Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue*

Sergey Zeltyn and Avishai Mandelbaum†

Faculty of Industrial Engineering & Management
Technion
Haifa 32000, ISRAEL

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

June 17, 2005

*An Internet Supplement to the present paper is downloadable from
`http://iew3.technion.ac.il/serveng/References/references.html`

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

Contents

1 Introduction **1**

1.1 The quality/efficiency tradeoff in call centers 1

1.2 Patience in invisible queues: the M/M/n+G model 2

1.3 Operational measures of performance 5

1.4 The ED, QD and QED operational regimes 6

2 Structure of the paper and summary of results **8**

2.1 Structure and summary of the paper 8

2.2 Summary of the Internet Supplement [48] 9

3 Literature review **9**

3.1 Relevant exact results for M/M/n+G 9

3.2 Relevant asymptotic results 10

4 The QED operational regime **12**

4.1 Formulation of results 12

4.1.1 Main case: patience distribution with positive density at the origin 12

4.1.2 Patience with balking 16

4.1.3 Patience with scaled balking 18

4.2 Numerical example 19

5 The Quality-Driven operational regime **21**

6 The Efficiency-Driven operational regime **22**

6.1 Formulation of results 22

6.2 Numerical example 24

7 Conclusions **25**

8 Ongoing and future research **27**

9 M/M/n+G queue: summary of performance measures **28**

10 Asymptotic behavior of integrals **30**

10.1 The Laplace method 30

10.2 Asymptotic results 31

11 Selected proofs **32**

11.1 Proof of Theorem 4.1, a-e 32

11.2 Proof of Theorem 5.1, a-e 36

Abstract

The subject of the present research is the M/M/n+G queue. This queue is characterized by Poisson arrivals at rate λ, exponential service times at rate μ, n service agents and generally distributed patience times of customers. The model is applied in the call center environment, as it captures the tradeoff between operational efficiency (staffing cost) and service quality (accessibility of agents).

In our research, three asymptotic operational regimes for medium to large call centers are studied. These regimes correspond to the following three staffing rules, as λ and n increase indefinitely and μ held fixed:
Efficiency-Driven (ED): $n \approx (\lambda/\mu)\cdot(1-\gamma)\,, \quad \gamma > 0,$
Quality-Driven (QD): $n \approx (\lambda/\mu)\cdot(1+\gamma)\,, \quad \gamma > 0,$ and
Quality and Efficiency Driven (QED): $n \approx \lambda/\mu + \beta\sqrt{\lambda/\mu}\,, \quad -\infty < \beta < \infty.$

In the ED regime, the probability to abandon and average wait converge to constants. In the QD regime, we observe a very high service level at the cost of possible overstaffing. Finally, the QED regime carefully balances quality and efficiency: agents are highly utilized, but the probability to abandon and the average wait are small (converge to zero at rate $1/\sqrt{n}$).

Numerical experiments demonstrate that, for a wide set of system parameters, the QED formulae provide excellent approximation for exact M/M/n+G performance measures. The much simpler ED approximations are still very useful for overloaded queueing systems.

Finally, empirical findings have demonstrated a robust linear relation between the fraction abandoning and average wait. We validate this relation, asymptotically, in the QED and QD regimes.

1 Introduction

1.1 The quality/efficiency tradeoff in call centers

During the last two decades, there has been an explosive growth in the number of companies that provide services via the telephone, as well as in the variety of telephone services provided. In the U.S. alone, the call center industry is estimated to employ several million agents which, in fact, outnumber agriculture [17, 13, 40]. In Europe, the number of call center employees in 1999-2000 was estimated, for example, by 600,000 in the UK (2.3% of the total workforce) and 200,000 in Holland (almost 3%) [4].

A central challenge in designing and managing a service operation in general, and telephone-based services in particular, is to achieve a desired balance between *operational efficiency* and *service quality*. In our research, we treat the staffing aspect of the *quality/efficiency tradeoff*, namely having the right number of agents in place. "The right number" means not too many, saving operating costs, and not too few, avoiding excessive customers' wait and abandonment.

The quality and efficiency levels of a well-run modern call center can be extraordinarily high. Indeed, in a large performance-leader enterprise, many hundreds of agents could serve many thousands of calling customers per hour; agents' utilization levels could exceed 90-95%, yet the service level could be very high. To reach these levels of performance, one presumes that planning would require sophisticated stochastic queueing models. However, one actually observes in practice that simple deterministic approaches lead to surprisingly good results. This puzzling, often professionally frustrating phenomenon, can be explained via our models, as the following example demonstrates.

Consider a call center with an average of 6000 calls per hour and service time of 4 minutes. Such a call center gets an average of $(6000 : 60) \cdot 4 = 400$ minutes of work per minute. The deterministic approach then prescribes 400 service agents to cope with this load, which is a questionable recommendation according to standard queueing models. For example, Erlang-C (M/M/n) would then be unstable, and waiting times and queue-lengths would increase indefinitely. But now assume that customers abandon, as they actually do, and assign a reasonable parameter to their patience, say on the order of a service time. Then, about 50% of the customers would be answered immediately upon calling, the average wait would be a mere 5 seconds, agents' utilization would be 98%, and all this at the cost of 2% abandonment – a remarkable performance indeed. (See Remark 4.6 on page 15 for a formal discussion that explains how sophisticated models do sometimes, but not always, lead to simple answers that are consistent with the deterministic approach.)

Most call centers are far from achieving the levels of service cited above. To these, our models would help climb the performance ladder, and this is where our contribution lies.

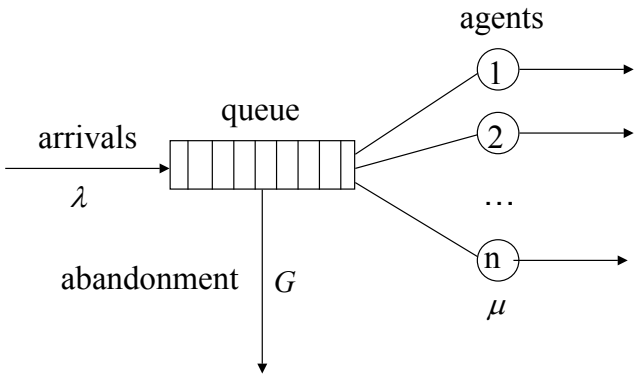
In light of the quality/efficiency tradeoff challenge, it is natural to model call centers by

queueing systems. Note that unlike many other queues, call center queues are *invisible*: callers cannot observe how long a queue is and their progress in it. Consequently, the abandonment behavior of customers in tele-queues is different from face-to-face queues. See, for example [17, 28, 39, 49].

1.2 Patience in invisible queues: the M/M/n+G model

As mentioned, the classical M/M/n queueing model, also called **Erlang-C**, is the model most frequently used in *workforce management* of call centers. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed service times with a rate μ, and n independent statistically-identical agents. (Time-varying arrival rates are accommodated via piecewise constant approximations.) But Erlang-C does not allow abandonment, which is a significant deficiency: customer abandonment is not a minor, let alone negligible, aspect of call center operations; see Garnett et al. [18]. Specifically, and as demonstrated above, ignoring abandonment can lead to wrong staffing decisions and distorted definitions of service level.

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



In this paper, we enrich Erlang-C to the **M/M/n+G queue**, which has the following additional feature: associated with each arriving caller there is a generally distributed *patience time* τ with a common distribution G. An arriving customer encounters an *offered waiting time* V, defined as the time that this customer would have to wait given that her or his patience is infinite. If the offered wait exceeds the customer's patience time, the call is then abandoned, otherwise the customer awaits service. In both cases, the actual waiting time W is equal to $\min(V, \tau)$. Our treatment of the M/M/n+G model is based on Baccelli and Hebuterne [3].

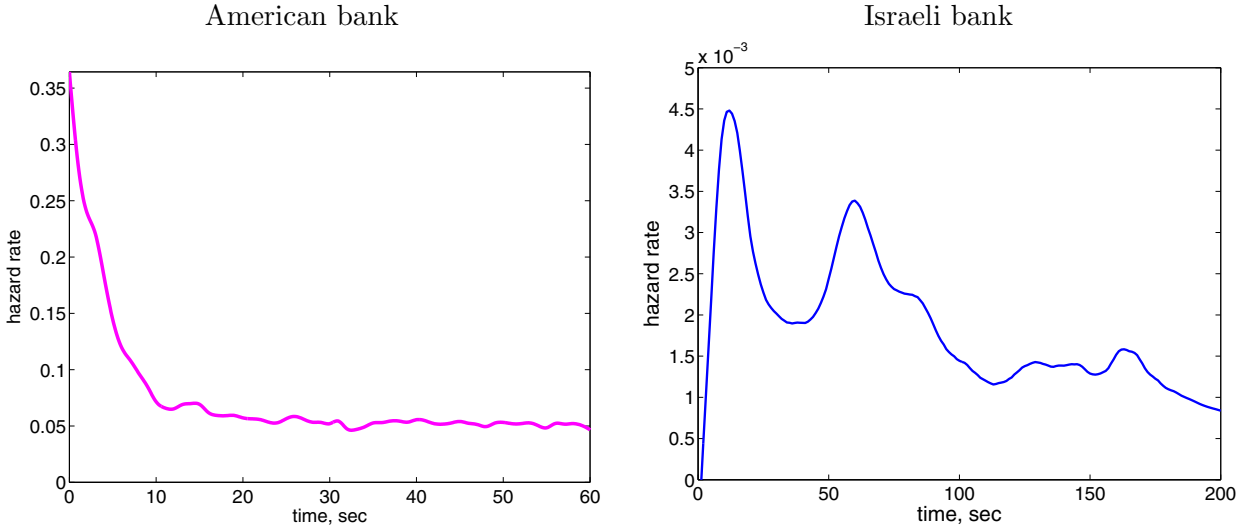
M/M/n+G generalizes the M/M/n+M (Erlang-A) model with exponentially distributed patience times, which is the most tractable model with abandonment. (This model was introduced

by Palm [37]. See also [18], [30] and [33].) M/M/n+M is already used in well-run call centers. We now explain why it is important to study its generalization.

In Figure 2 we display hazard-rate estimates of the customers' patience for two banks: a large U.S. bank and a small Israeli one. In the two cases we observe different, but clearly non-exponential patterns. (Recall that the hazard rate of an exponential random variable is a constant.) American customers are very impatient at the beginning of their wait, but their patience stabilizes after approximately 10 seconds. In contrast, Israeli customers have two clear peaks of abandonment: approximately at 15 and at 60 seconds. (It turns out that these two surges of abandonment take place immediately after two recorded messages to which customers are exposed: the first one when they enter the queue and the second after approximately 1 minute.)

Therefore, at least in some applications (according to our experience, in most), customers' patience times are non-exponential.

Figure 2: **Bank data: hazard rates of patience times**



We now show that a non-exponential distribution of patience can significantly affect system performance, when compared to the Erlang-A system with the same average patience. Consider M/M/n+G with $n = 100$ and service rate $\mu = 1$. Three patience distributions with the same average patience are compared: exponential with average 2, constant 2 and uniform on [0,4]. We varied the arrival rate λ from 10 to 500, in step 2.5, plotting in Figure 3 two performance measures: the probability to abandon and average wait. (We assume that the unit of time is minutes.)

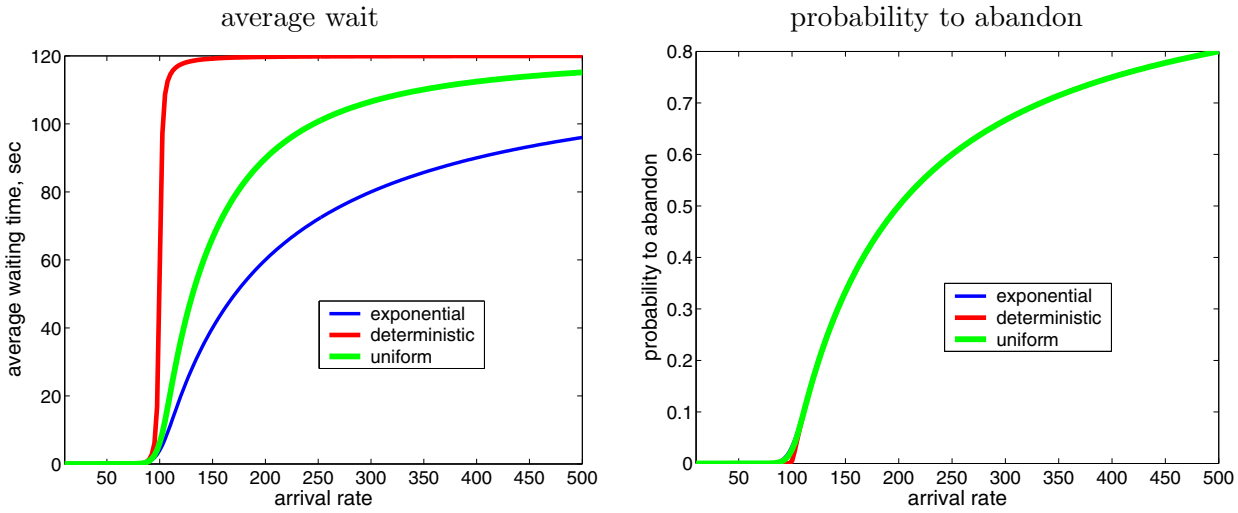
Observe that the two plots are not similar. The three average-wait curves are very different, except for small values of λ when the wait is negligible. In contrast, the probability-to-abandon

curves seem almost identical. Only if we zoom around $\lambda = 100$, there is a noticeable difference between deterministic patience and the two other distributions.

