costs per interval are weighted according to the arrival rates on these intervals.

Now modify definition (4.2) of the optimal staffing level $\bar{n}^*(\lambda) = (n_1^*(\lambda), \dots, n_K^*(\lambda))$ into

$$\bar{n}^*(\lambda) = \arg \min \sum_{i=1}^K c_i n_i(\lambda), \quad \text{subject to } U(\bar{n}, \lambda) \leq M.$$

The notion (4.4) of asymptotically optimal staffing level $\bar{n}(\lambda) = (n_1(\lambda), \dots, n_K(\lambda))$ changes to

$$\left| \sum_{i=1}^K c_i n_i(\lambda) - \sum_{i=1}^K c_i n_i^*(\lambda) \right| = o(f(\lambda)) = o(f(R)).$$

Finally, definition (4.3) of asymptotic feasibility is unchanged.

Below we present several special cases of global constraints that give rise to different operational regimes. Two cases will be treated in detail: a scaled constraint on the probability to abandon, in Section 5.1 (QED regime) and a constraint on the average wait, in Section 5.2 (ED regime). Several other special cases will be reviewed briefly.

5.1 Global constraints in the QED regime

Assume that the performance cost function at interval i , $1 \leq i \leq K$, is given by

$$U_i(n_i, \lambda) = C_{ab} \cdot P_i\{\text{Ab}\} \cdot \sqrt{\lambda}, \quad (5.1)$$

where $P_i\{\text{Ab}\}$ is the steady-state probability to abandon over interval i . It is easy to verify that this cost function gives rise to the following constraint on the overall probability to abandon:

$$P\{\text{Ab}\} \leq \frac{\alpha}{\sqrt{\lambda}}, \quad (5.2)$$

where the overall probability to abandon $P\{\text{Ab}\} = \sum_{i=1}^K r_i P_i\{\text{Ab}\}$, $\alpha = M/C_{ab}$ and M is a cost constraint.

Theorem 5.1 (Global constraint on the probability to abandon in the QED regime) Consider the global constraint optimization problem characterized by the cost function (5.1) or the equivalent constraint (5.2). Assume that the patience density at the origin g_0 exists and is positive.

Define the following optimization problem with respect to $\bar{\beta} = \{\beta_i, 1 \leq i \leq K\}$:

$$\left\{ \begin{array}{l} \min_{\bar{\beta}} \sum_{i=1}^K c_i \beta_i \sqrt{r_i}, \\ \text{s.t. } \sum_{i=1}^K \sqrt{r_i} P_w(\beta_i) P_a(\beta_i) = \alpha. \end{array} \right. \quad (5.3)$$

Then at least one solution of (5.3) exists. Denote by δ^* the minimal value that is attained with this solution in (5.3).

a. The optimal staffing level with respect to condition (5.2) satisfies

$$n_i^* = R_i + O(\sqrt{R}), \quad 1 \leq i \leq K, \quad (5.4)$$

$$\sum_{i=1}^K c_i n_i^* = \sum_{i=1}^K c_i \cdot R_i + \delta^* \sqrt{R} + o(\sqrt{R}). \quad (5.5)$$

where $R_i = (r_i \lambda) / \mu$ is the offered load at the i -th interval.

If the solution of (5.3) is unique and given by $\{\beta_i^*\}$, $1 \leq i \leq K$, then (5.4)-(5.5) can be replaced by

$$n_i^* = R_i + \beta_i^* \cdot \sqrt{R_i} + o(\sqrt{R}), \quad 1 \leq i \leq K. \quad (5.6)$$

b. Consider the staffing level

$$\tilde{n}_i^* = \lceil R_i + \beta_i^* \cdot \sqrt{R_i} \rceil, \quad 1 \leq i \leq K, \quad (5.7)$$

where $\{\beta_i^*\}$ is a solution of (5.3). Then the staffing level (5.7) is asymptotically feasible and asymptotically optimal in the sense that

$$\left| \sum_{i=1}^K c_i \tilde{n}_i^* - \sum_{i=1}^K c_i n_i^* \right| = o(\sqrt{R}). \quad \blacksquare$$

See Section 4.1 in the Online Appendix [28] for the proof of Theorem 5.1.

