Internet Supplement to 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

May 23, 2005

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

Contents

Abstract of the main paper

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 Summary of the Internet Supplement

The goal of this supplement is to elaborate on material presented in the main paper. To facilitate the reading, statements of results from the main paper are repeated here. (Note that some results are in fact expanded in the Supplement.) Table 1 displays the correspondence between results; for example, Theorem 4.1 of the main paper is Theorem 6.1 in the Supplement.

Table 1: **Relation between the statements from the main paper and the Internet Supplement**

Main Paper	Internet Supplement
Theorem 4.1	Theorem 6.1
Theorem 4.2	Theorem 6.6
Theorem 4.3	Theorem 6.7
Theorem 5.1	Theorem 7.1
Theorem 6.1	Theorem 8.1
Lemma 10.1	Lemma 4.1
Lemma 11.1	Lemma 6.1

In Section 2 we briefly describe the results of Baccelli and Hebuterne [2] that are used in the following proofs. Sections 3 and 4 contain proofs of the results from Sections 9 and 10 of the main paper, 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: the QED, Quality-Driven (QD) and Efficiency-Driven (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.) 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.

2 Summary of Baccelli-Hebuterne's results on the M/M/n+G queue

The analysis in [2] is based on a Markov process $\{(N(t), \eta(t)),\ t \geq 0\}$, where $N(t)$ is the number of busy agents and $\eta(t)$ is the virtual offered waiting time (the offered wait of a virtual customer

that arrives at time t). Then the steady-state characteristics are defined by:

$$\begin{cases} v(x) & \triangleq \lim_{t\to\infty}\lim_{\epsilon\to 0}\dfrac{\mathrm{P}\{N(t)=n,\ x<\eta(t)\le x+\epsilon\}}{\epsilon}, & x\ge 0 \\ \pi_j & \triangleq \lim_{t\to\infty}\mathrm{P}\{N(t)=j,\ \eta(t)=0\}, & 0\le j\le n-1 \end{cases} \tag{2.1}$$

Here $v(x)$ is the density of the virtual offered waiting time. The unique solution of the steady-state equations is given by

$$\pi_j \ = \ \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\,\pi_0\,, \quad 0\le j\le n-1 \tag{2.2}$$

$$v(x) \ = \ \lambda\,\pi_{n-1}\exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}\,, \tag{2.3}$$

where

$$\pi_0 \ = \ \left[1+\frac{\lambda}{\mu}+\cdots+\left(\frac{\lambda}{\mu}\right)^{n-1}\frac{1}{(n-1)!}(1+\lambda J)\right]^{-1}\,, \tag{2.4}$$

$$J \ \triangleq \ \int_0^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}dx\,. \tag{2.5}$$

Moreover, probability to abandon can be calculated by

$$\mathrm{P}\{\mathrm{Ab}\} \ = \ \left(1-\frac{n\mu}{\lambda}\right)\left(1-\sum_{j=0}^{n-1}\pi_j\right)+\pi_{n-1}\,. \tag{2.6}$$

3 The M/M/n+G queue: summary of performance measures

Here we summarize *exact* formulae for M/M/n+G performance measures. Recall the definitions from the main paper.

M/M/n+G primitives.
The M/M/n+G model requires four input parameters:
λ – arrival rate,
μ – service rate,
n – number of agents,
G – patience distribution ($\bar{G}$ – survival function).

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

$$J \ \triangleq \ \int_0^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx\,, \tag{3.1}$$

$$J_1 \triangleq \int_0^\infty x \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{3.2}$$

$$J_H \triangleq \int_0^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,. \tag{3.3}$$

In addition, let

$$J(t) \triangleq \int_t^\infty \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{3.4}$$

$$J_1(t) \triangleq \int_t^\infty x \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,, \tag{3.5}$$

$$J_H(t) \triangleq \int_t^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,. \tag{3.6}$$

Finally, define

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}} = \int_0^\infty e^{-t}\left(1+\frac{t\mu}{\lambda}\right)^{n-1} dt\,. \tag{3.7}$$

List of performance measures:

Recall notation from the main paper:

P{Ab} – probability to abandon, P{Sr} – probability to be served,

Q – queue length, W – waiting time,

V – offered wait (time that a customer with infinite patience would wait).

Then

$$\mathrm{P}\{V>0\} = \frac{\lambda J}{\mathcal{E}+\lambda J}\,, \tag{3.8}$$

$$\mathrm{P}\{W>0\} = \frac{\lambda J}{\mathcal{E}+\lambda J}\cdot \bar{G}(0)\,, \tag{3.9}$$

$$\mathrm{P}\{\mathrm{Ab}\} = \frac{1+(\lambda-n\mu)J}{\mathcal{E}+\lambda J}\,, \tag{3.10}$$

$$\mathrm{P}\{\mathrm{Ab} \mid \mathrm{V}>0\} = \frac{1+(\lambda-n\mu)J}{\lambda J}\,, \tag{3.11}$$

$$\mathrm{P}\{\mathrm{Sr}\} = \frac{\mathcal{E}+n\mu J-1}{\mathcal{E}+\lambda J}\,, \tag{3.12}$$

$$\mathrm{E}[V] = \frac{\lambda J_1}{\mathcal{E}+\lambda J}\,, \tag{3.13}$$

$$\mathrm{E}[V \mid V>0] = \frac{J_1}{J}\,, \tag{3.14}$$

$$\mathrm{E}[W] = \frac{\lambda J_H}{\mathcal{E}+\lambda J}\,, \tag{3.15}$$

$$\mathrm{E}[Q] = \frac{\lambda^2 J_H}{\mathcal{E}+\lambda J}\,, \tag{3.16}$$

$$\mathrm{E}[V \mid \mathrm{Ab}] = \frac{(\lambda-n\mu)J_1+J}{(\lambda-n\mu)J+1}\,, \tag{3.17}$$

$$\mathrm{E}[W \mid \mathrm{Ab}] = \frac{J + \lambda J_H - n\mu J_1}{(\lambda - n\mu)J + 1}, \tag{3.18}$$

$$\mathrm{E}[V \mid \mathrm{Sr}] = \mathrm{E}[W \mid \mathrm{Sr}] = \frac{n\mu J_1 - J}{\mathcal{E} + n\mu J - 1}, \tag{3.19}$$

$$\mathrm{P}\{V > t\} = \frac{\lambda J(t)}{\mathcal{E} + \lambda J}, \tag{3.20}$$

$$\mathrm{P}\{W > t\} = \frac{\lambda \bar{G}(t) J(t)}{\mathcal{E} + \lambda J}, \tag{3.21}$$

$$\mathrm{E}[V \mid V > t] = \frac{J_1(t)}{J(t)}, \tag{3.22}$$

$$\mathrm{E}[W \mid W > t] = \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t) J(t)}, \tag{3.23}$$

$$\mathrm{P}\{\mathrm{Ab} \mid V > t\} = \frac{\lambda - n\mu}{\lambda} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda J(t)}, \tag{3.24}$$

$$\mathrm{P}\{\mathrm{Ab} \mid W > t\} = \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)}. \tag{3.25}$$

3.1 Proofs of (3.8)-(3.25)

Here we present the proofs of (3.8)-(3.25), one by one.

(3.8). First, (2.2)-(2.5) and definition (3.7) imply the useful formula

$$\pi_{n-1} = \frac{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}}{\sum_{j=0}^{n-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^{j} + \frac{\lambda J}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-1}} = \frac{1}{\mathcal{E} + \lambda J}. \tag{3.26}$$

Then use that

$$\mathrm{P}\{V > 0\} = 1 - \sum_{j=0}^{n-1} \pi_j = \int_0^\infty v(x)dx = \lambda \pi_{n-1} J, \tag{3.27}$$

where the last equality of (3.27) follows from (2.3) and (2.5).

(3.9). Follows from

$$\mathrm{P}\{W > 0 | V > 0\} = \bar{G}(0).$$

(3.10). Formula (2.6) implies that

$$\mathrm{P}\{\mathrm{Ab}\} = \left(1 - \frac{n\mu}{\lambda}\right) \cdot \mathrm{P}\{V > 0\} + \frac{1}{\mathcal{E} + \lambda J}.$$

Now substitute (3.8).

(3.11). Immediate consequence of (3.8) and (3.10).

(3.12).

$$\mathrm{P}\{\mathrm{Sr}\} = 1 - \mathrm{P}\{\mathrm{Ab}\}.$$

(3.13). Results from (3.26) and

$$\mathrm{E}[V] \;=\; \lambda\pi_{n-1}\cdot\int_0^\infty x\exp\{\lambda H(x)-n\mu x\}dx\,.$$

(3.14). Formulae (3.8) and (3.13).

(3.15). According to formula (2.3), the survival function of the virtual wait is given by

$$\mathrm{P}\{V>t\} \;\stackrel{\Delta}{=}\; \bar V(t) \;=\; \lambda\pi_{n-1}\int_t^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx\,.$$

Hence, the average wait is equal to

$$\mathrm{E}[W] \;=\; \int_0^\infty \bar G(t)\bar V(t)dt \;=\; \lambda\pi_{n-1}\int_0^\infty \bar G(t)\cdot\int_t^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dxdt$$

and integrating by parts

$$\mathrm{E}[W] \;=\; \lambda\pi_{n-1}\int_0^\infty H(t)\cdot\exp\left\{\lambda H(t)-n\mu t\right\}dt\,.$$

Then use formula (3.26) and definition (3.3).

(3.16). Follows from (3.15) and Little's formula.

(3.17).

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{\mathrm{E}[V\cdot 1_{\{\tau\le V\}}]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty xv(x)G(x)dx}{\mathrm{P}\{\mathrm{Ab}\}}\,, \tag{3.28}$$

where from (2.3) and (3.26)

$$v(x) \;=\; \frac{\lambda\exp\{\lambda H(x)-n\mu x\}}{\mathcal{E}+\lambda J}\,. \tag{3.29}$$

Integration by parts implies that

$$\int_0^\infty x[\lambda\bar G(x)-n\mu]\exp\{\lambda H(x)-n\mu x\}dx$$

$$=\; \int_0^\infty xd\left[\exp\{\lambda H(x)-n\mu x\}\right] \;=\; -\int_0^\infty \exp\{\lambda H(x)-n\mu x\}dx \;=\; -J\,,$$

and

$$\int_0^\infty xG(x)\exp\{\lambda H(x)-n\mu x\}dx \;=\; \frac{(\lambda-n\mu)J_1+J}{\lambda}\,, \tag{3.30}$$

which, combined with (3.28), (3.29) and (3.10) implies

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{(\lambda-n\mu)J_1+J}{(\lambda-n\mu)J+1}\,.$$

(3.18). Similar to the previous calculation

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \frac{\mathrm{E}[\tau\cdot 1_{\{\tau\le V\}}]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty x\bar V(x)dG(x)}{\mathrm{P}\{\mathrm{Ab}\}}\,.$$

Note that

$$d[xG(x) + H(x) - x] = xdG(x)\,. \tag{3.31}$$

Then

$$\frac{\int_0^\infty x\bar{V}(x)dG(x)}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty v(x)\cdot[xG(x)+H(x)-x]dx}{\mathrm{P}\{\mathrm{Ab}\}}$$

(use (3.29) and (3.10))

$$=\; \frac{\lambda\int_0^\infty[xG(x)+H(x)-x]\cdot\exp\{\lambda H(x)-n\mu x\}dx}{1+(\lambda-n\mu)J} \;=\; \frac{J+\lambda J_H - n\mu J_1}{1+(\lambda-n\mu)J}\,,$$

where the last equality follows from (3.30) and the definitions of J_1 and J_H.