From this example we conclude that:

- Patience-distribution can significantly affect performance of the M/M/n+G queue;
- The effect of the patience distribution strongly depends on the performance measure we consider (average wait and probability to abandon, in our example);
- The effect of the patience distribution depends on the load that the system is working under. Specifically, it is very important whether the *offered load per agent* $\frac{\lambda}{n\mu}$ is significantly below 1, around 1, or significantly above 1.

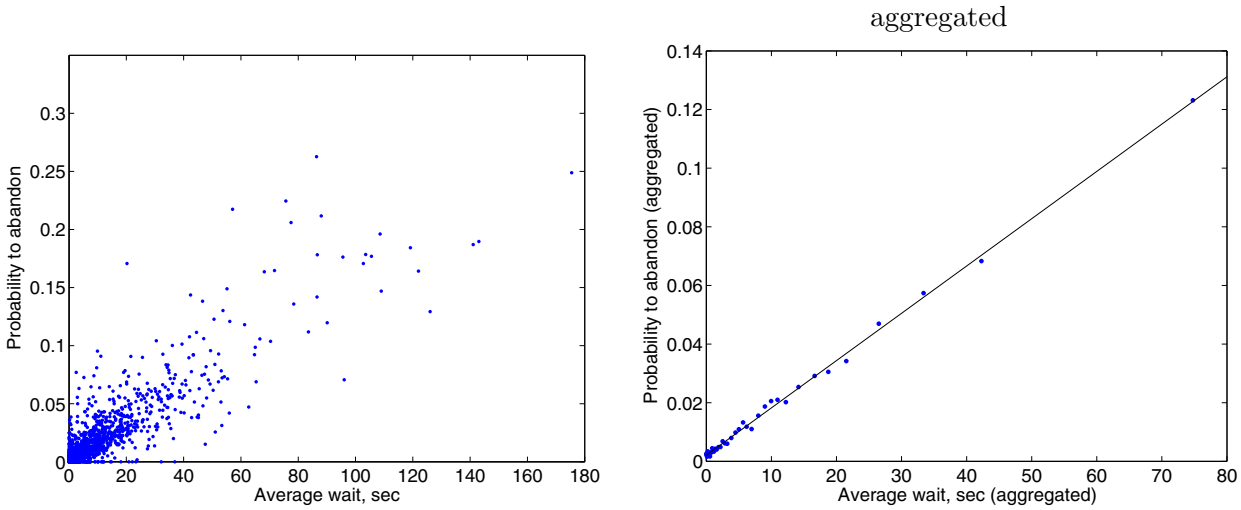
Figure 3: **Dependence of performance on patience distribution**



Another aspect of the abandonment phenomenon was explored in Mandelbaum and Zeltyn [32], where we studied an empirical and theoretical relationship between the probability to abandon P{Ab} and average wait E[W]. As an example, consider data from our large U.S. bank call center. First, P{Ab} and E[W] were computed for 1649 hourly intervals that constitute 5 months of work between 7am and 24pm during weekdays. The left plot of Figure 4 presents the resulting "cloud" of points, as they scatter on the plane. For the right plot, we are using an aggregation procedure that is designed to emphasize dominating patterns. Specifically, the 1649 intervals were ordered according to their average waiting times, and adjacent groups of 40 points were aggregated (further averaged): this forms the 41 points of the second plot in Figure 4. (The last point of the aggregated plot is an average of 49 hour intervals.)

We observe a remarkable linear fit that theoretically prevails for models with *exponential* patience, but we see that it practically arises for non-exponential patience as well. The paper [32] contains experimental and theoretical research on this issue. In the present paper, we add support as to why the linear relation can arise in the M/M/n+G model with non-exponential patience.

Figure 4: **Telesales customers. Probability to abandon vs. average waiting time**



1.3 Operational measures of performance

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 thousands of agents), the implementation of these calculations must be both scalable and numerically stable.

The most popular measure of operational (positive) performance is $\mathrm{P}\{W \leq T, \mathrm{Sr}\}$, where W is the waiting time, $\{\mathrm{Sr}\}$ is the event "customer gets service" and T is a target time that is determined by Management/Marketing. However, as explained before, performance measures must take into account those customers who abandon. Indeed, if forced into choosing a *single* number as a proxy for operational performance, we recommend it to be the *probability to abandon* $\mathrm{P}\{\mathrm{Ab}\}$, the fraction of customers who explicitly declare that the service offered is not worth its wait. Some managers actually opt for the refinement $\mathrm{P}\{W > \epsilon; \mathrm{Ab}\}$, for some small $\epsilon > 0$, for example $\epsilon = 3$ seconds. The justification is that those who abandon within 3 seconds can not be characterized as poorly served. There is also a practical rationale that arises from physical limitations, specifically that such "immediate" abandonment could in fact be a malfunction or an inaccuracy of the measurement devices.

The single abandonment measure P{Ab} can be refined to account explicitly for those customers who were in fact well-served. We propose the following four-dimensional service measure:

- $P\{W \leq T; \mathrm{Sr}\}$ - fraction of well-served;
- $P\{W > T; \mathrm{Sr}\}$ - fraction served with a potential for improvement;
- $P\{W > \epsilon; \mathrm{Ab}\}$ - fraction of poorly-served;
- $P\{W \leq \epsilon; \mathrm{Ab}\}$ - fraction of those whose service-level is undetermined, as explained above.

The approximations that are developed in our paper cover this multi-dimensional performance measure. See Remark 4.4 on page 14.

1.4 The ED, QD and QED operational regimes

Table 1: **Example of Half-Hour ACD Report**

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
Total	20,577	19,860	3.5%	30	307	95.1%	
8:00	332	308	7.2%	27	302	87.1%	59.3
8:30	653	615	5.8%	58	293	96.1%	104.1
9:00	866	796	8.1%	63	308	97.1%	140.4
9:30	1,152	1,138	1.2%	28	303	90.8%	211.1
10:00	1,330	1,286	3.3%	22	307	98.4%	223.1
10:30	1,364	1,338	1.9%	33	296	99.0%	222.5
11:00	1,380	1,280	7.2%	34	306	98.2%	222.0
11:30	1,272	1,247	2.0%	44	298	94.6%	218.0
12:00	1,179	1,177	0.2%	1	306	91.6%	218.3
12:30	1,174	1,160	1.2%	10	302	95.5%	203.8
13:00	1,018	999	1.9%	9	314	95.4%	182.9
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**
14:00	1,173	1,082	7.8%	78	313	99.5%	188.9
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**
15:00	1,137	1,122	1.3%	15	320	96.9%	205.8
15:30	1,169	1,137	2.7%	17	311	97.1%	202.2
16:00	1,107	1,059	4.3%	46	315	99.2%	187.1
16:30	914	892	2.4%	22	307	95.2%	160.0
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**
17:30	420	420	0.0%	0	328	73.8%	103.5
18:00	49	49	0.0%	14	180	84.2%	5.8

Consider Table 1, which displays a typical daily ACD (Automatic Call Distributor) report of a moderate-to-large call center in the U.S., from the Health Insurance industry. For every half-hour interval, the report depicts the number of incoming calls, abandonment fraction, the Average Speed of Answer (ASA), the Average Handling Time (AHT), the agents' occupancy and the average number of agents over the interval in consideration.

We observe that the performance level, presented by ASA and Abn%, varies significantly over the day. We shall concentrate on three time intervals, highlighted in bold: 13:30, 14:30 and 17:00.

The first interval is characterized by 100% occupancy, relatively high abandonment rate (9.4%) and considerable ASA (more than 1 minute). During this half-hour, the call center is working in the *Efficiency-Driven (ED) regime*, in the sense that the emphasis is on agents' utilization, or efficiency. (Note that the number of agents is smaller than in the adjacent intervals. A probable cause could be lunch break.)

The interval that starts at 17:00 presents a contrasting service pattern. There is no abandonment and the average wait is negligible (2 sec). The agents' occupancy is far below 100% (83%). Such a service regime will be called *Quality-Driven (QD)*, in the sense that the emphasis is on customers' service quality.

Finally, the last interval (14:30) demonstrates an intermediate service regime: utilization is high (96.6%), and abandonment and waiting are neither negligible nor high. Since in this half-hour, high efficiency and service level are achieved simultaneously, this operational regime has been called *QED (Quality and Efficiency-Driven)*.

The examples above show that there exist clear differences in operational-performance which, as will now become clear, could be pre-designed (though we do not claim that this is the case here). We shall now present formal definitions of the three operational regimes.

First, calculate the *offered load* $R = \dfrac{\lambda}{\mu}$ for the three intervals. We get

$$R_{ED} \;=\; 1061 : \frac{1800}{306} \;=\; 180.37$$

for the ED-interval (1800 is the number of seconds in an interval), $R_{QD} = 112.07$ for the QD-interval, and, finally, $R_{QED} = 204.69$ for the QED-interval.

In the ED regime, we observe that the offered load R_{ED} (180.37) is considerably larger than the number of agents n (163.4). This implies that agents could not have coped with the offered load unless abandonment took place. The formal characterization of the ED regime is in terms of the following relationship between n and R_{ED}:

$$n \;=\; R_{ED} \cdot (1 - \gamma)\,, \tag{1.1}$$

where the constant $\gamma > 0$ is interpreted as a *service grade*: larger γ will imply larger wait and

abandonment. In our example, $\gamma = 1 - n/R_{ED} = 0.094$, which is equal to %Abn. This is not a coincidence, and an explanatory asymptotic statement will be presented in Theorem 6.1.

For the QD-regime, we have $R_{QD} = 112.07$ and $n = 135$. The characterization of this regime is

$$n \;=\; R_{QD} \cdot (1+\gamma)\,, \qquad \gamma > 0\,. \tag{1.2}$$

Now we proceed to, what we believe is, the most important operational regime: QED at 14:30. In this regime, the difference between n (206.1) and R_{QED} (204.69) is relatively small and should not be quantified in units of R, as in (1.1) and (1.2). Furthermore, this difference can be either positive or negative. The appropriate characterization turns out to be

$$n \;=\; R_{QED} + \beta\sqrt{R_{QED}}\,, \qquad -\infty < \beta < \infty\,, \tag{1.3}$$

where, in our example, the service grade $\beta = (n - R_{QED})/\sqrt{R_{QED}} = 0.10$

Above we established the need to study models with general patience of customers. In concert with this, Sections 4-6 will be dedicated to the operational regimes (1.1)-(1.3), within the M/M/n+G queue framework.

2 Structure of the paper and summary of results

2.1 Structure and summary of the paper

Section 3 contains a Literature Review. Subsections 3.1 and 3.2 cover exact and asymptotic results, respectively.

Sections 4-6 present a systematic treatment of the three operational regimes, described in Subsection 1.4: the QED, Quality-Driven (QD) and Efficiency-Driven (ED) regimes, respectively. We assume that the arrival rate λ and the number of agents n increase indefinitely, and the service rate μ is held fixed.

The QED operational regime is studied in Section 4. Theorem 4.1 considers patience distributions with a positive density at the origin. Approximations for performance measures are derived and an asymptotic linear relation between the probability to abandon and average wait is established. Theorems 4.2 and 4.3 treat the balking phenomenon: customers that do not get service immediately balk with probability P{Blk}.

Our main theoretical results on the QD operational regimes are summarized in Theorem 5.1. Then Theorem 6.1 explores the ED operational regime. Here, in contrast to the other two regimes, patience behavior near the origin does not determine the values of asymptotic performance measures.

The conclusions are presented in Section 7. Section 8 outlines some promising directions of ongoing and future research.

Our proofs are carried out within the framework of exact M/M/n+G calculations, presented in Section 9. We show there that all the essential performance measures can be calculated via several building blocks defined in formulae (9.1)-(9.5). These building blocks have an integral form. Hence, for asymptotic analysis, we need a technique for asymptotic calculations of integrals. Here the *Laplace method* is helpful, and its necessary background is developed in Section 10. Finally, selected proofs are presented in Section 11. Readers are referred to our Internet Supplement [48] for more details.

2.2 Summary of the Internet Supplement [48]

We start with the present summary in Section 1. Then, in Section 2 we briefly describe the results of Baccelli and Hebuterne [3] that are used in the following proofs. Sections 3 and 4 contain proofs of the results from Sections 9 and 10, respectively. Section 5 discusses relevant properties of the hazard rate of the standard normal random variable. Sections 6-8 contain proofs and additional numerical experiments for the three operational regimes: QED, QD and ED, respectively. We also study two additional special cases in the framework of the QED regime. (See Subsections 6.1.2 and 6.1.3 of the Internet Supplement.) In both cases, the density of the patience distribution vanishes at the origin.

Then Section 9 explores the Economies-of-Scale (EOS) problem for the three regimes. Specifically, assuming that the arrival rate increases by a factor $m > 1$, we apply the corresponding operational regime and check how the most important performance measures change in these circumstances. Finally, in Section 10 our models are applied to call center data of a large bank in the USA.

3 Literature review

3.1 Relevant exact results for M/M/n+G

The seminal work on queueing systems with impatient customers is Palm [36, 37]. In particular, Palm introduced the basic Erlang-A (M/M/n+M) queueing system with exponential patience times.