Remark 5.1 The intuition for the constraint in (5.3) is the following. In the QED regime, the probability to abandon at the i -th interval can be approximated, according to formula (4.10), by

$$P_i\{\text{Ab}\} \sim \frac{P_w(\beta_i) P_a(\beta_i)}{\sqrt{r_i \lambda}}.$$

Since $P\{\text{Ab}\} = \sum_{i=1}^K r_i P_i\{\text{Ab}\}$,

$$\sum_{i=1}^K \sqrt{r_i} P_w(\beta_i) P_a(\beta_i) \approx P\{\text{Ab}\} \sqrt{\lambda} \approx \alpha.$$

Remark 5.2 (Conjecture on the solution of (5.3)) Our numerical experiments with the problem (5.3) revealed the following interesting phenomenon. Assume the case of equal costs c_i , $1 \leq i \leq K$, at all intervals. Then (5.3) has a unique solution with equal QoS grades $\beta_1^* = \beta_2^* = \dots = \beta_K^*$. This implies, in particular, higher service levels for intervals with higher arrival rates. (Recall Figure 3 in Section 2.5.) One can show that this result would follow from convexity of the function $P_{ab}(\beta) = P_w(\beta) P_a(\beta)$. Moreover, convexity of $P_{ab}(\beta)$ would imply uniqueness of solution for arbitrary staffing costs c_i . (See the proof of Theorem 5.2, where convexity of the constraint function is used in order to derive similar results.) However, convexity of $P_{ab}(\beta)$ is an open problem. Since

$P_{ab}(\beta)$ approximates the probability to abandon, this open problem is closely related to convexity of the probability to abandon in M/M/n+G, as a function of n . As far as we know, this fact has been proved only for Erlang-A system if the patience parameter θ is smaller or equal than the service rate μ . See Armony et al. [1] for the details of the proof and Koole [25] for a comprehensive discussion on the dynamic programming approach to monotonicity and convexity properties of queueing systems.

Remark 5.3 (Other types of constraints on performance measures) Here we briefly cover several other types of constraints that can be treated by the same methods as in Theorem 5.1.

Scaled constraint on the average wait. Since the QED approximations (4.10) and (4.11) for probability to abandon and average wait, respectively, are closely related, the global constraint $E[W] \leq T/\sqrt{\lambda}$ will give rise to the QED regime at all intervals. The QoS grades could be derived from an optimization problem that is very similar to (5.3): the objective function is the same and the constraint is replaced by $\sum_{i=1}^K \sqrt{r_i} P_w(\beta_i) P_a(\beta_i) = g_0 \cdot T$.

Constraint on the delay probability. Assume that one needs to satisfy a global constraint on the delay probability over k intervals:

$$P\{W > 0\} \leq \alpha.$$

Then the solution will have the following properties. At some intervals the QED regime should be used. At other intervals there should be essentially no staffing ($n_i = o(\sqrt{R_i})$). In addition, there can be multiple solutions to the optimization problem. See Theorem 3.1 in the Online Appendix [28] for details.

Constraint on the tail probability: unscaled and scaled cases. Here one can either use the scaled constraint $P\{W > T\} \leq \alpha/\sqrt{\lambda}$, as in Theorem 4.1, or the unscaled constraint $P\{W > T\} \leq \alpha$, as in Theorem 4.4. In both cases, intervals with essentially no staffing can arise. Other intervals will be staffed according to the QED regime in the first case and according to the ED+QED regime in the second case.

Multiple constraints. In call centers and other service systems it is often desirable to introduce several performance constraints. (For example, for the probability to abandon as well as the tail probability.) Generalization of our approach to this case is straightforward both in single-interval and global settings. Recall that in Section 2.5 we presented a numerical example for the case of two constraints.

5.2 Global constraints in the ED Regime

Consider the following constraint on the overall average wait:

$$E[W] \leq T, \tag{5.8}$$

which is equivalent to a constraint satisfaction for the performance cost functions

$$U_i(n_i, \lambda) = C_w \cdot E_i[W]; \tag{5.9}$$

here $E_i[W]$ is the steady-state average wait at interval i and $E[W] = \sum_{i=1}^K r_i E_i[W]$. The relation between the performance constraint T and the cost constraint M is given via $T = M/C_w$.

Theorem 5.2 (Global constraint on average wait in the ED regime) Consider the global constraint optimization problem characterized by the cost functions (5.9), or the equivalent constraint (5.8). Assume that

$$T < \bar{\tau}, \quad (5.10)$$

and that the cumulative distribution function $G(\cdot)$ of patience times is continuous and strictly increasing over the distribution support.