(3.19). This formula for $\mathrm{E}[V|\mathrm{Sr}]$ can be checked via

$$\mathrm{E}[V] = \mathrm{E}[V|\mathrm{Sr}]\cdot\mathrm{P}\{\mathrm{Sr}\} + \mathrm{E}[V|\mathrm{Ab}]\cdot\mathrm{P}\{\mathrm{Ab}\}\,.$$

Since the event {Sr} is equivalent to {W=V},

$$\mathrm{E}[V|\mathrm{Sr}] \;=\; \mathrm{E}[W|\mathrm{Sr}]\,.$$

(3.20). Follows from (3.29).

(3.21). Consequence of

$$\mathrm{P}\{W>t\} \;=\; \mathrm{P}\{V>t\}\cdot\mathrm{P}\{\tau>t\}\,.$$

(3.22). Follows from definitions of $J(t)$ and $J_1(t)$.

(3.23).

$$\mathrm{E}[W|W>t] \;=\; \frac{\int_t^\infty xw(x)dx}{\mathrm{P}\{W>t\}}\,, \tag{3.32}$$

where $w(x)$ is the waiting-time density. The denominator of (3.32) is equal to

$$\mathrm{P}\{W>t\} \;=\; \bar{G}(t)\bar{V}(t) \;=\; \lambda\pi_{n-1}\bar{G}(t)\int_t^\infty \exp\{\lambda H(x)-n\mu x\}dx\,. \tag{3.33}$$

Calculating the numerator of (3.32):

$$\begin{aligned}\int_t^\infty xw(x)dx \;&=\; \int_t^\infty xv(x)\bar{G}(x)dx + \int_t^\infty x\bar{V}(x)dG(x)\\ &=\; \lambda\pi_{n-1}\left[\int_t^\infty x\bar{G}(x)\exp\{\lambda H(x)-n\mu x\}dx\;+\right.\\ &\left.\quad+\int_t^\infty x\left(\int_x^\infty \exp\{\lambda H(u)-n\mu u\}du\right)dG(x)\right]\end{aligned} \tag{3.34}$$

Use (3.31) to show that the double integral in (3.34) is equal to

$$\int_t^\infty [xG(x)+H(x)-x]\cdot\exp\{\lambda H(x)-n\mu x\}dx\;-$$

$$- \ [tG(t) + H(t) - t] \cdot \int_t^\infty \exp\{\lambda H(x) - n\mu x\} dx \,. \tag{3.35}$$

After some terms cancel, we get from (3.32), (3.33), (3.34) and (3.35) that

$$\begin{aligned} \mathrm{E}[W|W > t] \ &= \ \frac{\int_t^\infty [H(x) + t\bar{G}(t) - H(t)] \cdot \exp\{\lambda H(x) - n\mu x\} dx}{\bar{G}(t) \cdot \int_t^\infty \exp\{\lambda H(x) - n\mu x\} dx} \\ &= \ \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t) J(t)} \,. \end{aligned}$$

(3.24).

$$\begin{aligned} \mathrm{P}\{\mathrm{Ab}|V > t\} \ &= \ \frac{\mathrm{P}\{\tau \le V;\, V > t\}}{\mathrm{P}\{V > t\}} \ = \ \frac{\int_t^\infty G(x) v(x) dx}{\int_t^\infty v(x) dx} \\ &= \ \frac{\int_t^\infty G(x) \cdot \exp\{\lambda H(x) - n\mu x\} dx}{J(t)} \,. \end{aligned} \tag{3.36}$$

Using integration by parts

$$\int_t^\infty [\lambda \bar{G}(x) - n\mu] \cdot \exp\{\lambda H(x) - n\mu x\} dx \ = \ -\exp\{\lambda H(t) - n\mu t\} \,.$$

Hence,

$$\int_t^\infty G(x) \cdot \exp\{\lambda H(x) - n\mu x\} dx \ = \ \frac{(\lambda - n\mu) J(t) + \exp\{\lambda H(t) - n\mu t\}}{\lambda} \,. \tag{3.37}$$

Now (3.36) and (3.37) imply

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

(3.25).

$$\begin{aligned} \mathrm{P}\{\mathrm{Ab}|W > t\} \ &= \ \frac{\mathrm{P}\{\tau \le V;\, \tau > t\}}{\mathrm{P}\{\min(V, \tau) > t\}} \ = \ \frac{\int_t^\infty \bar{V}(x) dG(x)}{\bar{G}(t) \bar{V}(t)} \\ = \ \frac{G(x)\bar{V}(x)|_t^\infty + \int_t^\infty G(x) v(x) dx}{\bar{G}(t)\bar{V}(t)} \ &= \ \frac{\int_t^\infty G(x) \cdot \exp\{\lambda H(x) - n\mu x\} dx - G(t) J(t)}{\bar{G}(t) J(t)} \\ &= \ \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)} \,. \end{aligned}$$

4 Asymptotic behavior of integrals

Here we prove Lemma 4.1 (Lemma 10.1 from the main paper) and two additional lemmata that will be needed in the following proofs.

4.1 Asymptotic results

Lemma 4.1 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

$$\frac{k_1}{l_1} \; > \; \frac{k_2}{l_2}\,. \tag{4.1}$$

Then

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

$$= \; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} \; + \; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right), \qquad \lambda \to \infty\,. \tag{4.2}$$

and

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

$$= \; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 [b_1 r_1(\lambda)]^{\frac{m+1}{l_1}}} \; - \; \frac{b_2 \Gamma\left(\frac{m+l_2+1}{l_1}\right)}{l_1 b_1^{\frac{m+l_2+1}{l_1}}} \cdot \lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}} \; + \; o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right). \tag{4.3}$$

Lemma 4.2 In addition to assumptions of Lemma 4.1, let $k_1 > k_2$ and assume that the function $r_2 = \{r_2(\lambda), \quad \lambda > 0\}$ satisfies $r_2(\lambda) = o(\lambda^{k_2}), \quad \lambda \to \infty$. Then

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} dx$$

$$= \; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} \; + \; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right). \tag{4.4}$$

Remark 4.1 Note that $r_2(\lambda)$ does not need to be positive, which is in contrast to the corresponding term λ^{k_2} in Lemma 4.1.

Remark 4.2 We can generalize (4.2) to

$$\int_0^\infty x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - \sum_{i=2}^n b_i \lambda^{k_i} x^{l_i}\right\} dx \; = \; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} + o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right), \qquad \lambda \to \infty\,,$$

as long as $\dfrac{k_1}{l_1} \; > \; \dfrac{k_i}{l_i}$ prevails for $2 \le i \le n$.

Lemma 4.3 Let $b, k, l, \delta > 0$, integer $m \geq 0$, and $-\infty < n < \infty$. Assume that the function $r(\lambda) \sim \lambda^k, \quad \lambda \to \infty$. Define a function

$$S(\lambda) \triangleq \int_{\delta\lambda^n}^{\infty} x^m \cdot \exp\left\{-br(\lambda)x^l\right\} dx\,, \qquad \lambda > 0\,,$$

and assume

$$nl + k > 0\,. \tag{4.5}$$

Then there exists $\nu > 0$ such that

$$S(\lambda) \;=\; o(e^{-\lambda^\nu})\,. \tag{4.6}$$

4.2 Proofs of Lemmata 4.1-4.3

Proof of Lemma 4.1.

Define

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

and

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

Formula (10.2) from the main paper and straightforward calculations imply

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

Now

$$|I - I_A| \;=\; o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right) \tag{4.7}$$

will imply Lemma 4.1. If $x > 0$ and $\lambda^{k_2}x^{l_2} \leq 1$, then there exists $C > 0$ such that

$$|\exp\{-b_2\lambda^{k_2}x^{l_2}\} - (1 - b_2\lambda^{k_2}x^{l_2})| \leq C\lambda^{2k_2}x^{2l_2}\,.$$

Define $\delta \triangleq \lambda^{-k_2/l_2}$ and note that the condition $\lambda^{k_2}x^{l_2} \leq 1$ is equivalent to $x \leq \delta$. Now

$$\int_0^{\delta} x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1}\right\} \cdot |\exp\{-b_2\lambda^{k_2}x^{l_2}\} - (1 - b_2\lambda^{k_2}x^{l_2})|dx$$

$$\leq\; C \cdot \int_0^{\infty} \lambda^{2k_2}x^{m+2l_2} \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1}\right\} dx$$

$$= \frac{C\lambda^{2k_2}}{l_1[b_1 r_1(\lambda)]^{\frac{m+2l_2+1}{l_1}}} \cdot \Gamma\left(\frac{m+2l_2+1}{l_1}\right) \;=\; O\left(\lambda^{2k_2 - \frac{k_1(m+2l_2+1)}{l_1}}\right) \;=\; o\left(\lambda^{k_2 - \frac{k_1(m+l_2+1)}{l_1}}\right)\,,$$

where the last equality follows from (4.1). In order to complete the proof, we show that the remainder $\int_{\delta}^{\infty}$ of the integrals can be ignored. Specifically, there exists $\nu > 0$ such that

$$\int_{\delta}^{\infty} x^m \cdot \exp\left\{-b_1 r_1(\lambda)x^{l_1} - b_2\lambda^{k_2}x^{l_2}\right\} dx \;=\; o\left(e^{-\lambda^\nu}\right)$$

and

$$\int_{\delta}^{\infty} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 \lambda^{k_2} x^{l_2}] dx \;=\; o\left(e^{-\lambda^{\nu}}\right) .$$

The last two statements follow from Lemma 4.3. (Condition (4.5) applies due to (4.1).) ■

Proof of Lemma 4.2.

The proof is similar to the proof of Lemma 4.1. The integration domain is again divided by $\delta = \lambda^{-k_2/l_2}$. For large λ the inequality $x \le \delta$ implies $|x^{l_2} r_2(\lambda)| \le 1$, which, in turn, implies

$$|\exp\{-b_2 r_2(\lambda) x^{l_2}\} - (1 - b_2 r_2(\lambda) x^{l_2})| \;\le\; C[r_2(\lambda)]^2 x^{2l_2}$$

for some $C > 0$. Then one shows that

$$\left| \int_0^{\delta} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} dx - \int_0^{\delta} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \right|$$

$$= \; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right) ,$$

and

$$\int_0^{\infty} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \;=\; \frac{\Gamma\left(\frac{m+1}{l_1}\right)}{l_1 b_1^{\frac{m+1}{l_1}}} \cdot \lambda^{-\frac{k_1(m+1)}{l_1}} \;+\; o\left(\lambda^{-\frac{k_1(m+1)}{l_1}}\right) .$$

The last step is to prove "exponential bounds":

$$\int_{\delta}^{\infty} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} dx \;=\; o\left(e^{-\lambda^{\nu}}\right), \quad \nu > 0 , \tag{4.8}$$

and

$$\int_{\delta}^{\infty} x^m \cdot \exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] dx \;=\; o\left(e^{-\lambda^{\nu}}\right), \quad \nu > 0 .$$

In order to get (4.8), the condition $k_1 > k_2$ is needed. It enables us to find $0 < C_1 < 1$, such that for $x > \delta$ and λ large enough,

$$\exp\left\{-b_1 r_1(\lambda) x^{l_1} - b_2 r_2(\lambda) x^{l_2}\right\} \;<\; \exp\left\{-b_1 C_1 r_1(\lambda) x^{l_1}\right\} ,$$

and

$$\exp\left\{-b_1 r_1(\lambda) x^{l_1}\right\} \cdot [1 - b_2 r_2(\lambda) x^{l_2}] \;<\; \exp\left\{-b_1 C_1 r_1(\lambda) x^{l_1}\right\} ,$$

Now we can apply Lemma 4.3. (Its proof appears below.) ■

Proof of Lemma 4.3.