Gnedenko and Kovalenko [19] analyzed the M/M/n+D queueing system (deterministic patience times). Jurkevic [26] applied their methods to the general M/M/n+G system. Independently, the M/M/n+G queue was analyzed by Baccelli and Hebuterne [3] and Haugen and Skogan [23]. Boxma and de Waal [7] developed several approximations for the probability to abandon in the M/M/n+G queue and tested them via simulation.

The derivation of M/M/n+G performance measures continued in Brandt and Brandt [8, 9]. They considered the more general M(k)/M(k)/n+G system where arrival and service rates are

allowed to depend on the number k of calls in the system. However, some of the results in [8, 9] (for example, the distribution of total number-in-system) are new also for the M/M/n+G queue.

Another important branch of research is the estimation of the patience distribution in real tele-queue systems. Palm [36] introduced a mathematical model for irritation which postulated a Weibull distribution of patience times. Then he presented some real data that confirmed his hypothesis. Kort [27] also used the Weibull distribution to model patience while waiting for a dial tone. Baccelli and Hebuterne [3], using data from Roberts [38], fit it to an Erlang distribution with 3 phases. Brown et al. [10], in research on a bank call center, observed the patience times in the second plot of Figure 2. Finally, Daley and Servi [12] estimate Erlang-A parameters and performance characteristics (in particular, probability to abandon) given incomplete empirical data.

Concerning models of customers' impatience, readers are referred to the papers of Zohar, Mandelbaum and Shimkin [49] and Mandelbaum and Shimkin [28, 39], where it is assumed that customers adapt their patience to the waiting patterns they expect to encounter. For further references and a more complete survey see Gans, Koole and Mandelbaum [17].

3.2 Relevant asymptotic results

Although exact formulae for the Erlang-A and M/M/n+G queues are available, they are too complicated for developing guidelines for call center managers. These formulae cannot provide insight into practical questions of the type: "how many additional agents would one need if the arrival rate doubles?", "how sensitive is our model to a possible error in patience estimate?" etc.

Thus, approximations are useful for providing insight and simplifying computations. There exist two main types of approximations: *steady-state* (provide asymptotic expressions for steady-state performance measures like P{Ab} or E[W]) and *process-limit* (provide asymptotics for model processes like queue-length). In this paper, we develop steady-state approximations. Below we review relevant asymptotic results, emphasizing applications of the square-root QED staffing rule (Section 4) and of the ED operational regime (Section 6). In our research, which is oriented towards call centers, we are mainly interested in models with a large number of agents n. Formally, we assume that n and the arrival rate λ increase to infinity and, then, we index a sequence of models by either n or λ. (As a rule, we omit this indexing in formulae.)

The square-root staffing rule (1.3) was described already by Erlang [15], as early as 1924. 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 that paper, the authors establish an important relation: as λ increase indefinitely, sustaining the QED operational regime (1.3) with fixed $\beta > 0$ is equivalent to the delay probability converging to a fixed level α, $0 < \alpha < 1$.

Garnett, Mandelbaum and Reiman [18] studied the QED regime for Erlang-A with exponential abandonment, establishing results that are analogous to [21]. In this case, the service grade β can go negative, as in (1.3). QED analysis for the Erlang-B (M/M/n/n) model was carried out by Jagerman [24]. He showed that for the staffing rule (1.3) the blocking probability is of order $1/\sqrt{n}$. The GI/D/n queueing model (general iid interarrival times, deterministic service) in the QED framework was analyzed in Jelenković, Mandelbaum and Momčilović [25]. The M/M/n/k model with possible busy signals (n agents, $(k-n)$ waiting spaces in queue) was treated by Massey and Wallace [35].

In addition, it turns out that the QED staffing regime can be analyzed in some Skill-Based Routing (SBR) models. Armony and Mandelbaum [2], Gurvich [20] and Armony et al. [1] explore two classical and basic SBR models: $\bigwedge$ and $\bigvee$-designs.

In [46], Whitt develops and validates an approximation for the M/G/n+G model with generally distributed iid service times. In fact, he approximates it by an M/M/n+M(k) Markovian model with abandonment rates that depend on the number-in-system k. Whitt [47] provides additional insight into the approximation proposed in [46]. He compares between ED approximations of the two models: exact M/G/n+G and approximate M/M/n+M(k) of [46].

Whitt [43] also presents a general fluid model (ED approximation) for the G/G/n+G queue with general distributions of arrivals, services and patience times.

Since Erlang-A and other queueing models with abandonment are sensitive to changes in the arrival rate (see Whitt [44]), it is important to consider models with uncertainty about the arrival-rate. Recent papers of Whitt [45] and Harrison and Zeevi [22] study ED approximations for such models and develop asymptotic rules of optimal staffing. In addition, Bassamboo, Harrison and Zeevi [5] provides asymptotic routing methods. Note that the ED approximations are cruder than the QED ones, which enables the analysis of very general models.

Finally, Ward and Glynn [41, 42] use another type of scaling. They analyze one-server queues with abandonment (both exponential and general models), assuming that the arrival rate is close to the service rate and the patience times become large. In this case, the reflected Ornstein-Uhlenbeck process arises as a heavy traffic diffusion limit. In addition, the behavior of the patience distribution near the origin turns out to play an important role, which is in line with some of our results here.

4 The QED operational regime

4.1 Formulation of results

4.1.1 Main case: patience distribution with positive density at the origin

Denote the patience-time density by $g = \{g(x), x \geq 0\}$, assuming that the density exists at the origin and its value $g(0) \stackrel{\Delta}{=} g_0$ is strictly positive.

In most applications that we have encountered, a non-negligible abandonment rate during the first seconds of wait was observed. Hence, there is a practical motivation to treat the case $g_0 > 0$, as the main one. In addition, it will turn out that there are significant theoretical reasons for this emphasis. (E.g. see Remark 4.7.)

Consider an M/M/n+G queue. Fix the service rate μ and the patience distribution G. Assume that the arrival rate $\lambda \to \infty$ and the staffing level n is given by

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \qquad \lambda \to \infty\,, \quad -\infty < \beta < \infty\,. \tag{4.1}$$

Define the hazard rate of the standard normal distribution by

$$h(x) \;=\; \frac{\phi(x)}{1-\Phi(x)} \;=\; \frac{\phi(x)}{\bar{\Phi}(x)}\,. \tag{4.2}$$

($\Phi(x)$ 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.)

Let

$$\hat{\beta} \;\stackrel{\Delta}{=}\; \beta\sqrt{\frac{\mu}{g_0}}\,. \tag{4.3}$$

Finally, in Theorem 4.1 and later, $f \sim g$ stands for $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) = 1$.

Theorem 4.1 (QED performance measures) In the QED operational regime, namely $\lambda \to \infty$ and n as in (4.1), the performance measures of the M/M/n+G queueing system are approximated by:

a. The delay probability converges to a constant that depends on β and the ratio g_0/μ:

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{4.4}$$

In addition, if $\lambda \to \infty$ and $\mathrm{P}\{\mathrm{W} > 0\} \to \alpha$, with $0 < \alpha < 1$, then

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})\,, \tag{4.5}$$

where $\alpha = \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}$.

b. The probability to abandon of delayed customers decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.6}$$

The probability to abandon $\mathrm{P}\{\mathrm{Ab}\}$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (4.4) with (4.6).

c. The average offered wait of delayed customers decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.7}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (4.4) and (4.7).

d. The average waiting time is of the same order as the average offered wait:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W>0] \;\sim\; \mathrm{E}[V \mid V>0]\,. \tag{4.8}$$

e. The ratio between the probability to abandon and average wait converges to the (positive) value of patience density at the origin:

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}|W>0\}}{\mathrm{E}[W|W>0]} \;\sim\; g_0\,. \tag{4.9}$$

f. The average offered wait and the average actual wait of abandoning customers decrease at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.10}$$

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{2\sqrt{g_0\mu}} \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right), \tag{4.11}$$

or, in other words,

$$\mathrm{E}[W|\mathrm{Ab}] \;\sim\; \frac{1}{2} \cdot \mathrm{E}[V|\mathrm{Ab}]\,. \tag{4.12}$$

g. The asymptotic distribution of wait, or *Total Service Factor(TSF)*, is given by the product of the right-hand side of (4.4) with

$$\mathrm{P}\left\{\frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \,\middle|\, W>0\right\} \;\sim\; \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{\frac{g_0}{\mu}} \cdot t\right)}{\bar{\Phi}(\hat{\beta})}, \qquad t \geq 0\,, \tag{4.13}$$

where $\mathrm{E}[S] = 1/\mu$ is the average service time.

h. The probability to abandon, given delay in queue, is asymptotically equal to

$$\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, \frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{g_0}{\mu}}\right) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.14}$$

i. The average wait, given delay in queue, is asymptotically equal to

$$\mathrm{E}\left[W \,\middle|\, \frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right] \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{g_0\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{g_0}{\mu}}\right) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.15}$$

Parts **h** and **i** together imply a generalization of part **e**:

$$\frac{\mathrm{P}\left\{\mathrm{Ab} \,|\, W > t/\sqrt{n}\right\}}{\mathrm{E}\left[W \,|\, W > t/\sqrt{n}\right]} \;\sim\; g_0\,, \qquad t \geq 0. \tag{4.16}$$

Remark 4.1 (Role of g_0) The patience density at the origin g_0 plays a major role in our approximations, which is to be expected. Indeed, since the waiting time in the QED regime is small (converges to zero), the patience distribution near the origin determines asymptotic behavior of the system.

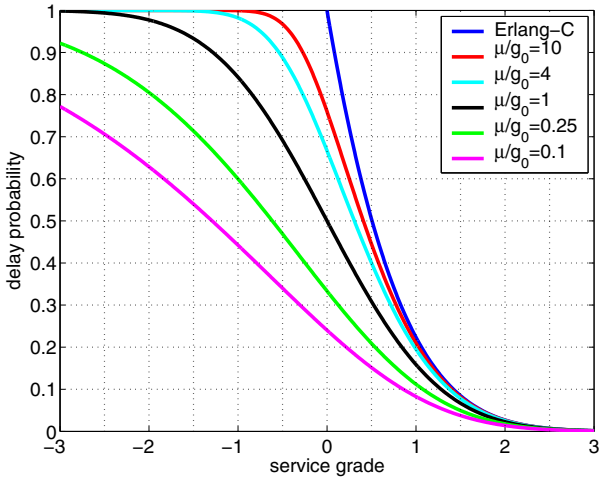
Remark 4.2 (Queue-length) According to the Little's Law $\mathrm{E}[Q] = \lambda \mathrm{E}[W]$, where Q is the steady-state queue-length. It follows that the average-in-queue is of order $O(\sqrt{n})$. The asymptotic distribution of Q is an interesting issue for future research.

Remark 4.3 The asymptotic statement (4.9), combined with results from Mandelbaum and Zeltyn [32], provides additional support for the practically observed linear relation between the probability to abandon and average wait. (Recall Figure 4.)

Remark 4.4 Theorem 4.1 enables one to calculate the four service measures from Subsection 1.3. Specifically, Parts **b**, **g** and **h** provide us with $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{P}\{W > T\}$ and $\mathrm{P}\{\mathrm{Ab}|W > T\}$, respectively. The product of the last two is equal to $\mathrm{P}\{W > T; \mathrm{Ab}\}$. The other three service measures are easily derived. For example, $\mathrm{P}\{W > T; \mathrm{Sr}\} \;=\; \mathrm{P}\{W > T\} - \mathrm{P}\{W > T; \mathrm{Ab}\}$.

Remark 4.5 Figure 5 illustrates the dependence (4.4) between the service grade β and the delay probability, over varying values of the ratio μ/g_0. In addition, we plotted the Halfin-Whitt curve [21] for the Erlang-C queue, which is meaningful for positive β only. Note that for large values of μ/g_0 (very patient customers) the Erlang-A curves are close to the Erlang-C curve.

Figure 5: **Asymptotic relations between service grade and delay probability**



Remark 4.6 (The special case $\beta = 0$) Note that when $\beta = 0$ in (4.1), the staffing level asymptotically corresponds to the simple rule that does not take into account stochastic considerations: assign the number of agents equal to the *offered load* λ/μ. In Erlang-C, this "naive" approach would lead to system instability. However, in M/M/n+G (which is a much better fit to the real world of call centers than Erlang-C) one would get a reasonable-to-good performance level. For example, if the service rate μ is equal to the individual abandonment rate θ, and $\beta = 0$, 50% of customers would get service immediately upon arrival. (Check Figure 5. Note that for Erlang-C, 50% delay probability corresponds to $\beta \approx 0.5$.) This suggests why some call centers that are managed using simplified deterministic models actually perform at reasonable service levels. (They obtain the "right answer" for the "wrong reasons".)

Remark 4.7 (Relation to Garnett, Mandelbaum and Reiman [18]) Formula (4.4) generalizes the statement for the Erlang-A queue (exponential patience), derived in Garnett et al. [18]. Namely,

$$\mathrm{P}\{W > 0\} \;\sim\; w\left(-\beta, \sqrt{\mu/\theta}\right) , \tag{4.17}$$

where

$$w(x,y) \;=\; \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}$$

and θ is the abandonment rate (parameter of the exponential patience). Straightforward calculations reveal the equivalence between formulae (4.17) and (4.4), if we substitute g_0 instead of θ in (4.17). (Note that θ is indeed the density of $\exp(\theta)$ at the origin.)