Define the following optimization problem with respect to $\bar{\gamma} = \{\gamma_i, 1 \leq i \leq K\}$:

$$\left\{ \begin{array}{l} \max_{\bar{\gamma}} \sum_{i=1}^K c_i \gamma_i r_i, \\ \text{s.t.} \sum_{i=1}^K r_i \cdot \int_0^{G^{-1}(\gamma_i)} \bar{G}(u) du = T, \\ \text{s.t.} 0 \leq \gamma_i \leq 1, 1 \leq i \leq K. \end{array} \right. \quad (5.11)$$

Then there exists at least one solution $\{\gamma_i^*\}$ that solves (5.11).

a. The optimal staffing level with respect to constraint (5.8) satisfies

$$\sum_{i=1}^K c_i n_i^* = \sum_{i=1}^K c_i \cdot (1 - \gamma_i^*) R_i + o(R).$$

If the solution of (5.11) is unique then

$$n_i^* = (1 - \gamma_i^*) R_i + o(R), \quad 1 \leq i \leq K.$$

b. Consider the staffing level

$$\tilde{n}_i^* = \lceil (1 - \gamma_i^*) R_i \rceil, \quad 1 \leq i \leq K, \quad (5.12)$$

where $\{\gamma_i^*\}$ is a solution of (5.11). Then the staffing level (5.12) is asymptotically feasible and asymptotically optimal in the sense that

$$\left| \sum_{i=1}^K c_i \tilde{n}_i^* - \sum_{i=1}^K c_i n_i^* \right| = o(R). \quad \blacksquare$$

Now assume that all staffing costs c_i are equal, hence our goal is to minimize the overall staffing level $\sum n_i$. In addition, assume that the hazard rate of the patience distribution exists and is strictly monotone over the distribution support. Propositions 5.1 and 5.2 below cover the Decreasing Hazard Rate (DHR) and the Increasing Hazard Rate (IHR) cases, respectively.

Proposition 5.1 (Global constraint on average wait in the ED regime: DHR) Assume that the staffing costs $c_i \equiv 1$, $1 \leq i \leq K$, and the mean patience $\bar{\tau}$ satisfies $T < \bar{\tau} \leq \infty$. Assume that G is a distribution with strictly *decreasing* hazard rate over the distribution support. Then the unique solution of (5.11) is given by $\gamma_i = \gamma^*$, $1 \leq i \leq K$, where γ^* solves

$$\int_0^{G^{-1}(\gamma)} \bar{G}(u) du = T$$

with respect to γ . ■

Proposition 5.2 (Global constraint on average wait in the ED regime: IHR) Assume that the staffing costs $c_i \equiv 1$, $1 \leq i \leq K$, and the mean patience $\bar{\tau}$ satisfies $T < \bar{\tau}$. Assume that G is a distribution with strictly *increasing* hazard rate over the distribution support. Let $\{\gamma_i^*, 1 \leq i \leq K\}$ denote a solution of (5.11). Define $\mathcal{K}_0 = \{i : \gamma_i^* = 0\}$, $\mathcal{K}_1 = \{i : \gamma_i^* = 1\}$, $\mathcal{K}_2 = \{i : 0 < \gamma_i^* < 1\}$. Then $\{\gamma_i^*, 1 \leq i \leq K\}$ has the following properties:

1. \mathcal{K}_2 consists of either zero or one element.
2. If $\mathcal{K}_2 = \{i\}$, then

$$\bar{\tau} \cdot \left(\sum_{k \in \mathcal{K}_1} r_k \right) < T < \bar{\tau} \cdot \left(\sum_{k \in \mathcal{K}_1} r_k + r_i \right). \quad (5.13)$$

3. If there exists a subset $\mathcal{K} \subset \{1, 2, \dots, K\}$ s.t. $\bar{\tau} \cdot \sum_{k \in \mathcal{K}} r_k = T$, then \mathcal{K}_2 is empty and $\bar{\tau} \cdot \sum_{k \in \mathcal{K}_1} r_k = T$.
4. If $\mathcal{K}_2 = \{i\}$, then for all $j \in \mathcal{K}_1$ such that

$$\bar{\tau} \cdot \left(\sum_{k \in \mathcal{K}_1 \setminus \{j\}} r_k + r_i \right) < T, \quad (5.14)$$

the inequality $r_i \leq r_j$ prevails.

5. If $\mathcal{K}_2 = \{i\}$, then for all $j \in \mathcal{K}_0$ such that $\bar{\tau} \cdot (\sum_{k \in \mathcal{K}_1} r_k + r_j) > T$, the inequality $r_i \leq r_j$ prevails. ■

See Section 4.2 in the Online Appendix [28] for the proofs of Theorem 5.2, Proposition 5.1 and Proposition 5.2.