We perform a change of variables

$$z = br(\lambda) x^l, \quad x = \left(\frac{z}{br(\lambda)}\right)^{1/l}, \quad dx = \frac{dz}{br(\lambda)} \left(\frac{z}{br(\lambda)}\right)^{1/l-1} ,$$

getting

$$S(\lambda) \;=\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \int_{C_1 r(\lambda)\lambda^{nl}}^{\infty} e^{-z} z^{\frac{m+1}{l}-1} dz\,, \tag{4.9}$$

where C_1 and C_2 are positive constants. Under condition (4.5), the lower bound $C_1 r(\lambda)\lambda^{nl}$ of the integral in (4.9) converges to infinity. Therefore, there exists $\alpha > 0$ such that for λ large enough,

$$S(\lambda) \;\leq\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \int_{C_1 r(\lambda)\lambda^{nl}}^{\infty} e^{-\alpha z} dz \;=\; \frac{C_2}{r(\lambda)^{\frac{m+1}{l}}} \cdot \exp\{-C_3 r(\lambda)\lambda^{nl}\}\,,$$

where C_3 is a positive constant. Since $r(\lambda)\lambda^{nl} \sim \lambda^{nl+k}$, we can easily find $\nu > 0$ such that (4.6) is satisfied. ■

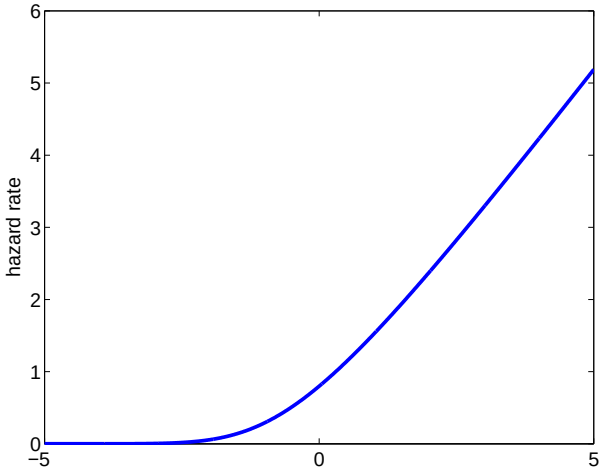
5 Some properties of the normal hazard-rate

In the sequel, we use some properties of the *hazard rate* function of the standard normal distribution:

$$h(x) \;=\; \frac{\phi(x)}{1-\Phi(x)} \;=\; \frac{\phi(x)}{\bar{\Phi}(x)}\,, \tag{5.1}$$

where $\Phi(x)$ is its cumulative distribution function, $\bar{\Phi}(x) = 1 - \Phi(x)$ is the survival function and $\phi(x) = \Phi^{'}(x)$ is the density.

Figure 1: **Normal hazard rate**



The derivative of the normal hazard rate is equal to

$$h^{'}(x) = h(x) \cdot (h(x) - x) \tag{5.2}$$

and, consequently,

$$h(x) - x = [\ln h(x)]^{'}\,.$$

The second derivative is

$$h''(x) = h(x) \cdot (2h^2(x) - 3xh(x) + x^2 - 1)\,. \tag{5.3}$$

Theorem 1.3 from Durrett [4] states that

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\phi(x) \leq \bar{\Phi}(x) \leq \frac{1}{x}\phi(x)\,, \qquad \text{for } x > 0.$$

Then it follows that:

$$h(x) \geq x, \qquad -\infty \leq x \leq \infty\,,$$

$$h(x) \leq \frac{x^3}{x^2 - 1}, \qquad x > 1,$$

and

$$|h(x) - x| \to 0\,, \qquad \text{as } x \to \infty\,. \tag{5.4}$$

It is well-known that h is an increasing function (see Gupta and Gupta [7] for a general treatment of multivariate normal case). Surprisingly, we have not found anywhere a proof that h is convex and we shall need this fact. So we constructed an indirect proof, based on the convexity of the Erlang-B formula [12], in the following way. Define the function

$$B(s,a) \;\stackrel{\Delta}{=}\; \left[a\int_0^\infty e^{-at}(1+t)^s dt\right]^{-1}.$$

For $a > 0$ and integer $s > 0$ it can be shown that

$$B(s,a) \;=\; \left[\sum_{i=0}^{s}\frac{a^i}{i!}\right]^{-1} \cdot \frac{a^s}{s!}\,.$$

The last expression is equal to the Erlang-B blocking probability in the M/M/s/s system with $a = \dfrac{\lambda}{\mu}$. It has been proved in [12] that $B(s,a)$ is convex in s in $[0,\infty)$, for all $a > 0$. Now define

$$\tilde{B}(\beta,a) \;\stackrel{\Delta}{=}\; \sqrt{a}\cdot B(a+\beta\sqrt{a},a)\,.$$

Obviously $\tilde{B}(\beta,a)$ is also convex in β over $(-\sqrt{a},\infty)$. The QED result for the Erlang-B system, derived by Jagerman [11], implies that

$$\tilde{B}(\beta,a) \to h(-\beta) \qquad (a \to \infty)\,.$$

The pointwise limit of a sequence of convex functions is convex as well, implying that h is convex.

Finally, formula (5.4) and the convexity of h imply

$$h'(x) < 1\,, \qquad -\infty < x < \infty\,. \tag{5.5}$$

6 QED operational regime

6.1 Formulation of results

6.1.1 Main case: patience distribution with a positive density at the origin

The first lemma continues Lemma 11.1 from the main paper.

Lemma 6.1 (Building blocks) Under the assumptions of Theorem 4.1 from the main paper, the building blocks J, $\mathcal{E}$ and J_1, defined in Section 3, are approximated by:

a.

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.1}$$

b.

$$J_1 \;=\; \frac{1}{n} \cdot \frac{1}{\mu g_0} \left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right). \tag{6.2}$$

c.

$$\mathcal{E} \;=\; \sqrt{n} \cdot \frac{1}{h(-\beta)} + o(\sqrt{n}). \tag{6.3}$$

d. Define

$$J_2 \;\triangleq\; \int_0^\infty x^2 \cdot \exp\left\{\lambda \int_0^x G(u)du - n\mu x\right\} dx. \tag{6.4}$$

Then

$$J_2 \;=\; \frac{1}{n^{3/2}} \cdot \frac{1}{(\mu g_0)^{3/2}} \left[\frac{\hat{\beta}^2 + 1}{h(\hat{\beta})} - \hat{\beta}\right] + o\left(\frac{1}{n^{3/2}}\right). \tag{6.5}$$

Theorem 6.1 (Performance measures) Under the assumptions of Theorem 4.1 from the main paper, the performance measures of the M/M/n+G queueing system in the QED regime are approximated by:

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

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{6.6}$$

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

$$n \;=\; \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \tag{6.7}$$

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 $\frac{1}{\sqrt{n}}$:

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.8}$$

The probability to abandon P{Ab} also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (6.6) with (6.8).

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

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.9}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (6.6) and (6.9).

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

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W>0] \;\sim\; \mathrm{E}[V \mid V>0]\,. \tag{6.10}$$

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

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;=\; \frac{\mathrm{P}\{\mathrm{Ab}|W>0\}}{\mathrm{E}[W|W>0]} \;\sim\; g_0\,. \tag{6.11}$$

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

$$\mathrm{E}[V|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.12}$$

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{2\sqrt{g_0\mu}} \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right), \tag{6.13}$$

or, in other words,

$$\mathrm{E}[W|\mathrm{Ab}] \;\sim\; \frac{1}{2} \cdot \mathrm{E}[V|\mathrm{Ab}]\,. \tag{6.14}$$

Moreover, the following inequality prevails:

$$\frac{1}{2} \cdot \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right] \;<\; h(\hat{\beta}) - \hat{\beta} \;<\; \frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\,, \qquad -\infty < \hat{\beta} < \infty. \tag{6.15}$$

(See also Remark 4.7 from the main paper.)

g. The distribution of wait, given delay in queue, is asymptotically equal to

$$\mathrm{P}\left\{\frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \,\middle|\, W>0\right\} \;\sim\; \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{\frac{g_0}{\mu}} \cdot t\right)}{\bar{\Phi}(\hat{\beta})}\,, \qquad t \geq 0\,. \tag{6.16}$$

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

$$\mathrm{P}\left\{\mathrm{Ab}\;\middle|\;\frac{W}{\mathrm{E}[S]}>\frac{t}{\sqrt{n}}\right\}\;=\;\frac{1}{\sqrt{n}}\cdot\sqrt{\frac{g_0}{\mu}}\cdot\left[h\left(\hat{\beta}+t\sqrt{\frac{g_0}{\mu}}\right)-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right). \tag{6.17}$$

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

$$\mathrm{E}\left[W\;\middle|\;\frac{W}{\mathrm{E}[S]}>\frac{t}{\sqrt{n}}\right]\;=\;\frac{1}{\sqrt{n}}\cdot\sqrt{\frac{1}{g_0\mu}}\cdot\left[h\left(\hat{\beta}+t\sqrt{\frac{g_0}{\mu}}\right)-\hat{\beta}\right]+o\left(\frac{1}{\sqrt{n}}\right). \tag{6.18}$$

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

$$\frac{\mathrm{P}\left\{\mathrm{Ab}\;|\,W>t/\sqrt{n}\right\}}{\mathrm{E}\left[W\;|\,W>t/\sqrt{n}\right]}\;\sim\;g_0\,,\qquad t\geq 0. \tag{6.19}$$

6.1.2 Patience distribution with density vanishing near the origin

We would like to cover models where customers are going through several stages of (im)patience before reneging. (See, for example, Ishay [10] or Baccelli and Hebuterne [2]; the latter fit an Erlang distribution with 3 phases to patience, using real data.) In such models, we cannot expect significant abandonment near the origin, which suggests patience distributions with density vanishing near the origin.

Lemma 6.2 (Building blocks) Assume that the density of patience time at the origin $g_0 = 0$; that the first $(k-1)$ derivatives vanish as well: $g^{(i)}(0) = 0,\ 1 \leq i \leq k-1$, and that the k-th derivative is positive: $g^{(k)}(0) \stackrel{\Delta}{=} g_{0k} > 0$.

For $\beta \neq 0$ (positive or negative) let the QED staffing level be

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda})\,. \tag{6.20}$$

If $\beta = 0$ let

$$n = \frac{\lambda}{\mu} + o\left(\lambda^s\right)\,, \tag{6.21}$$

for some $s < \dfrac{1}{k+2}$.