Approximations for other performance measures (for example, the probability to abandon and average wait) were also derived in [18]. However, they do not coincide exactly with our

approximations. The reason is that in Theorem 4.1 the lead asymptotic term is always presented explicitly with respect to n or λ. On the other hand, the approximation formulae in [18] do not display the lead term. For example, the analogue in [18] to our formula (4.6) is as follows:

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;\approx\; \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta}+\sqrt{\theta/(n\mu)})}\,.$$

Remark 4.8 The relation (4.12) can be explained in the following way. The offered wait in the QED regime is small. Since a positive density at the origin exists, the conditional waiting-time distribution of abandoning customers is approximately uniform on $[0, V]$. Therefore, $\mathrm{E}[W|\mathrm{Ab}] \sim (1/2)\cdot\mathrm{E}[V|\mathrm{Ab}]$.

Remark 4.9 In the Internet Supplement [48], we prove that

$$\frac{1}{2}\cdot\left[\frac{1}{h(\hat\beta)-\hat\beta}-\hat\beta\right] \;<\; h(\hat\beta)-\hat\beta \;<\; \frac{1}{h(\hat\beta)-\hat\beta}-\hat\beta\,, \qquad -\infty<\hat\beta<\infty. \tag{4.18}$$

In conjunction with Parts **c**, **d** and **f**, (4.18) implies corresponding asymptotic order relations between $\mathrm{E}[W|\mathrm{Ab}]$, $\mathrm{E}[W|W>0]$, $\mathrm{E}[V|V>0]$ and $\mathrm{E}[V|\mathrm{Ab}]$. In words, the average offered wait of *abandoning* customers exceeds (asymptotically) the average actual wait of *delayed* customers which, in turn, exceeds the average actual wait of abandoning customers.

4.1.2 Patience with balking

Assume that the patience-time distribution has an atom at the origin. In other words, if wait is encountered, customers abandon immediately with probability $\mathrm{P}\{\mathrm{Blk}\}>0$, or $\bar G(0)=1-\mathrm{P}\{\mathrm{Blk}\}$. From a practical point of view, this means that some customers refuse to wait at all. (Readers surely recall such a situation from personal experience.)

Remark 4.10 Consider two events:

- $\{V>0\}$, which means that "a customer did not get service immediately";
- $\{W>0\}$, which means "positive actual wait".

In fact, $\{W>0\}$ is identical to $\{V>0, \tau>0\}$.

In Theorem 4.1, we did not distinguish between $\{V>0\}$ and $\{W>0\}$ since $\mathrm{P}\{\tau=0\}=0$. However, in the models with balking one must be careful in this respect.

Assume, in addition, that the survival function $\bar G$ is differentiable at the origin: $\bar G'(0)=-g_0$. (Here g_0 is the right-side derivative of the patience-time distribution function G at the origin.)

Theorem 4.2 (QED performance measures with balking) Under the QED operational regime (4.1), the performance measures of the M/M/n+G queue with balking are approximated by:

a. Probability to encounter a queue decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{V > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.19}$$

The delay probability decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{W > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{(1 - \mathrm{P}\{\mathrm{Blk}\}) \cdot h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.20}$$

b. Conditional probability to abandon $\mathrm{P}\{\mathrm{Ab}|V > 0\}$ converges to the balking probability:

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \mathrm{P}\{\mathrm{Blk}\} + \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{4.21}$$

Conditional probability to abandon $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ decreases at rate $\frac{1}{n}$:

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\} \cdot (1 - \mathrm{P}\{\mathrm{Blk}\})} + o\left(\frac{1}{n}\right). \tag{4.22}$$

The unconditional probability to abandon decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1}{\sqrt{n}} \cdot h(-\beta) + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.23}$$

c. Conditional average offered wait $\mathrm{E}[V|V > 0]$ decreases at rate $\frac{1}{n}$:

$$\mathrm{E}[V|V > 0] \;=\; \frac{1}{n} \cdot \frac{1}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{4.24}$$

The average offered wait decreases at rate $\frac{1}{n^{3/2}}$:

$$\mathrm{E}[V] \;=\; \frac{1}{n^{3/2}} \cdot \frac{h(-\beta)}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^{3/2}}\right). \tag{4.25}$$

d. Conditional average waiting time $\mathrm{E}[W|W > 0]$ decreases at rate $\frac{1}{n}$:

$$\mathrm{E}[W|W > 0] \;=\; \frac{1}{n} \cdot \frac{1}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{4.26}$$

The average wait $\mathrm{E}[W]$ decreases at rate $\frac{1}{n^{3/2}}$:

$$\mathrm{E}[W] \;=\; \frac{1}{n^{3/2}} \cdot \frac{(1 - \mathrm{P}\{\mathrm{Blk}\}) \cdot h(-\beta)}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^{3/2}}\right). \tag{4.27}$$

Remark 4.11 In contrast to Theorem 4.1, the probabilities of wait (both $\mathrm{P}\{W > 0\}$ and $\mathrm{P}\{V > 0\}$) decrease at rate $O(1/\sqrt{n})$. If the balking probability $\mathrm{P}\{\mathrm{Blk}\} = 1$, (4.19) is equivalent to Jagerman's [24] QED result for M/M/n/n (Erlang-B). In addition, (4.23) demonstrates that the M/M/n+G queue with any positive fraction of balking implies, in the QED regime, the same fraction of lost customers as in M/M/n/n. In this sense, Balking turns out to be equivalent to Blocking.

Formula (4.21) provides insight into this striking similarity. We observe that in the M/M/n+G queue with balking, the fraction of customers that abandon after positive wait is negligible (the second term of (4.21)), which makes it similar to Erlang-B in which all lost customers abandon immediately.

Note that, given positive offered wait, a fixed proportion of the customers abandon, and the system is similar to M/M/n+G in the quality-driven regime (fixed ρ). This is the reason why the second term of (4.21) and formula (4.24) will have counterparts in the quality-driven results (5.4) and (5.5) later on.

4.1.3 Patience with scaled balking

Below we treat a special case of M/M/n+G which, in practical terms, corresponds to small yet non-negligible balking.

Assume that the balking probability depends on the system size n. Specifically, let $\mathrm{P}_n\{\mathrm{Blk}\} = \frac{p_b}{\sqrt{n}}$, for some $p_b > 0$. Assume that the derivative of the survival function $\bar{G}_n$ at the origin is independent of the system size: $\bar{G}_n'(0) = -g_0$.

Theorem 4.3 (Performance measures) Under the QED operational regime (4.1), the performance measures of the M/M/n+G queue with the scaled balking are approximated by:

a. The delay probability converges to a constant that depends on β, p_b and $\frac{g_0}{\mu}$:

$$\mathrm{P}\{V > 0\} \;\sim\; \mathrm{P}\{W > 0\} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{4.28}$$

where

$$\hat{\beta} \;\triangleq\; (\beta + p_b) \cdot \sqrt{\frac{\mu}{g_0}}\,. \tag{4.29}$$

b. Conditional probabilities to abandon decrease at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{4.30}$$

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1}{\sqrt{n}}\cdot\sqrt{\frac{g_0}{\mu}}\cdot\left[h(\hat{\beta})-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right). \tag{4.31}$$

The unconditional probability to abandon P{Ab} also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (4.30) and (4.28).

c. Conditional average offered wait $\mathrm{E}[V|V>0]$ decreases at rate $\frac{1}{\sqrt{n}}$:

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{g_0\mu}}\left[h(\hat{\beta})-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right). \tag{4.32}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (4.32) and (4.28).

d. The average waiting time is equivalent to the average offered wait:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V];\quad \mathrm{E}[W\mid W>0] \;\sim\; \mathrm{E}[V\mid V>0]\,. \tag{4.33}$$

e. The ratio between the probability to abandon of delayed customers and average wait of delayed customers converges to the value of the patience density at the origin:

$$\frac{\mathrm{P}\{\mathrm{Ab}|W>0\}}{\mathrm{E}[W|W>0]} \;\sim\; g_0\,. \tag{4.34}$$

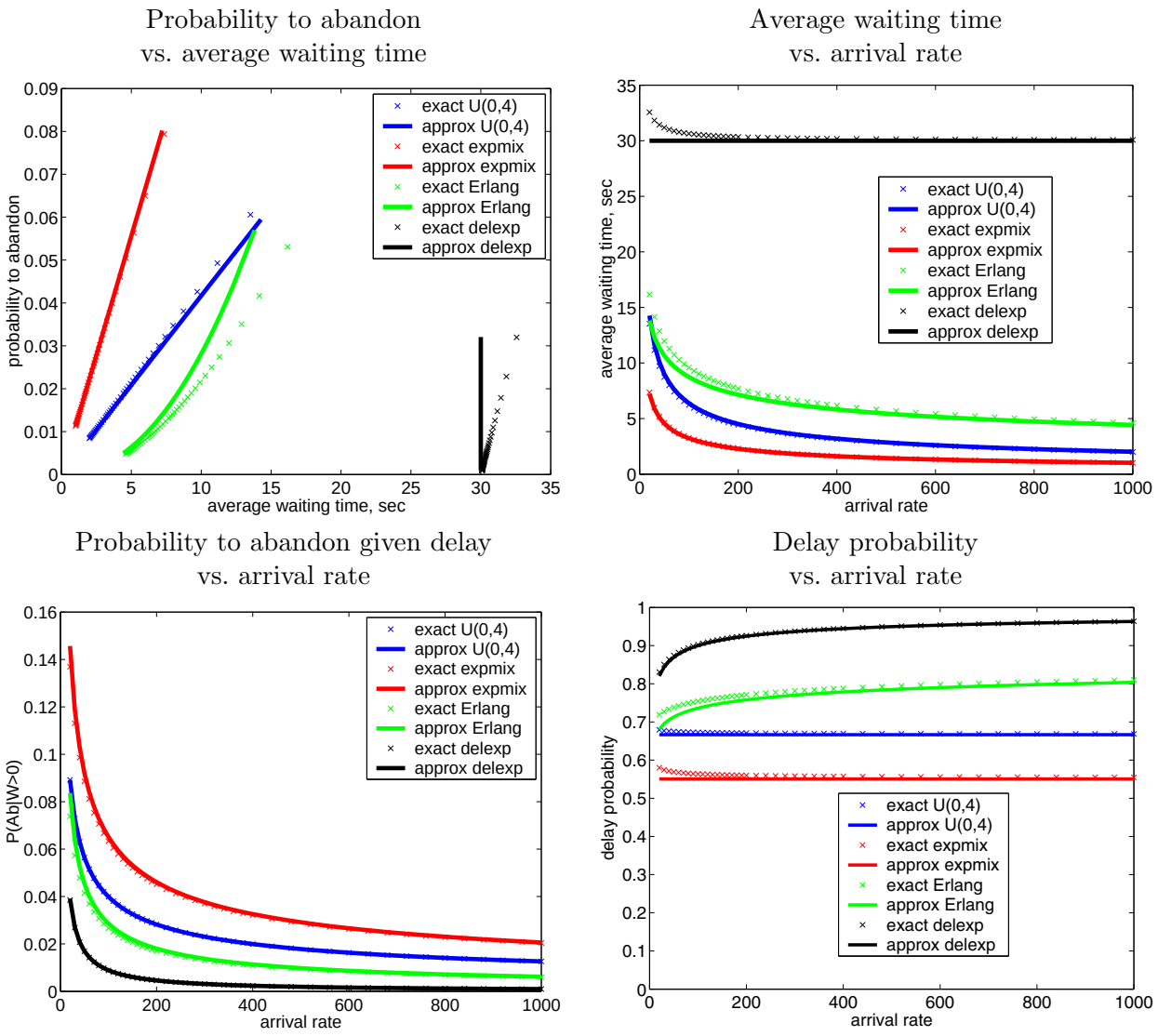
Remark 4.12 Under scaled balking, we observe a clear similarity with the main case described in Theorem 4.1 (positive patience density at the origin). Some results (formulae (4.28), (4.30) and (4.32)) have exact counterparts in Theorem 4.1: the service grade β should be replaced by $(\beta+p_b)$. We also derived the linear relation (4.34), although this result does not prevail for the corresponding unconditional performance measures.

4.2 Numerical example

We proceed with analyzing the quality of the QED approximations. Four patience distributions are considered, all with their means equal to 2:

- *Uniform* distribution on [0,4]: illustrates Theorem 4.1 with $g_0 = 0.25$;
- *Hyperexponential* distribution (mixture of two exponentials, with means 1 and 3 respectively): conforms to Theorem 4.1 with $g_0 = 2/3$;
- *Erlang* (Gamma) distributions, two exponential phases, each with the mean equal to 1: Theorem 6.2 from the Internet Supplement [48];
- *Delayed exponential* distribution equal to 1 + exp(mean=1): Theorem 6.4 from the Internet Supplement [48].

Figure 6: **Service grade $\beta = 0$, performance measures and approximations**



We assume the service grade $\beta = 0$. (See Subsection 6.2 of the Internet Supplement [48] for experiments with other values of β.) M/M/n+G queues with the service rate $\mu = 1$ are considered. Arrival rate λ increases from 20 to 1000 with a varying step (44 values of λ overall). The number of agents n increases according to the QED staffing rule:

$$n \;=\; \left[\frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}\right], \tag{4.35}$$

where, as usual, the square brackets in (4.35) denote the nearest integer value.

Then, for each M/M/n+G queue in consideration, exact (Section 9) and approximate calculations are performed. The results are presented in four graphs. The first graph presents

a scatterplot of the probability to abandon against average wait. The other three plots show three different performance characteristics, as they change with the arrival rate: average wait, the probability to abandon of delayed customers and the delay probability. Solid lines are for approximations, and x's are for exact values.