Remark 5.4 (Intuition for Propositions 5.1 and 5.2) If the patience distribution is DHR (customers become more patient as their waiting time increases), it will be shown that the fluid limit of average wait is convex in the ED parameter γ . Due to convexity of average wait and linearity of staffing costs in the number of servers, extremely low staffing levels would lead to increasing overall costs. Therefore, Proposition 5.1 recommends sustaining the same staffing parameters γ_i and, hence, the same service level at all intervals.

In contrast, the IHR case (customers lose patience in the process of waiting) gives rise to a concave limit of the average wait. In this case, Proposition 5.2 recommends different service levels over different intervals. Specifically, staffing around the offered load is recommended for some intervals ($\gamma_i^* = 0$) and essentially no staffing at the others ($\gamma_i^* = 1$). There can be, at most, one interval with “intermediate staffing” ($0 < \gamma_i^* < 1$). Statements 1-5 of Proposition 5.2 also elaborate on properties of this solution. Specifically, Statements 4 and 5 show that if one should apply “intermediate staffing” to one of two “candidate” intervals, given that the staffing regime at the other $K - 2$ intervals is fixed, the interval with the minimal arrival rate should be chosen for “intermediate staffing”. (However, it need not be true that the interval with the minimal arrival rate between *all* intervals should be chosen for “intermediate staffing”.)

Remark 5.5 (Global constraint on the probability to abandon in the ED regime) If the unscaled global constraint $P\{\text{Ab}\} \leq \alpha$ is considered, it is straightforward to show that asymptotically optimal staffing is ED:

$$n_i = (1 - \gamma_i)R_i + o(R), \quad 0 \leq \gamma_i \leq 1, \quad 1 \leq i \leq K.$$

In order to calculate $\{\gamma_i\}$, the following linear programming problem should be solved:

$$\begin{cases} \max_{\{\gamma_i\}} \sum_{i=1}^K c_i r_i \gamma_i, \\ \text{s.t.} \sum_{i=1}^K \gamma_i r_i = \alpha, \quad 0 \leq \gamma_i \leq 1. \end{cases} \quad (5.15)$$

If all the staffing costs c_i are equal, then any $\{\gamma_i\}$ with $\sum \gamma_i r_i = \alpha$ is asymptotically optimal. Otherwise assume, without loss of generality, that $c_1 \geq c_2 \geq \dots \geq c_K$. Then an optimal solution is obtained recursively: $\gamma_i = \min(1, \alpha_i / r_i)$, $1 \leq i \leq K$, where $\alpha_1 = \alpha$ and $\alpha_i = \alpha_{i-1} - \gamma_{i-1} r_{i-1}$, $2 \leq i \leq K$.

6 Numerical examples

In this section, we present three educating numerical examples on the three operational regimes studied in Section 4. Then the comparison of the three regimes is performed, supporting conclusions that were presented in Section 2.4. See Section 2.5 for a practically-oriented numerical experiment on global constraints and Online Appendix [28] for a comprehensive study on all types of constraints.

6.1 Constraint satisfaction in the QED regime

Consider a moderate-size call center with an average of 20 arrivals per minute over a given time interval. Let the average service time be 3 minutes ($\mu = 1/3$). Therefore, the offered load is equal to $R = \lambda/\mu = 20 \cdot 3 = 60$ Erlangs.

Assume that the customers of this call center constitute a 50-50% mixture of impatient customers with patience that is distributed $\text{Exp}(\text{mean}=1)$ and patient ones with $\text{Exp}(\text{mean}=5)$ patience. Formally, the distribution of overall patience is hyperexponential.

Suppose that the call center would like to maintain a high level of service. Three possible performance constraints are considered, unilaterally:

1. Probability to abandon should be less than 2%;
2. Average wait should be less than 5 seconds;
3. At least 90% of the customers should wait less than 20 seconds.

As explained in Section 2.4, the parameters of the problem (high service level, around $R=60$ agents needed) are appropriate for application of the QED regime (2.1). We thus calculate the optimal QoS parameter β^* via equation (4.15) from Section 4.2. Then the approximately optimal staffing level is given by formula (2.2). The optimal staffing n_{QED}^* for the three constraints above turns out to be 67, 62 and 61 agents, with the optimal QoS parameters β^* being 0.79, 0.14 and 0.04, respectively.

If we calculate the optimal staffing via exact M/M/n+G formulae [43], it turns out that the fit is perfect: the exact optimal staffing n^* is equal to n_{QED}^* in all three cases.