The asymptotic expression for $\mathcal{E}$ coincides with formula (11.3) from the main paper for all the theorems of Section 6. The approximations for J and J_1 are given by the following formulae:

a. If $\beta > 0$

$$J\;=\;\frac{1}{n\mu-\lambda}-\frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+3}}+o\left(\frac{1}{\lambda^{(k+1)/2}}\right), \tag{6.22}$$

$$J_1\;=\;\frac{1}{(n\mu-\lambda)^2}-\frac{(k+3)\cdot\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+4}}+o\left(\frac{1}{\lambda^{(k+2)/2}}\right). \tag{6.23}$$

b. If $\beta = 0$

$$J = \frac{1}{k+2} \cdot \left[\frac{(k+2)!}{\lambda g_{0k}}\right]^{1/(k+2)} \cdot \Gamma\left(\frac{1}{k+2}\right) + o\left(\frac{1}{\lambda^{1/(k+2)}}\right), \tag{6.24}$$

$$J_1 = \frac{1}{k+2} \cdot \left[\frac{(k+2)!}{\lambda g_{0k}}\right]^{2/(k+2)} \cdot \Gamma\left(\frac{2}{k+2}\right) + o\left(\frac{1}{\lambda^{2/(k+2)}}\right). \tag{6.25}$$

c. If $\beta < 0$

$$J \sim \exp\left\{\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\}$$
$$\cdot \sqrt{2\pi k!} \cdot (\lambda g_{0k})^{-1/(2k+2)} \cdot ((k+1)!(\lambda - n\mu))^{-k/(2k+2)}, \tag{6.26}$$

$$J_1 \sim \left(\frac{-\beta\sqrt{\mu}(k+1)!}{g_{0k}\sqrt{\lambda}}\right)^{1/(k+1)} \cdot J. \tag{6.27}$$

Remark 6.1 Expression (6.26) increases exponentially due to the $(\lambda - n\mu)^{(k+2)/(k+1)}$ term in the exponent.

Theorem 6.2 (Performance measures) Under the assumptions of Lemma 6.2, the performance measures of M/M/n+G are approximated by:

a. Delay probability.

If $\beta > 0$, the delay probability coincides (asymptotically) with the Erlang-C approximation from Halfin and Whitt [9]:

$$\mathrm{P}\{\mathrm{W} > 0\} \sim \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}. \tag{6.28}$$

If $\beta = 0$, the probability to get service immediately converges to zero at rate $\frac{1}{n^{k/(2k+4)}}$:

$$\mathrm{P}\{\mathrm{W} = 0\} = \frac{1}{n^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{k/(2k+4)}}\right). \tag{6.29}$$

If $\beta < 0$, the probability to get service immediately decreases to zero at an exponential rate:

$$\mathrm{P}\{\mathrm{W} = 0\} \approx \exp\left\{-\frac{k+1}{k+2} \cdot \left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)} \cdot (\lambda - n\mu)^{(k+2)/(k+1)}\right\}$$
$$\cdot \frac{g_{0k}^{1/(2k+2)} \cdot (-\beta(k+1)!)^{k/(2k+2)}}{\lambda^{k/(4k+4)} \cdot \mu^{(k+2)/(4k+4)} \cdot \sqrt{2\pi k!} \cdot h(-\beta)}. \tag{6.30}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} = \frac{1}{n^{(k+1)/2}} \cdot \frac{g_{0k}}{(\beta\mu)^{k+1}} + o\left(\frac{1}{n^{(k+1)/2}}\right). \tag{6.31}$$

If $\beta = 0$

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{1}{n^{(k+1)/(k+2)}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{(k+1)/(k+2)}}\right). \tag{6.32}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab}|V>0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.33}$$

c. Average offered waiting time.

If $\beta > 0$, the average offered wait is given by the Erlang-C approximation [9]:

$$\mathrm{E}[V \mid V>0] \;\sim\; \frac{1}{\beta\mu\sqrt{n}}. \tag{6.34}$$

If $\beta = 0$

$$\mathrm{E}[V \mid V>0] \;=\; \frac{1}{n^{1/(k+2)}} \cdot \frac{\Gamma\left(\frac{2}{k+2}\right)}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{(k+2)!}{\mu g_{0k}}\right]^{\frac{1}{k+2}} + o\left(\frac{1}{n^{1/(k+2)}}\right). \tag{6.35}$$

If $\beta < 0$

$$\mathrm{E}[V \mid V>0] \;=\; \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)} + o\left(\frac{1}{n^{1/(2k+2)}}\right). \tag{6.36}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W>0] \;\sim\; \mathrm{E}[V \mid V>0]. \tag{6.37}$$

Remark 6.2 The value $-\beta/\sqrt{n}$ in formula (6.33) is the minimal reneging rate that is required to avoid queue explosion. Indeed, one can check that $-\beta/\sqrt{n}$ is asymptotically equivalent to the "fluid limit" of the probability-to-abandon [6] $1 - 1/\rho$, given $n \to \infty$.

Remark 6.3 We do not study the case when all derivatives at the origin are zero but the density is positive near the origin. We think that the answers here would depend on the specific distribution (e.g. lognormal). In general, the case above is intermediate between those described in Theorems 6.2 and 6.4.

Example. Phase-type patience times. An important special case of distributions, relevant to Theorem 6.2, is phase-type (see Asmussen [1] or Ishay [10]). Here we study the behavior of the phase-type density near the origin, which is essential if one is to apply Theorem 6.2.

Definition. Consider a continuous-time Markov process $\{X = X_t,\ t \geq 0\}$ with a finite state-space $\{1, 2, \ldots, k, \Delta\}$, where $1, 2, \ldots, k$ are transient states and Δ is the absorbing state. The distribution of X is characterized by:

- *Initial distribution* $\bar{q} = (q_1, \ldots, q_k)$, where $q_i = \mathrm{P}\{X_0 = i\},\ 1 \le i \le k$ (the process cannot start from the absorbing state).
- *Phase-type generator* R, a $k \times k$ matrix of transition rates between the transient states. We know that $R_{ii} < 0$, $R_{ij} \ge 0$ for $i \ne j$, and $\sum_{j=1}^{k} R_{ij} \le 0$, where $1 \le i, j \le k$.
- *Absorption intensities* $\bar{r} = (r_1, \ldots, r_k)'$. Overall, the generator of X can be written as

$$Q \;=\; \begin{pmatrix} R & \bar{r} \\ 0, \ldots, 0 & 0 \end{pmatrix},$$

  where every row in Q sums up to zero: $\bar{r} = -R \cdot \bar{1}$. (Here $\bar{1}$ is the vector with all components equal to 1.)

Let

$$T \;\stackrel{\Delta}{=}\; \inf\{t > 0 :\ X_t = \Delta\}$$

denote the absorption time. Then $F_T(t) = \mathrm{P}_{\bar{q}}\{T \le t\}$ is a *phase-type distribution* with parameters $(\bar{q},\, R)$.

The cumulative distribution function of the phase-type distribution with parameters $(\bar{q}, R)$ is,

$$F_T(t) = 1 - \bar{q} \exp\{Rt\} \bar{1}\,,$$

and it has a density

$$f_T(t) = \bar{q} \exp\{Rt\} \bar{r}\,. \tag{6.38}$$

In order to apply Theorem 6.2, we must calculate the density at the origin and its derivatives. From (6.38), the density at the origin is

$$f_T(0) \;=\; \bar{q}\bar{r}$$

and its n-th derivative (for convenience, we denote also $f_T^{(0)}(t) \stackrel{\Delta}{=} f_T(t)$)

$$f_T^{(n)}(0) \;=\; \bar{q} R^n \bar{r}\,, \qquad n \ge 0\,. \tag{6.39}$$

Theorem 6.3 (Phase-Type patience) Represent the transient states of the underlying Markov process of a phase-type distribution by a directed graph. Two states j and k are connected if and only if $R_{jk} > 0$. For any initial state j $(q_j > 0)$, let L_j denote the number of states in a minimal path that connects j with the absorbing state Δ. Define

$$L \;\stackrel{\Delta}{=}\; \min_{j:q_j>0} L_j\,. \tag{6.40}$$

(For example, $L = n$ for the Erlang distribution with n phases and $L = 1$ for the hyperexponential distribution.)

Then $f_T^{(L-1)}(0) > 0$. Moreover, if $L \ge 2$, then $f_T^{(i)}(0) = 0$ for $0 \le i \le L - 2$. ■

Now Theorem 6.3 and formula (6.39) enable us to apply Theorem 6.2 to phase-type distributions.

6.1.3 Delayed distribution of patience

Assume that, up to a fixed time $c > 0$, customers do not abandon. For example, customers could be listening to a recorded announcement. Such situations inspire us to consider delayed distributions of patience, which can be represented by $c + \tau$, where τ represents (im)patience as before. The case of *deterministic patience* is important as well. As examples, one can consider overflowing[1], or Internet applications, where the waiting of jobs in queue is usually bounded.

Lemma 6.3 (Building blocks) Assume that the density of patience time vanishes over the interval $[0, c]$, for some $c > 0$. (That means that all customers are willing to wait at least c.) Assume that the density of patience time is positive at c: $g_c > 0$. For $\beta \neq 0$ (both negative and positive) consider the staffing level

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

For $\beta = 0$ let

$$n = \frac{\lambda}{\mu} + a\,, \qquad -\infty < a < \infty\,. \tag{6.41}$$

a. If $\beta > 0$

$$J \;=\; \frac{1}{n\mu - \lambda} - \frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{\beta\sqrt{\mu}} - \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} \;+\; o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}} \right), \tag{6.42}$$

$$\hat{\beta}_c \;\triangleq\; \beta\sqrt{\frac{\mu}{g_c}}\,. \tag{6.43}$$

If $\beta = 0$ and $a \neq 0$

$$J \;\sim\; \frac{1}{\mu a} \cdot (1 - e^{-\mu a c})\,. \tag{6.44}$$

If $\beta = 0$ and $a = 0$

$$J \;\sim\; c\,. \tag{6.45}$$

If $\beta < 0$

$$J \;\sim\; \frac{e^{c(\lambda-n\mu)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\}. \tag{6.46}$$

[1] Customers that do not get service within a deterministic target time are sent to another call center or to the VRU.

b. If $\beta > 0$

$$J_1 \;=\; \frac{1}{\lambda} \cdot \frac{1}{\beta^2 \mu} + o\left(\frac{1}{\lambda}\right) . \tag{6.47}$$

If $\beta = 0$ and $a \neq 0$

$$J_1 \;\sim\; \frac{1}{\mu^2 a^2} \cdot (1 - e^{-\mu a c}) - \frac{c e^{-\mu a c}}{\mu a} . \tag{6.48}$$

If $\beta = 0$ and $a = 0$

$$J_1 \;\sim\; \frac{c^2}{2} . \tag{6.49}$$

If $\beta < 0$

$$J_1 \;\sim\; \frac{c e^{c(\lambda - n\mu)}}{\sqrt{\lambda}} \cdot \left\{ \frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} . \tag{6.50}$$

Remark 6.4 In the case $\beta = 0$, performance measures are very sensitive to the remaining term $n - \lambda/\mu$. Therefore, in (6.41) this term is asymptotically small in comparison to $o(\sqrt{\lambda})$ in the other cases.

Theorem 6.4 (Performance measures) Under the assumptions of Lemma 6.3, the performance measures of the M/M/n+G system with delayed patience distribution are approximated by:

a. Delay probability.