Figure 6 demonstrates a very good fit between approximations and exact values for $\lambda > 100$ (and the fit is reasonable even for small arrival rates starting with $\lambda = 20$). Note the straight-line curves for the first two distributions in the first plot. (The two distributions with $g_0 = 0$ give rise to non-linear curves.)

5 The Quality-Driven operational regime

The quality-driven (QD) operational regime is defined by

$$n = \frac{\lambda}{\mu} \cdot (1+\gamma) + o(\sqrt{\lambda}), \qquad \gamma > 0, \tag{5.1}$$

as λ and n increase indefinitely. In the QD regime, the offered load per agent

$$\rho \;=\; \frac{\lambda}{n\mu} \;\to\; \frac{1}{1+\gamma} \;<\; 1.$$

If $o(\sqrt{\lambda}) \equiv 0$ in (5.1), the relation between ρ and γ is exact and given by:

$$\rho = \frac{1}{1+\gamma} \quad \text{and} \quad \gamma = \frac{1-\rho}{\rho}. \tag{5.2}$$

Theorem 5.1 (QD performance measures) Assume that the density of the patience time at the origin exists and is positive: $g_0 > 0$. Then the performance measures of the M/M/n+G queue in the QD regime are approximated by:

a. The delay probability decreases exponentially in n. Specifically,

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{\gamma} \cdot \left(\frac{1}{1+\gamma}\right)^{n-1} \cdot \exp\left\{\frac{\lambda\gamma}{\mu}\right\}. \tag{5.3}$$

b. Probability to abandon given wait:

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{1+\gamma}{\gamma} \cdot \frac{g_0}{\mu} + o\left(\frac{1}{n}\right) \;=\; \frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{g_0}{\mu} + o\left(\frac{1}{n}\right). \tag{5.4}$$

(Note that if the $o(\sqrt{\lambda})$ deviation term in (5.1) is not equal to zero, the two $o(1/n)$ terms in (5.4) will not be identical.)

c. Average offered waiting time:

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{n} \cdot \frac{1+\gamma}{\gamma} \cdot \frac{1}{\mu} + o\left(\frac{1}{n}\right) \;=\; \frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{1}{\mu} + o\left(\frac{1}{n}\right). \tag{5.5}$$

d. Average waiting time:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]\,. \tag{5.6}$$

e. Ratio between the probability to abandon and average wait:

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;\sim\; g_0\,. \tag{5.7}$$

f. Total Service Factor:

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{t}{n} \;\middle|\; W > 0 \right\} \;\sim\; e^{-(1-\rho)t}\,. \tag{5.8}$$

Remark 5.1 Assume that the staffing level (5.1) is kept exact: $n = \dfrac{\lambda}{\mu} \cdot (1+\gamma)$. Then the asymptotic formula for the delay probability transforms to:

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{1-\rho} \cdot (\rho e^{1-\rho})^n \qquad (n \to \infty)\,.$$

Remark 5.2 If the deviation in (5.1) is larger than $o(\sqrt{\lambda})$, for example,

$$n = \frac{\lambda}{\mu} \cdot (1+\gamma) + o(\lambda)\,,$$

formulae (5.4)-(5.8) still prevail. However, the approximation (5.3) can go wrong.

Subsection 7.3 of the Internet Supplement [48] contains numerical experiments, where the QD approximations are compared with the QED approximations and the exact M/M/n+G performance measures. The main conclusion is that, unless wait and abandonment are very small, the QED approximations are preferable over the QD ones. The QD formulae should be applied only if the delay probability is or should be less than 5-10%.

6 The Efficiency-Driven operational regime

6.1 Formulation of results

In the Efficiency-Driven (ED) operational regime, staffing is determined by:

$$n = \frac{\lambda}{\mu} \cdot (1-\gamma) + o(\sqrt{\lambda})\,, \qquad \gamma > 0\,. \tag{6.1}$$

The offered load per agent

$$\rho \;=\; \frac{\lambda}{n\mu} \;\to\; \frac{1}{1-\gamma} \;>\; 1\,.$$

If $o(\sqrt{\lambda}) \equiv 0$ in (6.1), the relation between ρ and γ is given by

$$\rho = \frac{1}{1-\gamma} \quad \text{and} \quad \gamma = \frac{\rho - 1}{\rho}\,. \tag{6.2}$$

Theorem 6.1 (ED performance measures) Assume that the equation

$$G(x) \;=\; \gamma$$

has a unique solution x^*. Assume further that the patience density at x^* is positive: $g(x^*) > 0$. Then the performance measures of the M/M/n+G queue in the ED operational regime are approximated by:

a. Probability to get service immediately decreases exponentially:

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; \frac{1}{\gamma} \cdot \sqrt{\frac{g(x^*)}{2\pi\lambda}} \cdot \exp\{-\lambda k(\gamma)\}\,. \tag{6.3}$$

where

$$k(\gamma) \;\stackrel{\Delta}{=}\; x^* \cdot \left(1 - \frac{n\mu}{\lambda}\right) - \int_0^{x^*} G(u)du\,. \tag{6.4}$$

b. Probability to abandon converges to the constant $\gamma \approx 1 - \dfrac{1}{\rho}$:

$$\mathrm{P}\{\mathrm{Ab}\} \;\sim\; \gamma\,. \tag{6.5}$$

c. The average offered wait $\mathrm{E}[V]$ converges to the constant x^*:

$$\mathrm{E}[V] \;\sim\; x^*\,. \tag{6.6}$$

The offered wait also converges to x^* in probability:

$$V \;\stackrel{p}{\to}\; x^*\,. \tag{6.7}$$

d. Define the distribution $G^* = \{G^*(x),\; x \geq 0\}$ by

$$G^*(x) \;=\; \begin{cases} \frac{G(x)}{G(x^*)} \;=\; \frac{G(x)}{\gamma}, & x \leq x^* \\ 1, & x > x^* \end{cases}$$

(In fact, G^* is the distribution of the random variable $\min(x^*, \tau)$, where τ is the patience time.) Then the average waiting time W weakly converges to the distribution G^*:

$$W \;\stackrel{w}{\to}\; G^*\,. \tag{6.8}$$

In addition,

$$\mathrm{E}[W] \;\to\; \mathrm{E}[\min(x^*, \tau)] \;=\; \int_0^{x^*} \bar{G}(u)du\,. \tag{6.9}$$

e. Total Service Factor:

The asymptotic distribution of wait is given by:

$$\mathrm{P}\{V > t\} \;\sim\; \begin{cases} 1, & t < x^* \\ 0, & t > x^* \end{cases}\,. \tag{6.10}$$

$$\mathrm{P}\{W > t\} \;\sim\; \begin{cases} \bar{G}(t), & t < x^* \\ 0, & t > x^* \end{cases} . \tag{6.11}$$

The distribution of wait around x^* can be approximated in the following way.
Let $-\infty < t < \infty$. Then

$$\mathrm{P}\left\{ \frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;\sim\; \bar{\Phi}\left(t \sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) . \tag{6.12}$$

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \;\middle|\; V > 0 \right\} \;\sim\; (1-\gamma) \cdot \bar{\Phi}\left(t \sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) . \tag{6.13}$$

Remark 6.1 In the ED regime, the average waiting time does not converge to zero. Therefore, in contrast to the QED and QD regimes, the patience density at the origin does not play an important role in the ED approximations.

Remark 6.2 ED limits for the probability to abandon and waiting time can be obtained using "fluid" (deterministic) considerations and hence are sometimes referred to as *fluid limits*. For example, see results in Whitt [43] that are closely related to (6.5) and (6.9).

Remark 6.3 Assume that the staffing level (6.1) is kept exact: $n = (\lambda/\mu) \cdot (1-\gamma)$. Then we can rewrite definition (6.4) as

$$k(\gamma) \;=\; \gamma x^* - \int_0^{x^*} G(u)du \, .$$

6.2 Numerical example

Three distributions that were used in Subsection 4.2, are studied again: uniform, hyperexponential and delayed exponential. Instead of the conditional probability $\mathrm{P}\{\mathrm{Ab}|W > 0\}$, we plot $\mathrm{P}\{\mathrm{Ab}\}$ (the delay probability is close to one and there are no reasons to distinguish between the two performance measures). Note that, in contrast to the QED regime, the ED approximation formulae are the same for distributions with both positive and zero densities at the origin.

The ED staffing rule is

$$n = \left[\frac{\lambda}{\rho\mu} \right] , \qquad \rho > 1 \, . \tag{6.14}$$

The value $\gamma = 1/6$, which corresponds to $\rho = 1.2$ is chosen. Other assumptions are the same as in Subsection 4.2.

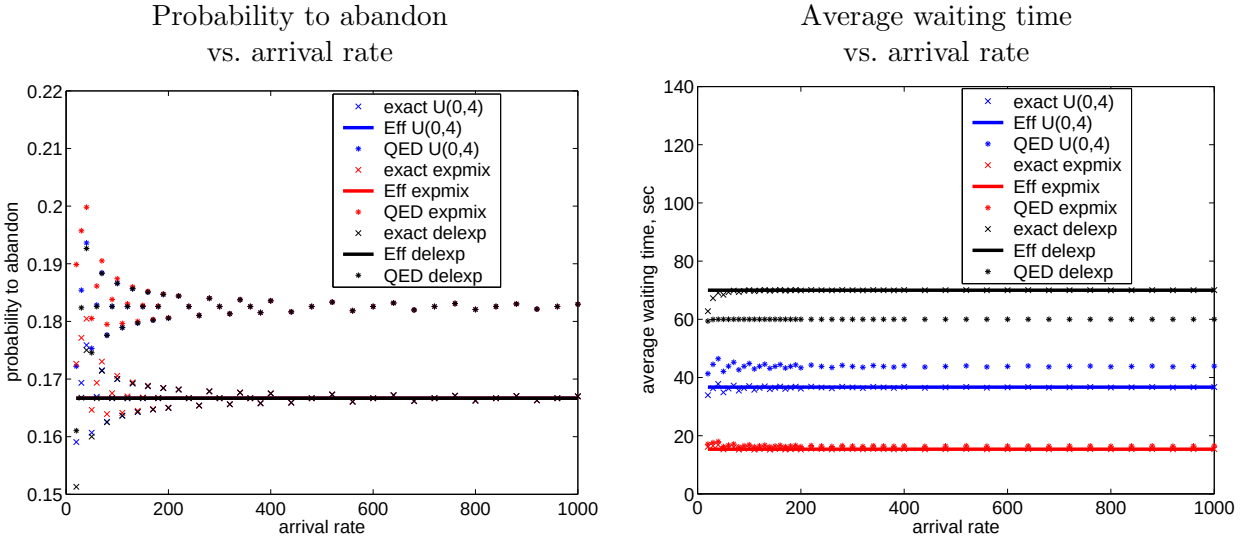
In addition, we calculate the QED approximations using

$$\beta \;=\; \frac{n - \lambda/\mu}{\sqrt{\lambda/\mu}} ,$$

and compare them with the ED approximations.

In Figure 7, we observe that the ED approximations for P{Ab} and E[W] converge to constants (fluid limits) predicted by Theorem 6.1. The corresponding QED approximations do not converge to fluid limits. In fact, the QED approximations for P{Ab} are close to $-\beta/\sqrt{n}$, which converges to $\sqrt{\rho}\cdot(1-1/\rho)$. It differs from the proper fluid limit by a factor $\sqrt{\rho}$. However, if ρ is close to one (e.g. 1.05), the QED approximations can be preferable, see Subsection 8.2 of the Internet Supplement [48].

Figure 7: **Offered load per server $\rho = 1.2$, performance measures and approximations**



7 Conclusions

Here we summarize our conclusions on the three operational regimes, analyzed in Sections 4-6, and on the relation P{Ab}/E[W] that has been explored in different contexts in this research.

QED regime. In contrast to the exact M/M/n+G formulae, our QED approximations can be applied using any software that provides the standard normal distribution (e.g. Excel). We observe that these approximations work very well for a wide range of M/M/n+G parameters.

Our rule-of-thumb recommendations for the use of QED formulae are the following:

- Number of servers n=10's to 1000's;
- Agents highly utilized but not overloaded (~90-98%);
- Delay probability 10-90%;

- Probability to abandon: 3-7% for small n, 1-4% for large n.

ED regime. These approximations are relatively simple to apply as well, although they require solving the equation $G(x) = \gamma$, and then integration (calculating $H(x^*)$). Both can be performed either numerically or analytically, depending on the patience distribution. We suggest to use the ED approximations if:

- Number of servers $n \geq 100$. (One can cautiously use n=10's, if the probability to abandon is large (>10%).)
- Agents very highly utilized (>95%);
- Delay probability: more than 85%;
- Probability to abandon: more than 5%.

QD regime. The QD approximations should be applied only for very high-performance systems. (For example, in emergency call centers.)

Linear P{Ab}/E[W] relation. In the QED and QD operational regimes, the linear relation P{Ab}/E[W] prevails. Such a relation is also observed in practice, as demonstrated in Figure 4. Summarizing these facts and those established in Mandelbaum and Zeltyn [32], we conclude that this phenomenon prevails in a very broad context: exact M/M/n+G performance measures, different approximations and real data.

On practical implementation of the QED approximations. Based on our theoretical results, the density of the patience time at the origin plays a critical role in the QED (and QD) approximations. What are the practical consequences of this fact in the context of call centers?