How does the patience distribution affect the optimal staffing level? Maintaining a mean patience of 3 minutes, assume that patience times are now uniformly distributed between 0 and 6 minutes. (This could correspond to a situation when after 6 minutes of wait customers are routed to other location.) Then both n^* and n_{QED}^* for the three constraints are 64, 66 and 63, respectively, again a perfect fit. We observe that if the $P\{\text{Ab}\}$ constraint is considered, more agents (67 vs. 64) are needed in the case of hyperexponential patience. However, if the wait in queue is controlled, the staffing level should be higher (66 vs. 62) for uniform patience.

As our theory reveals (formulae (4.5)-(4.12)), the patience density near the origin is a key characteristic that determines performance of queues with a high service level. Higher density near the origin implies more abandonment and smaller wait. The tail of the patience distribution and even its mean are less important. One can check that, of the two distributions mentioned above, the hyperexponential has a higher density near the origin ($3/5$ vs. $1/6$). Therefore, the staffing recommendations above (67 vs. 64) are consistent with [43].

6.2 Constraint satisfaction in the ED regime

Consider a very large call center with 400 arrivals per minute, average service time 3 minutes, hence $R = 1200$. Assume a hyperexponential patience distribution with the parameters, introduced in Section 6.1. Assume that management has an efficiency-driven view of the call center operations: utilization of agents should be close to 100% but at the cost of a certain compromise on service level.

Two alternative performance constraints are considered here:

1. Probability to abandon should be less than 10%;

2. Average wait should be less than 20 seconds.

For these parameters (large number of agents, “loose” service-level constraints) it is reasonable to apply ED staffing (2.4), where the values of parameter γ^* are established via equation (4.20) from Theorem 4.2. We get $n_{ED}^* = 1,080$ agents for the first constraint and 972 agents for the second one. The exact optimal solutions n^* are 1,081 and 972 agents, respectively.

Now consider a $U(0,6)$ patience distribution instead. In this case, our ED approximations prescribe 1,080 and 1,132 agents, and the exact solutions are 1,081 and 1,132. We observe the phenomenon that was mentioned above: staffing with respect to the $P\{Ab\}$ constraint in the ED regime does not depend on the patience distribution. However, if average wait is controlled, the influence of the patience distribution can be very significant: 972 vs. 1,132 agents.

Note that, in this example, we used only two types of constraints, as opposed to three in Section 6.1: the constraint on the tail probability $P\{W > T\}$ is not treated. The reason is that, as explained in Section 4.3, the ED regime does not provide an applicable approximation for the distribution of waiting time. However, as we know from Section 4.4, the ED+QED refinement enables such approximations.

6.3 Constraint satisfaction in the ED+QED regime

Consider the large call center from Section 6.2 with offered load $R = 1,200$. Assume the commonly-used service-level constraint: “at least 80% of the customers should wait less than 20 seconds”. Consider three patience distributions with the same mean: $\text{Exp}(\text{mean}=3)$, $U(0,6)$ and our previous hyperexponential mixture of $\text{Exp}(\text{mean}=1)$ and $\text{Exp}(\text{mean}=5)$. Applying the staffing formula (2.6) with values of γ^* and δ^* derived from Theorem 4.4, we get n_{ED+QED}^* equal to 1,099, 1,153 and 1,020, respectively. (The exact optimal values are 1,100, 1,153 and 1,021.) Theorem 4.4 implies that the γ^* ’s are different for the three distributions, hence the large variations in staffing levels. One concludes that the use of the exponential assumption on patience (the Erlang-A model), which is slowly becoming standard in call centers, can imply significant deviations from the optimum, under some circumstances.

Another important insight from Sections 6.2 and 6.3 is that a reasonable service level and beyond can be reached even if significant understaffing with respect to R takes place – given sufficient scale.

6.4 Comparison between operational regimes

Recall that in Section 2.4 we discussed existence of a single operational regime that is preferable over the others. Here we re-visit examples from Sections 6.1-6.3 in order to show that there is no such regime. Note that the same conclusion can be deduced if one summarizes comprehensive numerical examples from the Online Appendix [28].

Example 6.1 (Constraint satisfaction in a small call center) Consider the setting of Section 6.1 with hyperexponential patience, where the QED approximations provide us with a perfect fit.

Apply the ED staffing (2.4) and the corresponding approximations. It is straightforward to check, via $P_{ED}\{\text{Ab}\} = \gamma$, that the ED recommendation for the constraint $P\{\text{Ab}\} \leq 2\%$ is $n_{ED}^* = 59$. This is very far from the exact optimum $n^* = 67$ and would lead to $P\{\text{Ab}\} = 6.7\%$ – more than three-fold worse than the service goal. Therefore, the ED recommendations should not be used for small call centers.