If $\beta > 0$, the asymptotic delay probability coincides with the Erlang-C approximation [9]:

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1} . \tag{6.51}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{a}{1 - e^{-\mu a c}} + o\left(\frac{1}{\sqrt{n}}\right) . \tag{6.52}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \frac{\pi}{2} \cdot \frac{1}{\mu c} + o\left(\frac{1}{\sqrt{n}}\right) . \tag{6.53}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; e^{-c(\lambda - n\mu)} \cdot \frac{\frac{1}{h(-\beta)\sqrt{\mu}}}{-\frac{1}{\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}} . \tag{6.54}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;\sim\; \frac{e^{-c(n\mu - \lambda)}}{\sqrt{\lambda}} \cdot \left\{ \beta\sqrt{\mu} - \frac{\beta^2 \mu}{h(\hat{\beta}_c)\sqrt{g_c}} \right\} . \tag{6.55}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{a e^{-\mu a c}}{1 - e^{-\mu a c}} + o\left(\frac{1}{n}\right). \tag{6.56}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{1}{\mu c} + o\left(\frac{1}{n}\right). \tag{6.57}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.58}$$

(See Remark 6.2 on page 17.)

c. Average offered waiting time.

If $\beta > 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\beta\mu} + o\left(\frac{1}{\sqrt{n}}\right) \tag{6.59}$$

(Erlang-C approximation).

If $\beta = 0$ and $a \neq 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{1}{\mu a} - \frac{c e^{-\mu a c}}{1 - e^{-\mu a c}}. \tag{6.60}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{c}{2}. \tag{6.61}$$

If $\beta < 0$

$$\mathrm{E}[V] \;\sim\; \mathrm{E}[V \mid V > 0] \;\sim\; c. \tag{6.62}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]. \tag{6.63}$$

Remark 6.5 Formulae (6.60)-(6.63) imply that, for $\beta \leq 0$, average wait (both offered and actual) converges to positive constants. That distinguishes the case of delayed distributions from Theorems 6.1 and 6.2, where $\mathrm{E}[W]$ converged to zero.

The important case of deterministic patience times gives rise to similar statements:

Theorem 6.5 (Deterministic patience) Assume that patience time is deterministic and equal to $c > 0$.

a. Delay probability.

If $\beta > 0$:

$$\mathrm{P}\{\mathrm{W} > 0\} \;\sim\; \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}. \tag{6.64}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{a}{1 - e^{-\mu a c}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.65}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \frac{\pi}{2} \cdot \frac{1}{\mu c} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.66}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; e^{-c(\lambda - n\mu)} \cdot \frac{-\beta}{h(-\beta)}. \tag{6.67}$$

b. Probability to abandon.

If $\beta > 0$

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;\sim\; \frac{e^{-c(n\mu - \lambda)}}{\sqrt{\lambda}} \cdot \beta\sqrt{\mu}\,. \tag{6.68}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;=\; \frac{1}{n} \cdot \frac{a e^{-\mu a c}}{1 - e^{-\mu a c}} + o\left(\frac{1}{n}\right). \tag{6.69}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;=\; \frac{1}{n} \cdot \frac{1}{\mu c} + o\left(\frac{1}{n}\right). \tag{6.70}$$

If $\beta < 0$

$$\mathrm{P}\{\mathrm{Ab} | W > 0\} \;=\; \frac{-\beta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.71}$$

c. Average offered waiting time.

If $\beta > 0$

$$\mathrm{E}[V \mid V > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\beta\mu} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.72}$$

If $\beta = 0$ and $a \neq 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{1}{\mu a} - \frac{c e^{-\mu a c}}{1 - e^{-\mu a c}}\,. \tag{6.73}$$

If $\beta = 0$ and $a = 0$

$$\mathrm{E}[V \mid V > 0] \;\sim\; \frac{c}{2}\,. \tag{6.74}$$

If $\beta < 0$

$$\mathrm{E}[V] \;\sim\; \mathrm{E}[V \mid V > 0] \;\sim\; c\,. \tag{6.75}$$

d. Average waiting time.

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0]\,. \tag{6.76}$$

6.1.4 Patience with balking

Lemma 6.4 (Building blocks) Consider the QED operational regime

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \qquad \lambda \to \infty\,.$$

Assume that the patience-time distribution has an atom at zero. In other words, if wait is encountered, customers abandon immediately with probability $\mathrm{P}\{\mathrm{Blk}\} > 0$, or $\bar{G}(0) = 1 - \mathrm{P}\{\mathrm{Blk}\}$. Assume, in addition, that the survival function $\bar{G}$ is differential at the origin: $\bar{G}'(0) = -g_0$. (Here g_0 is the right-side derivative of the patience-time distribution function at the origin.) Then

a.

$$J \;=\; \frac{1}{\lambda \cdot \mathrm{P}\{\mathrm{Blk}\} + (n\mu - \lambda)} - \frac{g_0}{\lambda^2 \cdot \mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right). \tag{6.77}$$

b.

$$J_1 \;=\; \frac{1}{n^2\mu^2\mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^2}\right). \tag{6.78}$$

Theorem 6.6 (Performance measures) Under the assumptions of Lemma 6.4, the performance measures of the M/M/n+G queueing system in the QED regime can be approximated by:

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

$$\mathrm{P}\{V > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.79}$$

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

$$\mathrm{P}\{W > 0\} \;\sim\; \frac{1}{\sqrt{n}} \cdot \frac{(1 - \mathrm{P}\{\mathrm{Blk}\}) \cdot h(-\beta)}{\mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.80}$$

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

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \mathrm{P}\{\mathrm{Blk}\} + \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{6.81}$$

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

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\} \cdot (1 - \mathrm{P}\{\mathrm{Blk}\})} + o\left(\frac{1}{n}\right). \tag{6.82}$$

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

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1}{\sqrt{n}} \cdot h(-\beta) + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.83}$$

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

$$\mathrm{E}[V|V>0] \;=\; \frac{1}{n} \cdot \frac{1}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{6.84}$$

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

$$\mathrm{E}[V] \;=\; \frac{1}{n^{3/2}} \cdot \frac{h(-\beta)}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^{3/2}}\right). \tag{6.85}$$

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

$$\mathrm{E}[W|W>0] \;=\; \frac{1}{n} \cdot \frac{1}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right). \tag{6.86}$$

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

$$\mathrm{E}[W] \;=\; \frac{1}{n^{3/2}} \cdot \frac{(1-\mathrm{P}\{\mathrm{Blk}\}) \cdot h(-\beta)}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{n^{3/2}}\right). \tag{6.87}$$

6.1.5 Patience with scaled balking

Lemma 6.5 (Building blocks) Assume that the patience distribution depends on the system size n. Specifically, let the balking probability $\mathrm{P}_n\{\mathrm{Blk}\} = \frac{p_b}{\sqrt{n}}$, 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: $\bar{G}_n'(0) = -g_0$. Then

a.

$$J \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \tag{6.88}$$

where

$$\hat{\beta} \;\triangleq\; (\beta + p_b) \cdot \sqrt{\frac{\mu}{g_0}}. \tag{6.89}$$

b.

$$J_1 \;=\; \frac{1}{n\mu g_0}\left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right). \tag{6.90}$$

Theorem 6.7 (Performance measures) Under the assumptions of Lemma 6.5, the performance measures of the M/M/n+G queueing system in the QED regime can be approximated by:

a. The probability of delay and positive offered wait converge to a constant that depends on β, p_b and $\frac{g_0}{\mu}$:

$$\mathrm{P}\{V > 0\} \;\sim\; \mathrm{P}\{W > 0\} \;\sim\; \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \tag{6.91}$$

where $\hat{\beta}$ is defined by formula (6.89).

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

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.92}$$

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.93}$$

The unconditional probability to abandon P{Ab} also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (6.92) and (6.91).

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

$$\mathrm{E}[V|V > 0] \;=\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[h(\hat{\beta}) - \hat{\beta}\right] + o\left(\frac{1}{\sqrt{n}}\right). \tag{6.94}$$

The average offered wait $\mathrm{E}[V]$ also decreases at rate $\frac{1}{\sqrt{n}}$ and can be approximated by the product of (6.94) and (6.91).

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

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W > 0] \;\sim\; \mathrm{E}[V \mid V > 0] \tag{6.95}$$

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:

$$\frac{\mathrm{P}\{\mathrm{Ab}|W > 0\}}{\mathrm{E}[W|W > 0]} \;\sim\; g_0\,. \tag{6.96}$$

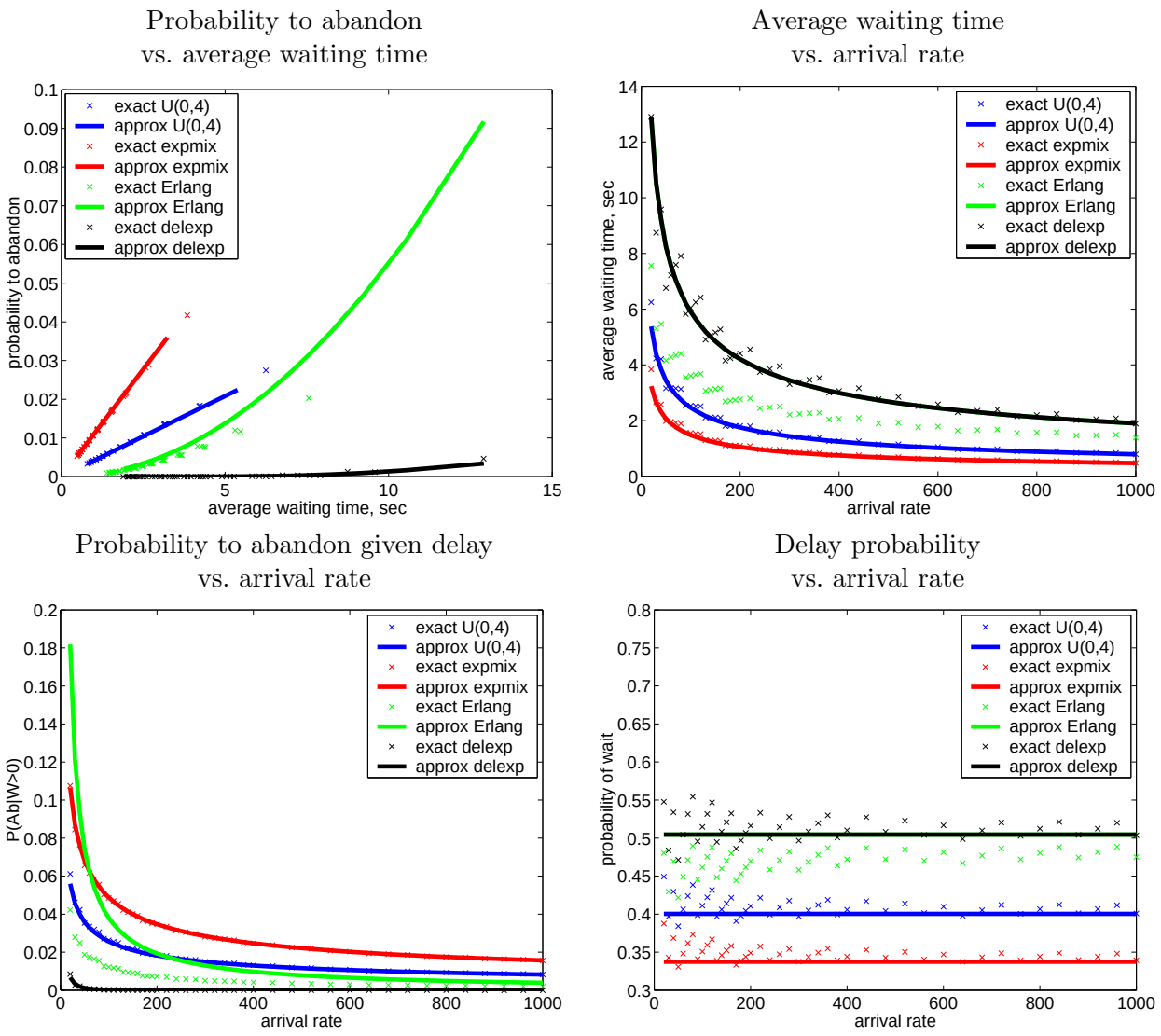
6.2 Numerical experiments

In the main paper we analyzed the quality of QED approximations for service grade $\beta = 0$. Here we perform experiments with several other service grades.