The good news is that approximate performance measures depend only on the patience distribution near the origin, which can be estimated via the Kaplan-Meier estimator (see, for example, [11] or the Appendix of [49]). In other words, one does not need to infer the tail of the patience distribution, the latter being usually a hard problem.

However, estimating the patience density at the origin and substituting this estimate into the QED formulae can lead to unsatisfactory results (see Section 10 of the Internet Supplement [48]). The reason is that the patience density can oscillate during the first seconds of wait and, therefore, the limit results may not apply even for $n = 100$'s.

We offer two approaches to deal with this problem: first, one can estimate g_0 as the *average* density during the first x (say, 5-10) seconds of wait. The second approach is based on the linear

relation (4.9), substituting the ratio P{Ab}/E[W] instead of g_0. (This approach was used in Section 10 of the Internet Supplement [48] and in [10].)

In addition, if there exists very significant abandonment during the first seconds of wait, models that incorporate balking can be applicable (Theorems 4.2 and 4.3).

8 Ongoing and future research

Finally, we outline some directions worthy of further research.

Dimensioning the M/M/n+G queue. In our context, the term *dimensioning* was introduced in Borst, Mandelbaum and Reiman [6]. The authors considered an optimization problem for the Erlang-C queue, where the goal is to minimize the sum of *staffing costs* and *waiting costs*. In other words, [6] developed a formal framework for the problem of the *quality-efficiency tradeoff*, discussed in Subsection 1.1 of the Introduction. The ongoing research [34] is dedicated to similar problems for the M/M/n+G queue. In addition to staffing and waiting costs, abandonment costs arise in this case. For a wide set of system parameters, a comparison between the asymptotic QED staffing and exact optimal staffing demonstrates that the two staffing rules are almost identical.

Queues with random arrival rate. In Brown et al. [10] it was shown that the Poisson arrival rate in an Israeli call center varies from day to day and its prediction 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 a random variable.

If $\mathrm{E}(\Lambda) \to \infty$ and its standard deviation is of the order $\sqrt{\mathrm{E}(\Lambda)}$, we expect that the QED operational regime and the square-root staffing rule will arise again. However, if $\sigma(\Lambda)$ is of the order $\mathrm{E}(\Lambda)$, the "cruder" ED regime seems to be the most appropriate; see Whitt [45], and Bassamboo, Harrison and Zeevi [5].

Queues with time-inhomogeneous arrival rates. 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 (as in this paper) during periods when the arrival rate is judged constant. An implicit assumption is that the arrival rate is slow-varying with respect to the durations of services.

Recently, Feldman et al. [16] developed an alternative simulation-based algorithm for staffing time-varying queues with abandonment. The algorithm is designed to achieve a given constant probability of delay, generalizing the QED operational regime to queues with non-homogeneous arrival rates. It is found in [16] that, with a proper definition of a time-varying offered load

$\{R_t,\, t \geq 0\}$, square-root staffing of the form

$$n_t \;\approx\; R_t + \beta\sqrt{R_t}$$

leads to a constant probability of delay α, where the relation between α and β is exactly that of a corresponding time-homogeneous queue.

Data analysis. Additional studies of customers' patience in tele-service should be performed. Currently data collection in two large banks in the U.S. and Israel, and in an Israeli cellular-phone company is in progress. If some special structures of patience are discovered, this will lead to more detailed and specialized results, for which the present paper is a natural starting point.

Generally distributed service times: M/G/n+G. In our research, we assumed exponential service times. However, this assumption seems to not apply for many call centers. In several application we encountered (see [10], for example), the lognormal distribution provides an excellent approximation for service times. Therefore, it is very important to study the M/G/n+G model with generally distributed service times. However, exact analysis of the M/G/n+G queue seems prohibitively difficult, hence one should probably resort to approximations (see Whitt [46]) and simulation (see Mandelbaum and Schwartz [29]).

Process-limit results for M/M/n+G. In this paper, we focused on steady-state results, both exact and approximate, for many M/M/n+G performance measures. Of interest are also analogous process-limit results, as in Garnett et al. [18] for Erlang-A.

9 M/M/n+G queue: summary of performance measures

Here we summarize *exact* formulae for M/M/n+G performance measures. The following definitions and statements are largely based on Baccelli and Hebuterne [3], but many of them are presented here for the first time.

Building blocks. Define $H(x) \;\triangleq\; \int_0^x \bar{G}(u)du$. Note that $H(\infty) = \bar{\tau}$, where $\bar{\tau}$ is the mean patience-time.
Introduce the integrals

$$J(t) \;\triangleq\; \int_t^\infty \exp\left\{\lambda H(x) - n\mu x\right\} dx\,, \tag{9.1}$$

$$J_1(t) \;\triangleq\; \int_t^\infty x \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx\,, \tag{9.2}$$

$$J_H(t) \;\triangleq\; \int_t^\infty H(x) \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx\,. \tag{9.3}$$

In addition, let

$$J \triangleq J(0),\ J_1 \triangleq J_1(0),\ J_H \triangleq J_H(0)\,. \tag{9.4}$$

Finally, define

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}} = \int_0^\infty e^{-t}\left(1+\frac{t\mu}{\lambda}\right)^{n-1} dt\ . \tag{9.5}$$

Remark 9.1 A convenient way to calculate $\mathcal{E}$ is via recursion: define

$$\mathcal{E}_k \triangleq \frac{\sum_{j=0}^{k} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^{k}}, \qquad k \geq 0,$$

and use

$$\mathcal{E}_0 = 1\,; \qquad \mathcal{E}_k = 1 + \frac{k\mu}{\lambda}\cdot\mathcal{E}_{k-1}\,, \quad 1 \leq k \leq n-1\,; \qquad \mathcal{E} = \mathcal{E}_{n-1}\,.$$

List of performance measures: Many important performance measures of the M/M/n+G queue can be conveniently expressed via the building blocks above and the patience distribution G. For example,

$$\mathrm{P}\{V>0\} = \frac{\lambda J}{\mathcal{E}+\lambda J}\,, \tag{9.6}$$

$$\mathrm{P}\{W>0\} = \frac{\lambda J}{\mathcal{E}+\lambda J}\cdot\bar{G}(0)\,, \tag{9.7}$$

$$\mathrm{P}\{\mathrm{Ab}\mid \mathrm{V}>0\} = \frac{1+(\lambda-n\mu)J}{\lambda J}\,, \tag{9.8}$$

$$\mathrm{E}[V] = \frac{\lambda J_1}{\mathcal{E}+\lambda J}\,, \tag{9.9}$$

$$\mathrm{E}[V\mid V>0] = \frac{J_1}{J}\,, \tag{9.10}$$

$$\mathrm{E}[W] = \frac{\lambda J_H}{\mathcal{E}+\lambda J}\,, \tag{9.11}$$

$$\mathrm{E}[V\mid \mathrm{Ab}] = \frac{(\lambda-n\mu)J_1+J}{(\lambda-n\mu)J+1}\,, \tag{9.12}$$

$$\mathrm{E}[W\mid \mathrm{Ab}] = \frac{J+\lambda J_H-n\mu J_1}{(\lambda-n\mu)J+1}\,, \tag{9.13}$$

$$\mathrm{P}\{V>t\} = \frac{\lambda J(t)}{\mathcal{E}+\lambda J}\,, \tag{9.14}$$

$$\mathrm{P}\{W>t\} = \frac{\lambda\bar{G}(t)J(t)}{\mathcal{E}+\lambda J}\,, \tag{9.15}$$

$$\mathrm{E}[V \mid V > t] = \frac{J_1(t)}{J(t)}\,, \tag{9.16}$$

$$\mathrm{E}[W \mid W > t] = \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t)J(t)}\,, \tag{9.17}$$

$$\mathrm{P}\{\mathrm{Ab} \mid V > t\} = \frac{\lambda - n\mu}{\lambda} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda J(t)}\,, \tag{9.18}$$

$$\mathrm{P}\{\mathrm{Ab} \mid W > t\} = \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t)J(t)}\,. \tag{9.19}$$

See Section 3 of the Internet Supplement [48] for the proofs of formulae (9.6)-(9.19). The handout of Mandelbaum and Zeltyn [31], prepared for a large U.S. bank, contains a summary of exact and approximate performance measures of M/M/n+G.

10 Asymptotic behavior of integrals

We have seen that the building blocks of the M/M/n+G model have an integral form (recall formulae (9.1)-(9.5)). In Sections 4-6 we shall calculate various approximations for these building blocks and, consequently, for the M/M/n+G performance measures. To this end, we now develop a general method and prove several lemmata that will help us in the task.

10.1 The Laplace method

In our proofs, we repeatedly derive asymptotic approximations for integrals that are expressed in the form

$$\int_0^\infty x^m \cdot e^{-f_\lambda(x)} dx\,, \qquad \lambda \to \infty\,. \tag{10.1}$$

As a rule, $f_\lambda(0) = 0$, for all $\lambda > 0$, and $f_\lambda(x) \to \infty$, as $\lambda \to \infty$, for all $x > 0$. Note that the exponential term rapidly converges to zero, for $x > 0$. Hence, one could expect that, as $\lambda \to \infty$, the value of (10.1) depends mainly on the behavior of the integrand near the origin.

An important special case is given by

$$\int_0^\infty x^m \cdot \exp\left\{-b\lambda^k x^l\right\} dx = \frac{\Gamma\left(\frac{m+1}{l}\right)}{l b^{\frac{m+1}{l}}} \cdot \lambda^{-\frac{k(m+1)}{l}}\,, \tag{10.2}$$

where $k \geq 0$, $l > 0$, $b > 0$ and $m \geq 0$. If $m = 0$ one gets

$$\int_0^\infty \exp\left\{-b\lambda^k x^l\right\} dx = \frac{\Gamma\left(\frac{1}{l}\right)}{l b^{1/l}} \cdot \lambda^{-k/l}\,. \tag{10.3}$$

But, generally, (10.1) cannot be calculated analytically, in which case we derive its approximation in the spirit of de Bruijn [14]. The general approach is to show that $\int_\delta^\infty x^m \cdot e^{-f_\lambda(x)} dx$ is negligible for some $\delta > 0$ (δ can depend on λ). Then $\int_0^\delta x^m \cdot e^{-f_\lambda(x)} dx$ is approximated using the Taylor expansion of $f_\lambda(x)$ near the origin and formulae (10.2)-(10.3) above.

This technique is referred to in [14] as the *Laplace method* for the calculation of integrals. We now apply it to derive several asymptotic statements.

10.2 Asymptotic results

Lemma 10.1 Let b_1, k_1, l_1, l_2 be positive numbers and let b_2, k_2, m be non-negative. In addition, assume that l_1 and l_2 are integers. Consider a function $r_1 = \{r_1(\lambda), \quad \lambda > 0\}$ such that $r_1(\lambda) \sim \lambda^{k_1}, \quad \lambda \to \infty$. Finally, assume that

$$\frac{k_1}{l_1} > \frac{k_2}{l_2}\,. \tag{10.4}$$

Then

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx$$

$$= \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} + o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right), \qquad \lambda \to \infty\,, \tag{10.5}$$

and

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx$$

$$= \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 [b_1 r_1(\lambda)]^{\frac{m+1}{l_1}}} - \frac{b_2 \Gamma\left(\frac{m+l_2+1}{l_1}\right)}{l_1 b_1^{\frac{m+l_2+1}{l_1}}} \cdot \lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}} + o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right). \tag{10.6}$$

Remark 10.1 Note that the main term in the right hand side of (10.5) is equal to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 \lambda^{k_1} x^{l_1}\right\} dx\,.$$

Thus, the relation (10.4) determines the "dominant" term in the exponent. Moreover, the second term in (10.6) is equal to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 \lambda^{k_1} x^{l_1}\right\} \cdot b_2 \lambda^{k_2} x^{l_2} dx\,.$$

Therefore, Lemma 10.1 states, informally, that

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 \lambda^{k_2} x^{l_2}\right\} dx \approx \int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 \lambda^{k_2} x^{l_2}] dx\,.$$

We have seen that the asymptotic value of the integral from Lemma 10.1 is determined by the inequality (10.4). What happens if the two ratios from (10.4) are equal? For our needs, it is sufficient to answer this question for the two special cases, presented in Lemmata 10.2 and 10.3. Their proofs can be obtained via straightforward calculations.

Lemma 10.2 Assume that b_1 is an arbitrary number and that $b_2 > 0$. Then

$$\int_0^\infty \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\}dx \;=\; \frac{1}{\sqrt{2b_2\lambda}} \cdot \frac{1}{h(b_1/\sqrt{2b_2})}\,,$$

where $h(\cdot)$ is the hazard rate of the standard normal distribution.

Lemma 10.3 Under the conditions of Lemma 10.2,

$$\int_0^\infty x \cdot \exp\{-b_1\sqrt{\lambda}x - b_2\lambda x^2\}dx \;=\; \frac{1}{2b_2\lambda} \cdot \left[1 - \frac{b_1}{\sqrt{2b_2}} \cdot \frac{1}{h(b_1/\sqrt{2b_2})}\right].$$

See Section 4 of the Internet Supplement [48] for the proofs.

11 Selected proofs

11.1 Proof of Theorem 4.1, a-e

We start with a lemma that provides asymptotics for several building blocks, introduced in Section 9. That will make it easy to derive approximations for the performance measures.

Lemma 11.1 (Building blocks) Under the assumptions of Theorem 4.1, the building blocks J, $\mathcal{E}$, J_1, and J_H, defined in Section 9, are approximated by:

a.