Now we check if the ED+QED regime is robust for a small call center. Applying it for $\text{Exp}(\text{mean}=3)$, $U(0, 6)$ and our hyperexponential distribution, we get respectively that $n_{ED+QED}^* = 63, 64$ and 61 , while $n^* = n_{QED}^* = 64, 66$ and 61 . We observe a perfect QED fit. Hence the ED+QED recommendations are not that bad but the QED ones, nevertheless, are preferable for small call centers.

Example 6.2 (Constraint satisfaction in a large efficiency-driven call center) Consider the large call center from Sections 6.2-6.3, where the ED and ED+QED approximations were found appropriate. We check if the QED approximations are robust in this case, considering the three patience distributions from the end of Example 6.1. First, consider the constraint $P\{\text{Ab}\} \leq 10\%$ from Section 6.2. QED recommends 1,081 agents for all distributions, which coincides with the exact optima.

In contrast, QED staffing for the constraint “ $E[W] \leq 20$ seconds” is 1,067, 1,134 and 961 vs. the exact optima of 1,067, 1,132 and 972. We observe that the fit for our hyperexponential patience is relatively poor. Considering the constraint on $P\{W > T\}$ from Section 6.3, we also get a poor fit of QED approximations, especially for the hyperexponential distribution: $n_{ED+QED}^* = 1,000$ vs. $n^* = 1,021$.

Hence, using QED approximations in large ED call centers can mislead if moderate-to-loose constraints on the waiting time are considered.

Remark 6.1 (QED approximations in an Efficiency-Driven setting) Example 6.2 demonstrates that the QED approximation for $P\{\text{Ab}\}$ provides an excellent fit for a large overloaded call center. In order to understand the reason, recall the QED approximation of the probability to abandon:

$$P\{\text{Ab}\} \approx \frac{1}{\sqrt{\lambda}} \sqrt{g_0} [h_\phi(\hat{\beta}) - \hat{\beta}].$$

For large negative $\hat{\beta}$, the normal hazard rate $h_\phi(\cdot)$ is negligible. Using the definition of $\hat{\beta}$ in (4.8), we can easily deduce that the QED approximation is then close to $(R - n)/R$, namely the ED approximation.

If the patience distribution is exponential, the QED approximation provides an excellent fit for the average wait as well. This can be explained by the high quality of $P\{\text{Ab}\}$ approximations and the relation $P\{\text{Ab}\} = \theta \cdot E[W]$, which prevails for both exact values and QED approximations. (Note that the exponential parameter $\theta = g_0$.) However, if the patience distribution is non-exponential, $P\{\text{Ab}\} = g_0 \cdot E[W]$ does not prevail in the ED regime and, as we show in Section 6.2 of the Online Appendix [28], QED approximations can have significant bias with respect to the exact optimal values.

7 Possible future research

To conclude, we outline several types of problems that we propose for future research.

- **Revenue/cost optimization.** As already discussed in the Introduction, optimization of revenues and/or costs constitutes an alternative to the approach of the present paper. The ongoing research [27] is dedicated to this problem for the $M/M/n+G$ queue, continuing the work of Borst et al. [6] on Erlang-C.
- **Additional research on global constraint satisfaction.** Section 5 of the present paper gives rise to interesting research problems. For example, one could try to verify the conjecture in Remark 5.2.

It would be also interesting to study the staffing level for *several joint constraints*, for example $P\{Ab\}$ and $P\{W > T\}$ (or rather $P\{W > T, Sr\}$). We believe that, in this case, unlike the single-interval problem, several constraints could be binding for an asymptotic solution.

- **Time-inhomogeneous arrival rate.** Such queues are prevalent in practice and their time-varying analysis poses a challenge. A common approach is to approximate the time-varying arrival-rate by a piecewise-constant function, and then apply steady-state results during periods when the arrival rate is assumed constant. An implicit assumption is that the arrival rate is slow-varying with respect to the durations of services. Recently, Feldman et al. [13] developed an alternative simulation-based algorithm for staffing time-varying queues with abandonment in order to achieve a constant delay probability. We believe that a similar approach can be applied to other constraint satisfaction problems such as those analyzed in the present paper.
- **Generally distributed service times.** The $M/M/n+G$ model assumes exponential services. However, this assumption does not apply for many call centers. For example, in several application (e.g. [10]) we encountered a lognormal distribution of service times. Therefore, it is important to study the $M/G/n+G$ model with a general service distribution. Whitt [40] suggests approximating, in steady state, $M/G/n+G$ by $M/M/n+G$ with the same service mean. Recently, Reed [32, 33] studies the $GI/GI/N$ queue in the QED (Halfin-Whitt) regime, but his results are sample-paths limits, as opposed to our steady-state limits. We believe that additional research in this direction is worthwhile.
- **Random arrival rate.** In Brown et al. [10] and Weinberg et al. [35] it was shown that Poisson arrival rates in two different call centers vary from day to day and the prediction of arrival rates raises statistical and practical challenges. Therefore, it is very important to study queueing models where the arrival rate Λ of a homogeneous Poisson arrival process is in fact a random variable. We expect that both the QED and ED regimes (and, maybe, some new regimes) can be relevant in this case, depending on the order of the variation of Λ . See Whitt [39], and Bassamboo, Harrison and Zeevi [4] for the “cruder” ED case.