Example 1 (Figure 2): $\beta = 0.5$. The approximations for the first two distributions are excellent again. The slopes of the two corresponding curves in the first plot remain the same as

in Figure 6 of the main paper: 0.25 and 2/3, respectively. Note that, in contrast to that figure, the difference between exact values and approximations does not decrease monotonically in λ. That is due to approximation of the QED staffing level in formula (4.35) from the main paper by the nearest integer value.

Figure 2: **Service grade $\beta = 0.5$, performance measures and approximations**

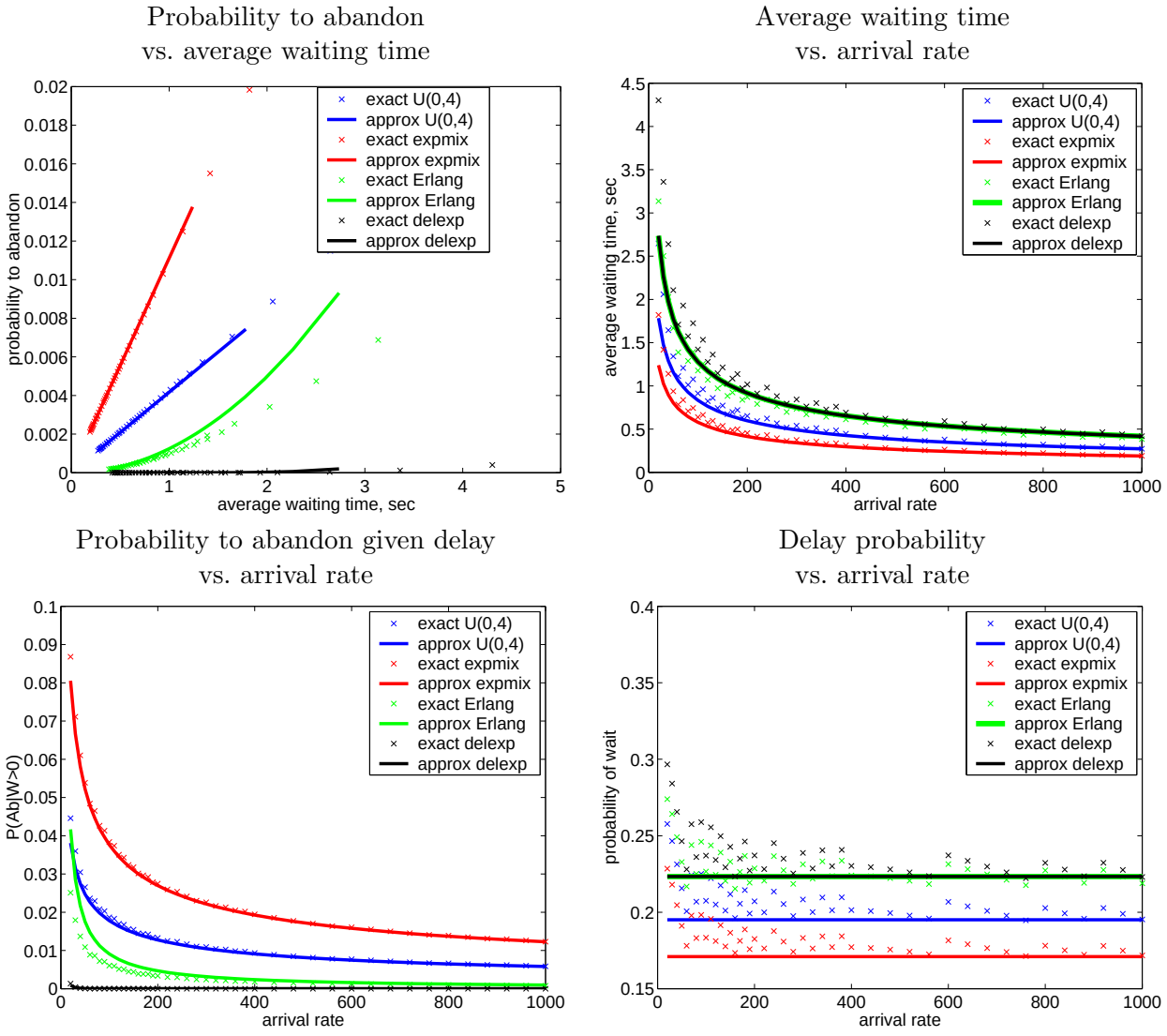


The delayed exponential distribution also demonstrates very good fit: the average wait and the delay probability are very close to the Erlang-C approximation, and the probability to abandon decreases exponentially.

However, quality of approximations for the Erlang distribution is not so good (in fact, the worst one among all special cases considered in this subsection). Approximations for the aver-

age wait and the delay probability coincide with Erlang-C formulae (and, therefore, with the approximation for delayed exponential). The fit of $\mathrm{P}\{W > 0\}$ is not bad at all. However, the fit of $\mathrm{E}[W]$ is less good and the fit of $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ is the worst of all. The reason seems to be unstableness of approximation (6.31) for small positive service grades β.

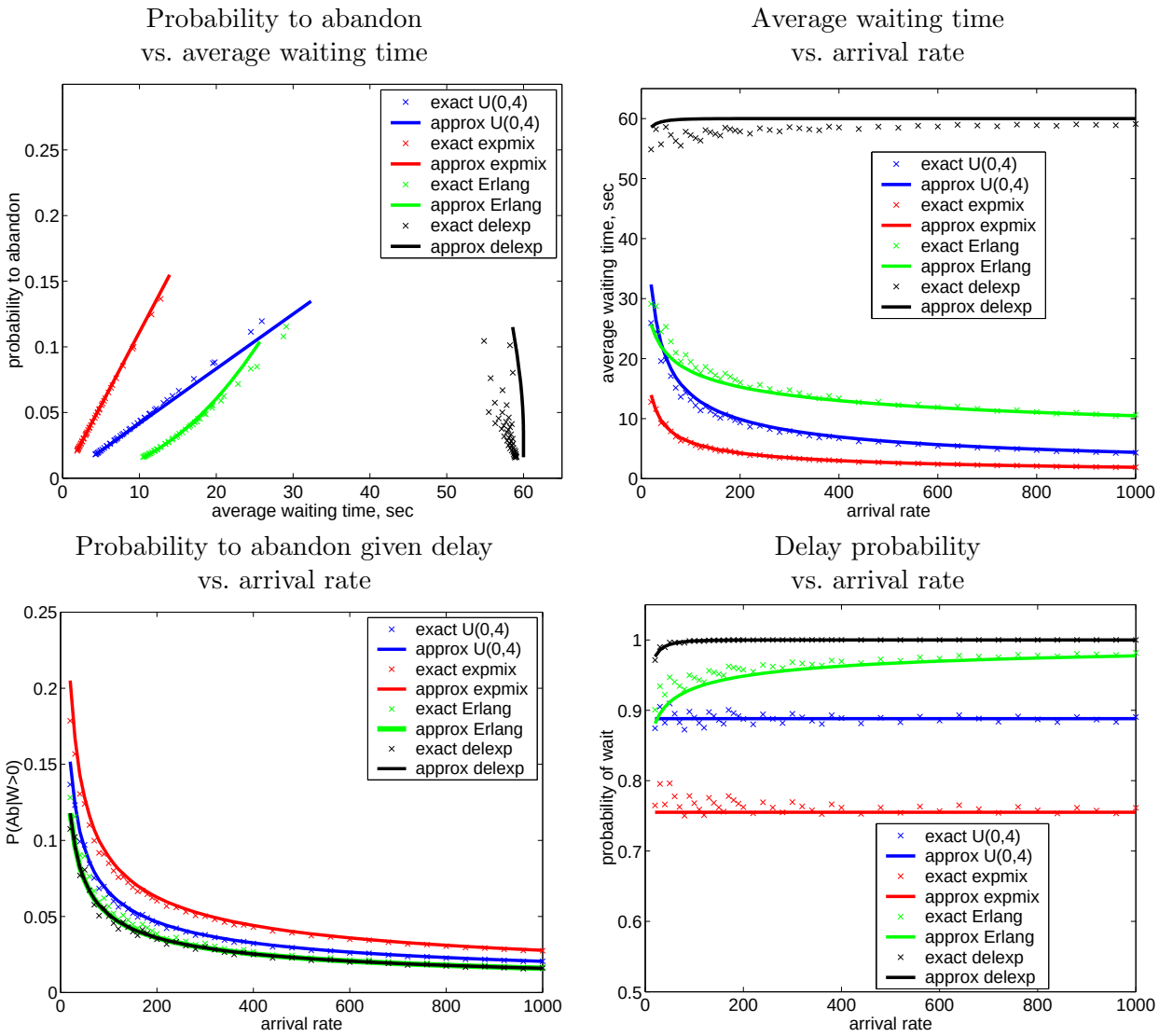
Figure 3: **Service grade $\beta = 1$, performance measures and approximations**



Example 2 (Figure 3): $\beta = 1$. Now the approximations for the Erlang distribution are much better than in Figure 2. In particular, the fit of $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ graph is reasonable for small values of λ and good for large values. (Recall from formula (6.31) that conditional probability