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \tag{11.1}$$

where $\hat{\beta}$ was defined in (4.3).

b.

$$J_1 \;=\; \frac{1}{n} \cdot \frac{1}{\mu g_0}\left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right), \qquad J_H \;=\; \frac{1}{n} \cdot \frac{1}{\mu g_0}\left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right). \tag{11.2}$$

c.

$$\mathcal{E} \;=\; \sqrt{n} \cdot \frac{1}{h(-\beta)} + o(\sqrt{n})\,. \tag{11.3}$$

Proof of Lemma 11.1.

a. First, we present the proof for the case when the QED staffing rule prevails *exactly*:

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}}\,. \tag{11.4}$$

Define

$$u_\lambda(x) \;\triangleq\; \int_0^x \left[\lambda(\bar{G}(u) - 1) - \beta\sqrt{\lambda\mu}\right] du\,. \tag{11.5}$$

Then, under the staffing rule (11.4)

$$J \;=\; \int_0^\infty \exp\{u_\lambda(x)\}dx\,. \tag{11.6}$$

Since $\bar{G}(0) = 1$ and $\bar{G}'(0) = -g_0$, $\forall \epsilon > 0\ \exists\, \delta > 0$ such that

$$|\bar{G}(u) - 1 + g_0 \cdot u| \le \epsilon u \quad \text{for } u \in [0, \delta]\,. \tag{11.7}$$

(See, for example, de Bruijn [14], page 65.) Using (11.5)-(11.7), we get

$$\begin{aligned} \int_0^\delta \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda(g_0+\epsilon)x^2}{2}\right\}dx \;&\le\; \int_0^\delta \exp\left\{u_\lambda(x)\right\}dx \\ &\le\; \int_0^\delta \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda(g_0-\epsilon)x^2}{2}\right\}dx\,. \end{aligned} \tag{11.8}$$

Now we construct a bound for $\int_\delta^\infty \exp\{u_\lambda(x)\}dx$, showing that, given $\lambda \to \infty$, the asymptotic behavior of $\int_0^\infty \exp\{u_\lambda(x)\}dx$ depends only on the values of $u_\lambda(x)$ near the origin.

Since $g_0 > 0$, the patience survival function $\bar{G}$ is strictly decreasing at the origin. Take

$$\alpha \;\triangleq\; 1 - \frac{1+\bar{G}(\delta/2)}{2} > 0\,. \tag{11.9}$$

Then, for λ large enough,

$$\begin{aligned} u_\lambda(x) \;&=\; \int_0^x \left[\lambda(\bar{G}(u)-1) - \beta\sqrt{\lambda\mu}\right]du \;=\; \int_0^{\delta/2}\cdots + \int_{\delta/2}^x \cdots \\ &\le\; -\frac{\delta}{2}\beta\sqrt{\lambda\mu} - \int_{\delta/2}^x \alpha\lambda du \;=\; -\frac{\delta}{2}\beta\sqrt{\lambda\mu} - \alpha\lambda\left(x - \frac{\delta}{2}\right). \end{aligned}$$

Integrating,

$$\begin{aligned} \int_\delta^\infty \exp\left\{u_\lambda(x)\right\}dx \;&\le\; \exp\left\{\frac{\alpha\lambda\delta}{2} - \frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}\cdot\frac{e^{-\alpha\lambda\delta}}{\alpha\lambda} \\ &=\; \frac{\exp\left\{-\frac{\alpha\lambda\delta}{2} - \frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}}{\alpha\lambda} = o\left(e^{-\nu\lambda}\right), \quad \nu > 0\,. \end{aligned}$$

In other words,

$$\left|\int_0^\delta \exp\left\{u_\lambda(x)\right\}dx - \int_0^\infty \exp\left\{u_\lambda(x)\right\}dx\right| \;=\; o\left(e^{-\nu\lambda}\right). \tag{11.10}$$

Using identical arguments, the same relation between $\int_0^\delta$ and $\int_0^\infty$ can be derived for the two other integrals from (11.8).

Now, applying Lemma 10.2 to (11.8), we derive that

$$\frac{1}{\sqrt{\lambda(g_0+\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0+\epsilon}}\right)} + o\left(e^{-\nu\lambda}\right) \;\le\; J \;\le\; \frac{1}{\sqrt{\lambda(g_0-\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0-\epsilon}}\right)} + o\left(e^{-\nu\lambda}\right).$$

If λ is large enough,

$$(1-\epsilon)\frac{1}{\sqrt{\lambda(g_0+\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0+\epsilon}}\right)} \;\le\; J \;\le\; (1+\epsilon)\frac{1}{\sqrt{\lambda(g_0-\epsilon)}}\frac{1}{h\left(\beta\sqrt{\frac{\mu}{g_0-\epsilon}}\right)}\,.$$

Since ϵ is arbitrary, and

$$\lambda \sim n\mu \qquad (\lambda, n \to \infty)$$

in the QED regime, we get the statement (11.1).

Finally, assume that the QED staffing rule prevails asymptotically in the sense of (4.1). Then $\forall \tilde{\epsilon} > 0$, for large λ we have

$$\frac{\lambda}{\mu} + (1-\tilde{\epsilon})\beta\sqrt{\frac{\lambda}{\mu}} \;\le\; n \;\le\; \frac{\lambda}{\mu} + (1+\tilde{\epsilon})\beta\sqrt{\frac{\lambda}{\mu}}\,,$$

and

$$\int_0^\infty \exp\left\{\int_0^x \left[\lambda\bar{G}(u) - (\lambda + (1+\tilde{\epsilon})\beta\sqrt{\lambda\mu})\right] du\right\} dx \;\le\; J$$
$$\le\; \int_0^\infty \exp\left\{\int_0^x \left[\lambda\bar{G}(u) - (\lambda + (1-\tilde{\epsilon})\beta\sqrt{\lambda\mu})\right] du\right\} dx\,.$$

Now we can proceed with the proof above for the exact QED staffing and get the same statement using that $\tilde{\epsilon}$ is arbitrary.

Remark 11.1 The following three main ideas, that are a part of the so-called *Laplace method* (see de Bruijn [14]), were applied in the proof above:

- *Taylor expansion* is used in order to approximate sub-integral expressions in the building blocks near the origin;
- *Exponential bounds* for $\int_\delta^\infty$ integrals are developed;
- In the end, we explain how to replace the exact staffing (11.4) by the asymptotic staffing (4.1).

These three steps should be used repeatedly in the proofs of Theorems 4.1, 4.2 and 4.3.

b. The proof is similar to Part **a**. Lemma 10.3 is used instead of Lemma 10.2.

c. In the QED regime,

$$\mathcal{E} \;=\; \int_0^\infty e^{-t}\left(1+\frac{\mu t}{\lambda}\right)^{\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1} dt\,.$$

Changing variables: ($t = \lambda x$, $x = \frac{t}{\lambda}$), we get

$$\mathcal{E} \;=\; \lambda\int_0^\infty \exp\left\{-\lambda x + \left(\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}-1\right)\ln(1+\mu x)\right\} dx\,. \tag{11.11}$$

It is well known that

$$\ln(1+\mu x) = \mu x - \frac{\mu^2 x^2}{2} + O(x^3)\,, \qquad x \to 0\,. \tag{11.12}$$

Substitute into formula (11.11) the first two terms of the Taylor expansion (11.12):

$$\begin{aligned}
\mathcal{E}_A \;&\triangleq\; \lambda \cdot \int_0^\infty \exp\left\{-\lambda x + \left(\frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} - 1\right)\left(\mu x - \frac{\mu^2 x^2}{2}\right)\right\} dx \\
&= \;\lambda \cdot \exp\left\{\frac{(\beta\sqrt{\lambda\mu} - \mu)^2}{\lambda\mu + \beta\sqrt{\lambda\mu^3} - \mu^2}\right\} \cdot \int_0^\infty \exp\left\{-\left(\lambda\mu + \beta\sqrt{\lambda\mu^3} - \mu^2\right)\left[\frac{x - \frac{\beta\sqrt{\lambda\mu}-\mu}{\lambda\mu+\beta\sqrt{\lambda\mu^3}-\mu^2}}{2}\right]^2\right\} dx \\
&\sim \;\exp\left\{\frac{\beta^2}{2}\right\} \cdot \lambda\sqrt{\frac{2\pi}{\lambda\mu + \beta\sqrt{\lambda\mu^3} - \mu^2}} \cdot \Phi\left(\frac{\beta\sqrt{\lambda\mu} - \mu}{\sqrt{\lambda\mu + \beta\sqrt{\lambda\mu^3} - \mu^2}}\right) \\
&\sim \;\sqrt{\frac{2\pi\lambda}{\mu}} \cdot \exp\left\{\frac{\beta^2}{2}\right\} \cdot \Phi(\beta) \;\sim\; \frac{\sqrt{n}}{h(-\beta)}\,.
\end{aligned}$$

The Laplace argument, based on the Taylor expansion (see Remark 11.1), ensures that $\mathcal{E} \sim \mathcal{E}_A$, given $\lambda \to \infty$.

Proof of Theorem 4.1.

a. Formula (9.6), Lemma 11.1, parts **a** and **b**, and the equivalence $\lambda \sim n\mu$ $(n \to \infty)$, imply that

$$\mathrm{P}\{W > 0\} \;=\; \frac{\lambda J}{\mathcal{E} + \lambda J} \;\sim\; \frac{\frac{\sqrt{n\mu}}{h(\hat{\beta})\sqrt{g_0}}}{\frac{\sqrt{n\mu}}{h(\hat{\beta})\sqrt{g_0}} + \frac{\sqrt{n}}{h(-\beta)}} \;=\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}.$$

Now we must prove the opposite direction: if the probability of wait converges to a constant, then the QED staffing prevails. The probability-of-wait function

$$P_{g_0,\mu}(\beta) \;\triangleq\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}$$

is monotonically decreasing in β. Hence, the inverse function $P^{-1}_{g_0,\mu}(\alpha)$, $0 < \alpha < 1$, is well-defined.

Assume that

$$\mathrm{P}_{\lambda,n_\lambda}\{W > 0\} \;\to \alpha, \quad 0 < \alpha < 1, \tag{11.13}$$

and take $\beta = P^{-1}_{g_0,\mu}(\alpha)$. We must show that for all $\epsilon > 0$ and λ large enough,

$$\frac{\lambda}{\mu} + (\beta - \epsilon)\sqrt{\frac{\lambda}{\mu}} \;\le\; n_\lambda \;\le\; \frac{\lambda}{\mu} + (\beta + \epsilon)\sqrt{\frac{\lambda}{\mu}}\,. \tag{11.14}$$

Consider the staffing levels

$$n^1_\lambda \;=\; \left\lceil \frac{\lambda}{\mu} + (\beta - \epsilon)\sqrt{\frac{\lambda}{\mu}} \right\rceil \quad \text{and} \quad n^2_\lambda \;=\; \left\lfloor \frac{\lambda}{\mu} + (\beta + \epsilon)\sqrt{\frac{\lambda}{\mu}} \right\rfloor .$$

According to formula (4.4),

$$\mathrm{P}_{\lambda,n_\lambda^1}\{W>0\} \;\to\; P_{g_0,\mu}(\beta-\epsilon) \;=\; \alpha+\delta_1, \quad \delta_1>0,$$

and

$$\mathrm{P}_{\lambda,n_\lambda^2}\{W>0\} \;\to\; P_{g_0,\mu}(\beta+\epsilon) \;=\; \alpha-\delta_2, \quad \delta_2>0.$$

Therefore, for λ large enough,

$$\mathrm{P}_{\lambda,n_\lambda^1}\{W>0\} > \alpha+\frac{\delta_1}{2} \quad \text{and} \quad \mathrm{P}_{\lambda,n_\lambda^2}\{W>0\} < \alpha-\frac{\delta_2}{2}\,.$$

We know that $\mathrm{P}\{W>0\}$ is monotonically decreasing in the staffing level n. This fact and (11.13) imply that for λ large enough $n_\lambda^1 < n_\lambda < n_\lambda^2$, which proves (11.14).

b. Note that the definition of the QED regime implies

$$\lambda - n\mu = -\beta\sqrt{\lambda\mu} + o(\sqrt{\lambda})\,.$$

Applying the above to formula (9.8) and using the approximation for J from Lemma 11.1, we get the expression for the conditional probability to abandon.

c. Direct consequence of (9.10).

d. Follows from (9.6), (9.10), (9.11) and Part **c** of Lemma 11.1.

e. Follows from **b**, **c** and **d**.