References

- [1] Armony M., Plambeck E. and Seshadri S. (2007) Sensitivity of Optimal Capacity to Customer Impatience in an Unobservable M/M/S Queue (Why You Shouldn't Shout at the DMV). *Manufacturing & Service Operations Management*, to appear. 5.2
- [2] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: Kylstra F.J. (Ed.), *Performance '81*. North-Holland Publishing Company, 159-179. 2, 3
- [3] Baron O., Milner J. (2008) Staffing to maximize profit for call centers with alternate service level agreements. *Operations Research*, to appear. 3
- [4] Bassamboo A., Harrison J.M. and Zeevi A. (2006) Design and control of a large call center: asymptotic analysis of an LP-based method. *Operations Research*, 54, 419-435. 3, 7
- [5] Bassamboo A., Harrison J.M. and Zeevi A. (2005) Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scaled fluid limits. *Queueing Systems: Theory and Applications (QUESTA)*, 51, 249-285. 3
- [6] Borst S., Mandelbaum A., and Reiman M. (2004), Dimensioning large call centers, *Operations Research*, 52(1), 17-34. 1, 1.3, 2.2, 2.4, 3, 3, 7
- [7] Boxma O.J. and de Waal P.R. (1994) Multiserver queues with impatient customers. *ITC*, 14, 743-756. 3
- [8] Brandt A. and Brandt M. (1999) On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35, 1-18. 2, 3
- [9] Brandt A. and Brandt M. (2002) Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems: Theory and Applications (QUESTA)*, 41, 73-94. 2, 3
- [10] Brown L.D., Gans N., Mandelbaum A., Sakov A., Shen H., Zeltyn S. and Zhao L. (2005) Statistical analysis of a telephone call center: a queueing science perspective. *Journal of the American Statistical Association (JASA)*, 100(469), 36-50. 2, 2.1, 3, 7
- [11] Delorey E. Correlating IVR Performance and Customer Satisfaction. Available at http://www.easyivr.com/tech-ivr-applications_108.htm. 1
- [12] Erlang A.K. (1948) On the rational determination of the number of circuits. In *The life and works of A.K.Erlang*. Brockmeyer E., Halstrom H.L. and Jensen A., eds. Copenhagen: The Copenhagen Telephone Company. 1.1, 3
- [13] Feldman Z., Mandelbaum A., Massey W. and Whitt W. (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2), 324-338. 3, 7
- [14] 4CallCenters Software (2002). Available at <http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm>. 1.1
- [15] Gans N., Koole G. and Mandelbaum A. (2003) Telephone call centers: a tutorial and literature review. Invited review paper, *Manufacturing and Service Operations Management*, 5 (2), 79-141. 1.1, 3
- [16] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a telephone call-center with impatient customers. *Manufacturing and Service Operations Management* 4, 208-227. 2.1, 3
- [17] Gilson K.A. and Khandelwal D.K. (2005) Getting more from call centers. The McKinsey Quarterly. Available at <http://www.marketingpower.com/content25334.php#1>. 1
- [18] Green L.V., Kolesar P.J. and Soares J. (2001) Improving the SIPP approach for staffing service systems that have cyclic demand. *Operations Research*, 49, 549564. 3
- [19] Green L.V., Kolesar P.J. and Soares J. (2003) An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management*, 12(1), 46-61. 3
- [20] Green L.V., Kolesar P.J. and Whitt W. (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1), 13-39. 2.1, 3