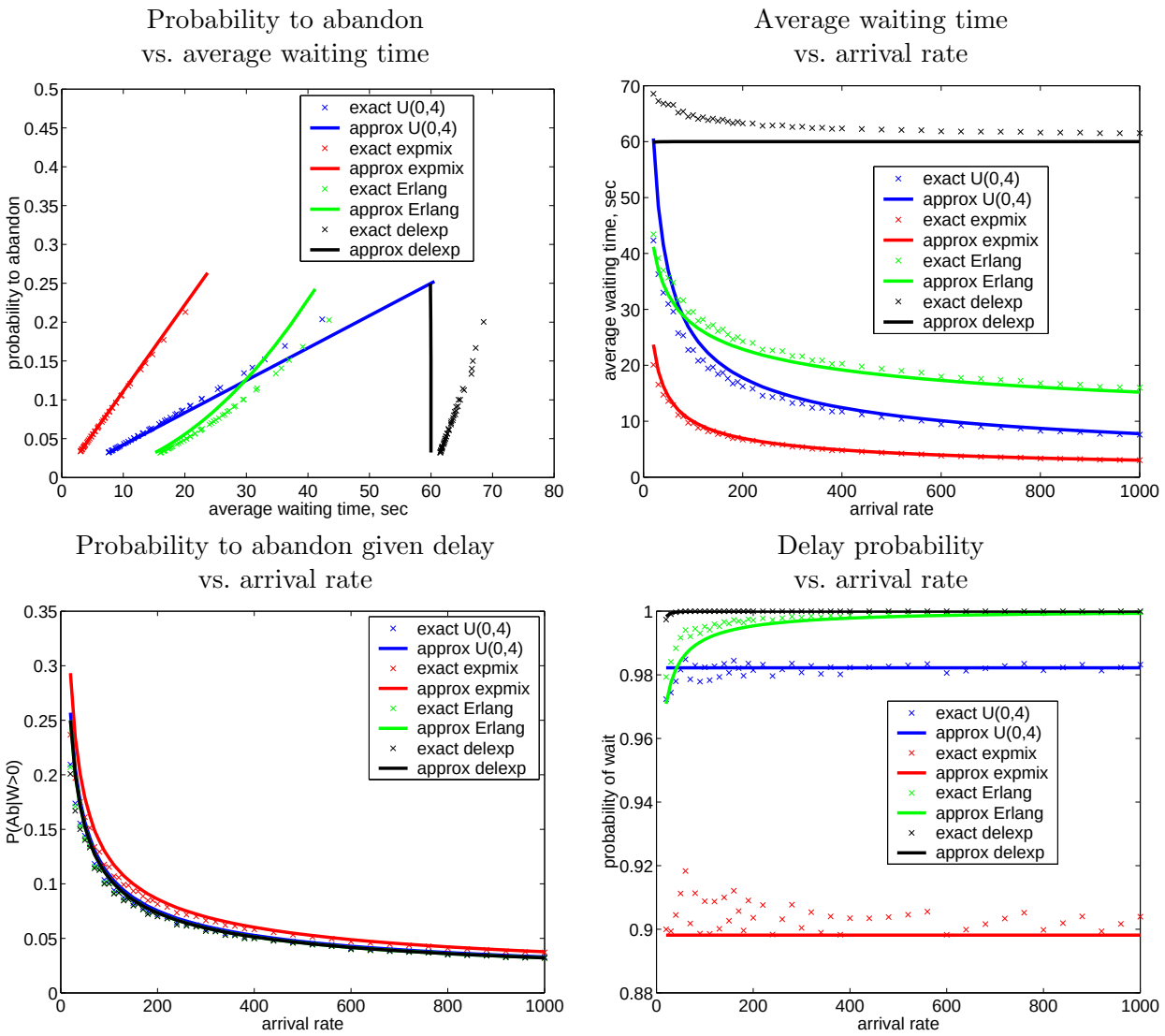
to abandon decreases at rate $1/n$.) In the delayed exponential case, the probability to abandon is negligible for all values of λ.

Figure 4: **Service grade $\beta = -0.5$, performance measures and approximations**



Example 3 (Figure 4): $\beta = -0.5$. The fit for the first two distributions ($g_0 > 0$) is fine. The approximation for $\mathrm{P}\{\mathrm{Ab}|W > 0\}$ coincides for the last two distributions with $g_0 = 0$. (See Remark 6.2.) The average wait decreases at rate $n^{-1/4}$ for the Erlang patience and converges to delay time in the delayed exponential case. Finally, in the last two cases delay probability converges to one exponentially (but with very different rates).

Figure 5: **Service grade $\beta = -1$, performance measures and approximations**



Example 4 (Figure 5): $\beta = -1$. Here we encounter two interesting phenomena. First, the conditional probabilities to abandon start to be very similar for the four distributions and close to $-\beta/\sqrt{n}$. (Recall formula (4.6) from the main paper and take into account that $h(\hat{\beta})$ is small for large negative β.)

Another interesting phenomenon is observed in the last plot: $\mathrm{P}\{W > 0\}$ curve for exponential mixture is relatively far from the uniform one. To explain it, note that for large negative β

$$\mathrm{P}\{W = 0\} \;\approx\; \frac{\sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}}{1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}} \;\approx\; \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\,,$$

recall that the normal hazard-rate $h(\cdot)$ decreases rapidly for large negative $\hat{\beta}$, and that the absolute value of $\hat{\beta}$ is larger for the uniform distribution. (Recall definition (4.3) from the main paper.)

Conclusions.

- Overall, the QED approximations are very good even for moderate staffing levels. Below (Subsections 7.3 and 8.2) we compare them with the quality-driven and the efficiency-driven approximations observing that, in most cases, the QED approximations are preferable.
- In the main case ($g_0 > 0$), the linear P{Ab} / E[W] relation is confirmed for all values of the service grade.
- For relatively large positive β we observe convergence to the Erlang-C asymptotic formulae for the average wait and the delay probability.
- For relatively large negative β the probability to abandon converges to $-\beta/\sqrt{n}$ for all distributions in consideration. (Recall Remark 6.2 after Theorem 6.2.)

6.3 Proofs of the QED results

Proof of Lemma 6.1. We provide a detailed proof of **b** and prove a new asymptotic statement **d** that was not presented in the main paper.

b. In the QED regime,

$$J_1 \;=\; \int_0^\infty x\cdot\exp\{h_\lambda(x)\}dx \;=\; \int_0^\infty x\cdot\exp\left\{\int_0^x\left[\lambda(\bar{G}(u)-1)-\beta\sqrt{\lambda\mu}\right]du\right\}dx\,.$$

Straightforward calculations imply that

$$\begin{aligned} J_{1A} \;\triangleq\; \int_0^\infty x\cdot\exp\left\{-x\beta\sqrt{\lambda\mu}-\frac{\lambda g_0 x^2}{2}\right\}dx \;&=\; \frac{1}{\lambda g_0}-\frac{\beta}{\lambda g_0}\sqrt{\frac{2\pi\mu}{g_0}}\exp\left\{\frac{\beta^2\mu}{2g_0}\right\}\left[1-\Phi\left(\beta\sqrt{\frac{\mu}{g_0}}\right)\right] \\ &=\; \frac{1}{n\mu g_0}\left[1-\frac{\hat{\beta}}{h(\hat{\beta})}\right]+o\left(\frac{1}{n}\right)\,. \end{aligned} \tag{6.97}$$

Asymptotic equivalence between J_1 and J_{1A} is demonstrated via the Laplace argument, using inequality (11.7) from the main paper. Approximation for $\int_0^\delta$ integrals is proved very similarly to **a**. Consider the second part of the argument, exponential bounds for the $\int_\delta^\infty$ integrals above:

$$\int_\delta^\infty x\cdot\exp\{h_\lambda(x)\}dx \;\le\; \int_\delta^\infty x\cdot\exp\left\{-\alpha\lambda\left(x-\frac{\delta}{2}\right)-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}dx\,,$$

(α was defined by formula (11.9) in the main paper)

$$=\; \exp\left\{\alpha\lambda\frac{\delta}{2}-\frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\}\cdot\int_\delta^\infty xe^{-\alpha\lambda x}dx$$

$$= \exp\left\{\alpha\lambda\frac{\delta}{2} - \frac{\delta}{2}\beta\sqrt{\lambda\mu}\right\} \cdot \left[\frac{\delta}{\alpha\lambda}e^{-\alpha\lambda\delta} + \frac{1}{(\alpha\lambda)^2}e^{-\alpha\lambda\delta}\right] = o(e^{-\nu\lambda}), \quad \nu > 0\,. \tag{6.98}$$

The tail part of the J_{1A} integral,

$$\int_{\delta}^{\infty} x \cdot \exp\left\{-x\beta\sqrt{\lambda\mu} - \frac{\lambda g_0 x^2}{2}\right\} dx\,,$$

can be treated as a special case of J (linear survival function near the origin).

d. First, calculate the integral

$$J_{2A} \triangleq \int_0^{\infty} x^2 \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx = \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} \cdot \int_0^{\infty} x^2 \exp\left\{\frac{-\lambda g_0\left(x + \frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}\right)^2}{2}\right\} dx\,,$$

(changing variables)

$$= \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} \cdot \int_{\frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}}^{\infty} \left(y - \frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}\right)^2 \exp\left\{\frac{-\lambda g_0 y^2}{2}\right\} dy\,.$$

and, after exact calculations,

$$= \frac{\sqrt{2\pi}}{(\lambda g_0)^{3/2}}\left(1 + \frac{\beta^2\mu}{g_0}\right)\left[1 - \Phi\left(\beta\sqrt{\frac{\mu}{g_0}}\right)\right] \cdot \exp\left\{\frac{\beta^2\mu}{2g_0}\right\} - \frac{1}{(\lambda g_0)^{3/2}} \cdot \beta\sqrt{\frac{\mu}{g_0}}$$

$$= \frac{1}{(n\mu g_0)^{3/2}}\left[\frac{\hat{\beta}^2 + 1}{h(\hat{\beta})} - \hat{\beta}\right] + o\left(\frac{1}{n^{3/2}}\right)\,.$$

The last equality follows from the definition of $\hat{\beta}$ and $\lambda \sim n\mu$ $(\lambda, n \to \infty)$.

Finally, in the same way as in Parts **a** and **b** of Lemma 11.1 from the main paper, we can validate the approximation of

$$J_2 = \int_0^{\infty} x^2 \cdot \exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\} dx$$

by

$$J_{2A} = \int_0^{\infty} x^2 \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\} dx\,.$$

Proof of Theorem 6.1.

d. Here an alternative approach to the proof in the main paper is presented.

We must prove that

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[W]}{\mathrm{E}_\lambda[V]} = 1\,,$$

where the performance measures are indexed by the arrival rate in the QED regime. Recall that $W = \min(V, \tau)$, where V and τ are independent.

It can be derived from the proof of Lemma 6.1 (Part **b**) that, $\forall\,\delta > 0$,

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \delta]}{\mathrm{E}_\lambda[V]} \;=\; \frac{\int_\delta^\infty x v_\lambda(x) dx}{\int_0^\infty x v_\lambda(x) dx} \to 0\,. \tag{6.99}$$

(Specifically, formula (6.98) shows that an exponential bound is available for $\int_\delta^\infty$.)
Now,

$$\begin{aligned} & \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \tau]}{E_\lambda[V]} \\ = \;& \lim_{\lambda\to\infty} \left(\frac{\mathrm{E}_\lambda[V; V > \tau; \tau > \delta]}{\mathrm{E}_\lambda[V]} + \frac{\mathrm{E}_\lambda[V; V > \tau; \tau < \delta]}{\mathrm{E}_\lambda[V]} \right) \\ \le \;& \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; \tau < \delta]}{\mathrm{E}_\lambda[V]} \;=\; P\{\tau < \delta\}\,. \end{aligned} \tag{6.100}$$

The first term of (6.100) converges to zero due to (6.99). The last equality follows from the independence between V and τ. The probability $P\{\tau < \delta\}$ can be made arbitrarily small, since τ has no mass at zero. Hence,

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V > \tau]}{\mathrm{E}_\lambda[V]} = 0 \;\text{ and }\; \lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V; V \le \tau]}{\mathrm{E}_\lambda[V]} = 1\,. \tag{6.101}$$

Now,

$$\mathrm{E}_\lambda[W] \;=\; \mathrm{E}_\lambda[\min(V;\tau)] \;=\; \mathrm{E}_\lambda[V; V \le \tau] + \mathrm{E}_\lambda[\tau; \tau < V] \;\sim\; \mathrm{E}_\lambda[V]\,.$$