11.2 Proof of Theorem 5.1, a-e

Lemma 11.2 (Building blocks) Under the assumptions of Theorem 5.1:

a.

$$J \;=\; \frac{1}{n\mu-\lambda} - \frac{g_0}{\lambda^2\gamma^3} + o\left(\frac{1}{\lambda^2}\right). \tag{11.15}$$

b.

$$J_1 \;=\; \frac{1}{(n\mu-\lambda)^2} - \frac{3g_0}{\lambda^3\gamma^4} + o\left(\frac{1}{\lambda^3}\right). \tag{11.16}$$

c.

$$\mathcal{E} \;\sim\; \sqrt{2\pi n}\cdot(1+\gamma)^{n-1}\cdot\exp\left\{-\frac{\lambda\gamma}{\mu}\right\}. \tag{11.17}$$

Proof of Lemma 11.2.

a. Lemma 10.1 with

$$m=0, \quad k_1=1, \quad l_1=1, \quad k_2=1, \quad l_2=2,$$

implies that

$$J_A \;\triangleq\; \int_0^\infty \exp\left\{-\lambda\gamma x - f(\lambda)\mu x - \frac{\lambda g_0 x^2}{2}\right\} dx$$

$$= \frac{1}{\lambda\gamma + f(\lambda)\mu} - \frac{1}{\lambda^2}\,\frac{g_0}{\gamma^3} + o\left(\frac{1}{\lambda^2}\right) = \frac{1}{n\mu - \lambda} - \frac{1}{\lambda^2}\,\frac{g_0}{\gamma^3} + o\left(\frac{1}{\lambda^2}\right), \tag{11.18}$$

where the relation $n\mu - \lambda = \lambda\gamma + f(\lambda)\mu$ follows from the staffing rule (5.1), and $f(\lambda)$ denotes the deviation term $o(\sqrt{\lambda})$ from (5.1).

Standard Laplace arguments from Lemma 11.1 ensure that

$$J = \int_0^\infty \exp\left\{\int_0^x \left[\lambda(\bar{G}(u) - 1)\right] du - \lambda\gamma x - f(\lambda)\mu x\right\} dx$$

can be substituted into (11.18) instead of J_A.

b. The proof is very similar to part **a**.

c. Recall that

$$\mathcal{E} = \int_0^\infty e^{-t}\left(1 + \frac{t\mu}{\lambda}\right)^{\frac{\lambda}{\mu}(1+\gamma)+f(\lambda)-1} dt = \lambda\int_0^\infty e^{-\lambda x}(1+\mu x)^{\frac{\lambda}{\mu}(1+\gamma)+f(\lambda)-1} dx\,. \tag{11.19}$$

Now perform the change of variables $y = x - \gamma/\mu$. (If the "$f(\lambda) - 1$" term in the power is not taken into account, the expression under the integral (11.19) reaches a maximum at γ/μ.)

$$\mathcal{E} = \lambda\exp\left\{-\lambda\frac{\gamma}{\mu}\right\}\cdot\int_{-\gamma/\mu}^\infty \exp\left\{-\lambda y + \left[\frac{\lambda}{\mu}(1+\gamma) + f(\lambda) - 1\right]\cdot\ln(1+\gamma+\mu y)\right\} dy. \tag{11.20}$$

We approximate $\mathcal{E}$, replacing the logarithm in (11.20) by the first three terms of the Taylor expansion above and changing integral limits to $\int_{-\infty}^{\infty}$:

$$\mathcal{E}_A = \lambda\exp\left\{-\lambda\frac{\gamma}{\mu}\right\}\cdot(1+\gamma)^{n-1}\cdot\int_{-\infty}^\infty \exp\left\{(f(\lambda)-1)\cdot\left[\frac{\mu y}{1+\gamma} - \frac{\mu^2 y^2}{2(1+\gamma)^2}\right] - \frac{\lambda\mu y^2}{2(1+\gamma)}\right\} dy\,.$$

The last term in the exponent above determines the asymptotic value of the integral. For example, using $f(\lambda) = o(\sqrt{\lambda})$,

$$\int_{-\infty}^\infty \exp\left\{f(\lambda)\cdot\frac{\mu y}{1+\gamma} - \frac{\lambda\mu y^2}{2(1+\gamma)}\right\} dy = \int_{-\infty}^\infty \exp\left\{-\frac{\lambda\mu}{2(1+\gamma)}\cdot\left[y - \frac{f(\lambda)}{\lambda}\right]^2 + \frac{\mu(f(\lambda))^2}{2\lambda(1+\gamma)}\right\} dy$$
$$\sim \sqrt{\frac{2\pi}{\lambda}}\,\sqrt{\frac{1+\gamma}{\mu}}\,.$$

Therefore, taking into account $n \sim \dfrac{\lambda}{\mu}\cdot(1+\gamma), \quad \lambda\to\infty,$

$$\mathcal{E}_A \sim \sqrt{2\pi n}\cdot(1+\gamma)^{n-1}\cdot\exp\left\{-\frac{\lambda\gamma}{\mu}\right\}.$$

In order to validate the same result for $\mathcal{E}$ we apply the Laplace argument, based on the following inequality: $\forall\,\epsilon > 0 \quad \exists\,\delta > 0$ such that

$$\left|\ln(1+\mu y+\gamma) - \ln(1+\gamma) - \frac{\mu y}{1+\gamma} + \frac{\mu^2 y^2}{2(1+\gamma)^2}\right| \le \frac{\epsilon\mu^2 y^2}{2(1+\gamma)^2} \quad \text{for } y\in[-\delta,\delta].$$

Proof of Theorem 5.1. In most cases, the proof is a straightforward application of the formulae in Lemma 11.2.

a. Note that

$$\frac{1}{n\mu-\lambda} \sim \frac{1}{\lambda\gamma}.$$

Then

$$\mathrm{P}\{W>0\} = \frac{\lambda J}{\mathcal{E}+\lambda J} \sim \frac{\lambda J}{\mathcal{E}} \sim \frac{1}{\sqrt{2\pi n}}\cdot\frac{1}{\gamma}\cdot\left(\frac{1}{1+\gamma}\right)^{n-1}\cdot\exp\left\{\frac{\lambda\gamma}{\mu}\right\}.$$

b.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} = \frac{1+(\lambda-n\mu)J}{\lambda J} \sim \frac{(n\mu-\lambda)g_0}{\lambda^3\gamma^3 J} \sim \frac{g_0}{\lambda^2\gamma^2 J} \sim \frac{g_0}{\lambda\gamma}.$$

Recall that $\lambda\sim n\mu\rho$ and $\gamma\sim\dfrac{1-\rho}{\rho}$, $(n\to\infty)$. This implies

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \sim \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{g_0}{\mu}.$$

c.

$$\mathrm{E}[V|V>0] = \frac{J_1}{J} \sim \frac{1}{n\mu-\lambda} \sim \frac{1}{\lambda\gamma} \sim \frac{1}{n}\cdot\frac{1}{1-\rho}\cdot\frac{1}{\mu}.$$

d. The proof is similar to part **d** of Lemma 11.1.

e. A direct consequence of parts **b-d**.

References

[1] Armony M., Gurvich I. and Mandelbaum A. (2004) Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers. Working paper. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 3.2

[2] Armony M. and Mandelbaum A. (2004) Design, staffing and control of large service systems: The case of a single customer class and multiple server types. Working paper. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 3.2

[3] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: Kylstra F.J. (Ed.), *Performance '81*. North-Holland Publishing Company, 159-179. 1.2, 2.2, 3.1, 9

[4] Bain P. and Taylor P. (2002) Consolidation, "Cowboys" and the developing employment relationship in British, Dutch and US call centres. In: Holtgrewe U., Kerst C. and Shire K. (Ed.), *Re-Organising Service Work*. Ashgate Publishing Limited, 42-62. 1.1

[5] Bassamboo A., Harrison J.M. and Zeevi A. (2004) Design and control of a large call center: asymptotic analysis of an LP-based method. Submitted for publication. 3.2, 8

[6] Borst S., Mandelbaum A. and Reiman M. (2004). Dimensioning large call centers. *Operations Research*, 52(1), 17-34. 8

[7] Boxma O.J. and de Waal P.R. (1994) Multiserver queues with impatient customers. *ITC*, 14, 743-756. 3.1

[8] Brandt A. and Brandt M. (1999) On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35, 1-18. 3.1

[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. 3.1

[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. 3.1, 7, 8, 8

[11] Cox D.R. and Oakes D. (1984) Analysis of Survival Data, Chapman and Hall. 7

[12] Daley D.J. and Servi L.D. (2000) Estimating customer loss rates from transactional data. In: Shanthikumar J.G. and Sumita U. (ed.): *International Series in Operations Research and Management Science*, 313-332. 3.1

[13] Datamonitor. `http://www.datamonitor.com`. 1.1

[14] de Bruijn N.G. (1981) *Asymptotic Methods in Analysis*, Dover. 10.1, 11.1, 11.1

[15] 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. 3.2

[16] Feldman Z., Mandelbaum A., Massey W. and Whitt W. (2004) Staffing of time-varying queues to achieve time-stable performance. Submitted to *Management Science*. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 8

[17] Gans N., Koole G. and Mandelbaum A. (2003) Telephone call centers: a tutorial and literature review. Invited review paper, *Manufacturing and Service Operations Management*, 5 (2), 79-141. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 1.1, 3.1

[18] Garnett O., Mandelbaum A. and Reiman M. (2002) Designing a telephone call-center with impatient customers. *Manufacturing and Service Operations Management* 4, 208-227. 1.2, 1.2, 3.2, 4.7, 4.7, 8

[19] Gnedenko B.W. and Kovalenko I.N. (1968) *Introduction to Queueing Theory*, Jerusalem, Israel Program for Scientific Translations. 3.1

[20] Gurvich I. (2004) Design and control of the M/M/N queue with multi-class customers and many servers. M.Sc. Thesis, Technion, 2004. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 3.2

[21] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588. 3.2, 4.5

[22] Harrison J.M. and Zeevi A. (2004) A method for staffing large call centers using stochastic fluid models. To appear in *Manufacturing & Service Operations Management*. 3.2

[23] Haugen R.B. and Skogan E. (1980) Queueing systems with stochastic time out. *IEEE Trans. Commun.* COM-28, 1984-1989. 3.1

[24] Jagerman D.L. (1974) Some properties of the Erlang loss function. *Bell Systems Technical Journal*, 53, 525-551. 3.2, 4.11

[25] Jelencović P., Mandelbaum A. and Momčilović P. (2004) Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications (QUESTA)*, 47, 53-69. 3.2

[26] Jurkevic O.M. (1971) On many-server systems with stochastic bounds for the waiting time (in Russian), *Izv. Akad. Nauk SSSR Techniceskaja kibernetika*, 4, 39-46. 3.1

[27] Kort B.W. (1983) Models and methods for evaluating customer acceptance of telephone connections. *GLOBECOM '83*, IEEE, 706-714. 3.1

[28] Mandelbaum A. and Shimkin N. (2000) A model for rational abandonment from invisible queues. *Queueing Systems: Theory and Applications (QUESTA)*, 36, 141-173. 1.1, 3.1

[29] Mandelbaum A. and Schwartz R. (2002) Simulation experiments with M/G/100 queues in the Halfin-Whitt (QED) regime. Technical Report, Technion. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 8

[30] Mandelbaum A. and Zeltyn S. (2004) The Palm/Erlang-A queue, with applications to call centers. Teaching note to *Service Engineering* course. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 1.2

[31] Mandelbaum A. and Zeltyn S. (2004) The M/M/n+G queue: summary of performance measures. Teaching note to *Service Engineering* course. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 9

[32] Mandelbaum A. and Zeltyn S. (2004) The impact of customers patience on delay and abandonment: some empirically-driven experiments with the M/M/N+G queue. *OR Spectrum*, 26 (3), 377-411. Special Issue on Call Centers. 1.2, 4.3, 7

[33] Mandelbaum A. and Zeltyn S. (2004) Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. Invited chapter to *IAO* book project. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 1.2

[34] Mandelbaum A. and Zeltyn S. (2005) Dimensioning M/M/n+G queue. Working paper. 8

[35] Massey A.W. and Wallace B.R. (2004) "An optimal design of the M/M/C/K queue for call centers", to appear in *Queueing Systems*. 3.2

[36] Palm C. (1953) Methods of judging the annoyance caused by congestion. *Tele*, 4, 189-208. 3.1

[37] 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.*) 1.2, 3.1

[38] Roberts J.W. (1979). Recent observations of subscriber behavior. *In Proceedings of the 9th International Tele-traffic Conference*. 3.1

[39] Shimkin N. and Mandelbaum A. (2004) Rational abandonment from tele-queues: non-linear waiting costs with heterogeneous preferences. *Queueing Systems: Theory and Applications (QUESTA)*, 47, 117-146. 1.1, 3.1

[40] U.S. Bureau of Labor Statistics. Table B-1: Employees on Nonfarm Payrolls by Major Industry, 1950 to Date. As reported on `www.bls.gov`. 1.1

[41] Ward A.R. and Glynn P.W. (2003) A diffusion approximation for a Markovian queue with reneging. *Queueing Systems: Theory and Applications (QUESTA)*, 43, 103-128. 3.2

[42] Ward A.R. and Glynn P.W. (2005) A diffusion approximation for a GI/GI/1 queue with balking or reneging. Draft. 3.2

[43] Whitt W. (2004) Fluid models for many-server queues with abandonments. *Operations Research*, to appear. 3.2, 6.2

[44] Whitt W. (2004) Sensitivity of performance in the Erlang A model to changes in the model parameters. Submitted to *Operations Research*. 3.2

[45] Whitt W. (2004) Staffing a call center with uncertain arrival rate and absenteeism. Submitted to *Management Science*. 3.2, 8

[46] Whitt W. (2005) Engineering Solution of a Basic Call-Center Model. *Management Science*, 51(2), 221-235. 3.2, 8

[47] Whitt W. (2005) Two fluid approximations for multi-server queues with abandonments. *Operations Research Letters*, 33, 363-372. 3.2

[48] 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`. (document), 2.1, 2.2, 4.9, 4.2, 4.2, 5.1, 6.2, 7, 9, 10.2

[49] Zohar E., Mandelbaum A. and Shimkin N. (2002) Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science*, 48, 566-583. 1.1, 3.1, 7