- [21] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588. [3](#), [4.4](#)
- [22] Harrison J.M. and Zeevi A. (2005) A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management*, 7, 20-36. [3](#)
- [23] Holman D., Batt R., and Holtgrewe U. (2007) The global call centre report: international perspectives on management and employment. ILR Collection, Research Studies and Reports, Cornell University. Available at <http://digitalcommons.ilr.cornell.edu/reports/13/>. [1](#)
- [24] Koole G. and van der Sluis E. (2003) Optimal shift scheduling with a global service level constraint. *IEE Transactions*, 35 (11), 1049-1055. [3](#)
- [25] Koole G. (2006) Monotonicity in Markov reward and decision chains: Theory and applications. *Foundations and Trends in Stochastic Systems*, 1, 1-76. [5.2](#)
- [26] Mandelbaum A. and Zeltyn S. (2007) Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. In: Spath D., Fähnrich, K.-P. (Eds.), *Advances in Services Innovations*, 17-48, Springer-Verlag. [3](#)
- [27] Mandelbaum A. and Zeltyn S. (2007) Dimensioning call centers with abandonment: cost optimization. In preparation. [1](#), [7](#)
- [28] Mandelbaum A. and Zeltyn S. (2008) Staffing many-server queues with impatient customers: constraint satisfaction in call centers. Online Appendix. Available at <http://iew3.technion.ac.il/serveng/References/constraintsupp.pdf>. [1.3](#), [1.4](#), [2.1](#), [2.4](#), [2.4](#), [4.1](#), [4.2](#), [4.3](#), [4.4](#), [4.4](#), [5.1](#), [5.3](#), [5.2](#), [6](#), [6.4](#), [6.1](#)
- [29] Palm C. (1957) Research on telephone traffic carried by full availability groups. *Tele*, vol.1, 107 pp. (English translation of results first published in 1946 in Swedish in the same journal, which was then entitled *Tekniska Meddelanden fran Kungl. Telegrafstyrelsen*.) [1.1](#), [3](#)
- [30] Randhawa R.S. and Kumar S. (2007) Multi-server loss systems with subscribers. Working paper. [3](#)
- [31] Randhawa R.S. and Kumar S. (2007) Usage restriction and subscription services: operational benefits with rational users. *Manufacturing and Service Operations Management*, to appear. [3](#)
- [32] Reed J.E. (2007) The G/GI/N queue in the Halfin-Whitt regime I: Infinite server queue system equations. Submitted to publication. [7](#)
- [33] Reed J.E. (2007) The G/GI/N queue in the Halfin-Whitt regime II: Idle time system equations. Submitted to publication. [7](#)
- [34] de Véricourt F. and Jennings O.B. (2008) Dimensioning large-scale membership services. *Operations Research*, 56(1), 173-187. [3](#)
- [35] Weinberg J., Brown L.D and Stroud J.R. (2007) Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, 102, 1185-1198. [7](#)
- [36] Whitt W. (1992) Understanding the efficiency of multi-server service systems. *Management Science*, 38(5), 708-723. [3](#)
- [37] Whitt W. (2006) Fluid models for many-server queues with abandonments. *Operations Research*, 54(1), 37-54. [3](#)
- [38] Whitt W. (2006) Sensitivity of performance in the Erlang A model to changes in the model parameters. *Operations Research*, 54(2), 247-260. [3](#)
- [39] Whitt W. (2006) Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1), 88-102. [3](#), [7](#)
- [40] Whitt W. (2005) Engineering Solution of a Basic Call-Center Model. *Management Science*, 51(2), 221-235. [3](#), [7](#)

- [41] Whitt W. (2005) Two fluid approximations for multi-server queues with abandonments. *Operations Research Letters*, 33, 363-372. [3](#)
- [42] Whitt W. (2006) A multi-class fluid model for a contact center with scaled-based routing. *International Journal of Electronics and Communications*, 60(2), 95-102. [3](#)
- [43] Zeltyn S. and Mandelbaum A. (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems: Theory and Applications (QUESTA)*, 51, 361-402. [1.1](#), [1.3](#), [2](#), [2.1](#), [2.2](#), [2.3](#), [2.2](#), [2.1](#), [3](#), [3](#), [4.2](#), [4.3](#), [4.4](#), [6.1](#)
- [44] Zeltyn S. and Mandelbaum A. (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. Internet Supplement. Available at <http://iew3.technion.ac.il/serveng/References/references.html>. [2.2](#)

Acknowledgements. The research of both authors was supported by BSF (Binational Science Foundation) grant 2001685/2005175, ISF (Israeli Science Foundation) grants 388/99, 126/02, 1046/04, IBM and by the Technion funds for the promotion of research and sponsored research. This paper grew out of joint research with Sam Borst and Marty Reiman – their contribution and encouragement are greatly appreciated. The authors thank the associate editor and the referees for their constructive detailed feedback.