The second statement of **d** follows from $\mathrm{P}\{W > 0\} \sim \mathrm{P}\{V > 0\}$.

f. Use formula (3.17) and the QED asymptotics for J and J_1:

$$\begin{aligned} \mathrm{E}[V \mid \mathrm{Ab}] \;&=\; \frac{(\lambda - n\mu)J_1 + J}{(\lambda - n\mu)J + 1} \\ &\sim\; \frac{-\beta\mu\sqrt{n}J_1 + J}{-\beta\mu\sqrt{n}J + 1} \\ &\sim\; \frac{1}{\sqrt{n}} \cdot \frac{-(\beta/g_0)\cdot(1 - \hat\beta/h(\hat\beta)) + 1/(\sqrt{\mu g_0}\cdot h(\hat\beta))}{1 - \hat\beta/h(\hat\beta)} \\ &\sim\; \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{g_0\mu}} \left[\frac{1}{h(\hat\beta) - \hat\beta} - \hat\beta \right]. \end{aligned}$$

Now we shall prove formula (6.13). Note that

$$\mathrm{E}[W|\mathrm{Ab}] \;=\; \mathrm{E}[\tau|\tau < V] \;=\; \frac{\mathrm{E}[\tau; \tau < V]}{\mathrm{P}\{\mathrm{Ab}\}} \;=\; \frac{\int_0^\infty \mathrm{E}[\tau; \tau < x] v(x) dx}{\mathrm{P}\{\mathrm{Ab}\}}$$

$$= \;\frac{\int_0^\infty v(x) \left(\int_0^x t dG(t)\right) dx}{\mathrm{P}\{\mathrm{Ab}\}}\,,$$

where $v(x)$ is the density of the offered wait (recall formula (2.3)). Then

$$\int_0^x t dG(t) \;=\; xG(x) - \int_0^x G(t) dt\,.$$

Note that $\forall \epsilon > 0\ \exists \delta > 0$ such that for $x \in [0, \delta]$,

$$(g_0 - \epsilon)x^2 \ \leq \ xG(x) \ \leq \ (g_0 + \epsilon)x^2$$

and

$$\frac{(g_0 - \epsilon)x^2}{2} \ \leq \ \int_0^x G(t)dt \ \leq \ \frac{(g_0 + \epsilon)x^2}{2}\,.$$

Then, for $x \in [0, \delta]$,

$$\frac{x^2}{2}(g_0 - 3\epsilon) \ \leq \ \int_0^x tdG(t) \ \leq \ \frac{x^2}{2}(g_0 + 3\epsilon)\,.$$

Since $\int_0^x tdG(t)$ is bounded by x^2, we can construct an exponential bound for $\int_\delta^\infty v(x)\left(\int_0^x tdG(t)\right)dx$ in the spirit of Lemma 6.1, part **b**. Then, based on the Laplace method, we deduce that

$$\int_0^\infty v(x)\left(\int_0^x tdG(t)\right)dx \ \sim \ \frac{g_0}{2}\int_0^\infty x^2 v(x)dx\,,$$

and

$$\mathrm{E}[W|\mathrm{Ab}] \ \sim \ \frac{g_0}{2\mathrm{P}\{\mathrm{Ab}\}}\cdot\lambda\pi_{n-1}J_2 \ \sim \ \frac{g_0}{2\mathrm{P}\{\mathrm{Ab}\}}\cdot\frac{\lambda J_2}{\mathcal{E}+\lambda J}\,. \tag{6.102}$$

From part **d** of Lemma 6.1 we observe that the numerator of (6.102) is equal to

$$\lambda g_0 J_2 \ = \ \frac{1}{\sqrt{n\mu g_0}}\cdot\left[\frac{1+\hat\beta^2}{h(\hat\beta)} - \hat\beta\right] + o\left(\frac{1}{\sqrt{n}}\right)\,.$$

For the denominator of (6.102)

$$2\mathrm{P}\{\mathrm{Ab}\}\cdot(\mathcal{E}+\lambda J) \ = \ 2(1+(\lambda - n\mu)J) \ \sim \ 2(1-\beta\mu\sqrt{n}J) \ \sim \ 2\cdot\left(1 - \frac{\hat\beta}{h(\hat\beta)}\right)\,.$$

Dividing the numerator of (6.102) by the denominator:

$$\begin{aligned}\mathrm{E}[V|\mathrm{Ab}] \ &= \ \frac{1}{\sqrt{n}}\cdot\frac{1}{2\sqrt{g_0\mu}}\cdot\frac{1+\hat\beta^2-\beta h(\hat\beta)}{h(\hat\beta)-\hat\beta} + o\left(\frac{1}{\sqrt{n}}\right)\\ &= \ \frac{1}{\sqrt{n}}\cdot\frac{1}{2\sqrt{g_0\mu}}\cdot\left[\frac{1}{h(\hat\beta)-\hat\beta} - \hat\beta\right] + o\left(\frac{1}{\sqrt{n}}\right)\,.\end{aligned}$$

Finally, we prove inequalities (6.15). The right one is a consequence of (5.2) and (5.5). The left inequality is equivalent to

$$2h(x) \ > \ x + \frac{1}{h(x)-x}$$

or

$$2h^2(x) - 3xh(x) + x^2 - 1 \ > \ 0\,,$$

which follows from convexity of h and formula (5.3).

g. From formula (2.3) for the density of the virtual offered wait it follows that

$$\mathrm{P}\left\{\frac{V}{\mathrm{E}[S]} > \frac{t\sqrt{\mu}}{\sqrt{\lambda}} \,\middle|\, V>0\right\} \;=\; \frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{\int_0^x \left[\lambda(\bar G(u)-1)-\beta\sqrt{\lambda\mu}\right] du\right\} dx}{J}. \qquad (6.103)$$

Then using the Laplace method we show that the last expression is equivalent to

$$\frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{-\beta\sqrt{\lambda\mu}x-\frac{\lambda g_0 x^2}{2}\right\} dx}{J} \;=\; \frac{\exp\left\{\dfrac{\beta^2\mu}{2g_0}\right\}\displaystyle\int_{\frac{t}{\sqrt{\mu\lambda}}+\frac{\beta}{g_0}\sqrt{\frac{\mu}{\lambda}}}^{\infty} \exp\left\{-\frac{\lambda g_0 y^2}{2}\right\} dy}{J}$$

(using the asymptotic expression for J, taken from Lemma 11.1 of the main paper, Part **a**)

$$\sim\; \frac{\bar\Phi\left(\hat\beta+\sqrt{\dfrac{g_0}{\mu}}\cdot t\right)}{\bar\Phi(\hat\beta)}.$$

Now, in order to complete the proof, we need to substitute $\sqrt{\dfrac{\lambda}{\mu}}$ by $\sqrt{n}$ and the virtual offered wait V by the waiting time W in the left-hand part of (6.103). The validity of the first substitution can be verified using $\lambda \sim n\mu$.

For the second substitution we must prove

$$\mathrm{P}\{W>0\} \;\sim\; \mathrm{P}\{V>0\} \quad \text{and} \quad \mathrm{P}\left\{\frac{W}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \;\sim\; \mathrm{P}\left\{\frac{V}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\}.$$

Both relations directly follow from $W=\min(V,\tau)$ and $V \xrightarrow{p} 0$ $(n\to\infty)$ (see part **d**).

h. Conditional probability to abandon:

$$\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, \frac{V}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \;=\; \frac{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)(1-\bar G(x))dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)dx}$$

$$\sim\; \frac{g_0\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} xv(x)dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} v(x)dx} \;\sim\; \frac{g_0\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} x\exp\left\{-\beta\sqrt{\lambda\mu}x-\frac{\lambda g_0 x^2}{2}\right\} dx}{\displaystyle\int_{t/\sqrt{\lambda\mu}}^{\infty} \exp\left\{-\beta\sqrt{\lambda\mu}x-\frac{\lambda g_0 x^2}{2}\right\} dx}. \qquad (6.104)$$

Calculating the numerator of (6.104), we get

$$\frac{1}{\lambda}\left[\exp\left\{-\frac{g_0 t^2}{2\mu}-\beta t\right\} - \hat\beta\cdot\frac{\bar\Phi\left(\hat\beta+\sqrt{\dfrac{g_0}{\mu}}\cdot t\right)}{\bar\Phi(\hat\beta)}\right].$$

The denominator of (6.104) is equal to

$$\frac{1}{\sqrt{\lambda g_0}} \cdot \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{\frac{g_0}{\mu}} \cdot t\right)}{\bar{\Phi}(\hat{\beta})} .$$

Dividing the numerator by the denominator, we get (6.17).

Proof of Lemma 6.2.

a. $\boldsymbol{\beta > 0}$. Here and in the proofs below we denote $o(\cdot)$ deviation terms in the staffing rules (6.20) and (6.21) by $f(\lambda)$. Apply Lemma 4.1 with

$$k_1 = \frac{1}{2}; \quad l_1 = 1; \quad k_2 = 1; \quad l_2 = k+2; \quad m = 0; \tag{6.105}$$

(condition $\dfrac{k_1}{l_1} > \dfrac{k_2}{l_2}$ is valid for $k > 0$) to derive that

$$J_A \triangleq \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \tag{6.106}$$

$$= \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \; - \; \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x\right\} \cdot \frac{\lambda g_{0k} x^{k+2}}{(k+2)!} dx + o\left(\frac{1}{\lambda^{(k+1)/2}}\right)$$

$$= \frac{1}{n\mu - \lambda} \; - \; \frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+3}} \; + \; o\left(\frac{1}{\lambda^{(k+1)/2}}\right) .$$

(We use that $n\mu - \lambda = \beta\sqrt{\lambda\mu} + f(\lambda)\mu$.) Now note that

$$J \; = \; \int_0^\infty \exp\left\{\lambda \int_0^x \bar{G}(u)du - \lambda x - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \, . \tag{6.107}$$

Under the assumptions of Lemma 6.2, $\forall \epsilon > 0 \; \exists \delta > 0$ such that, for $u \in [0, \delta]$,

$$1 - \frac{(g_{0k} + \epsilon)u^{k+1}}{(k+1)!} \; \leq \; \bar{G}(u) \; \leq \; 1 - \frac{(g_{0k} - \epsilon)u^{k+1}}{(k+1)!} \, . \tag{6.108}$$

From Lemma 4.3 $(m = 0, n = 0, k = 1/2, l = 1)$, there exists $\nu > 0$ such that

$$\int_\delta^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \; = \; o\left(e^{-\lambda^\nu}\right) \, . \tag{6.109}$$

Formulae (6.108) and (6.109) enable us to apply the Laplace method in order to show that J from (6.107) can be approximated by J_A from (6.106).

We use a similar reasoning in order to derive (6.23). (Lemma 4.1 is applied with $m = 1$ and other parameters taken from (6.105).) Specifically,

$$J_{1A} \triangleq \int_0^\infty x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \tag{6.110}$$

$$= \int_0^\infty x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \; - \int_0^\infty x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x\right\} \cdot \frac{\lambda g_{0k} x^{k+2}}{(k+2)!} dx + o\left(\frac{1}{\lambda^{(k+1)/2}}\right)$$

$$= \frac{1}{(n\mu-\lambda)^2} \; - \; \frac{(k+3)\cdot\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+4}} \; + \; o\left(\frac{1}{\lambda^{(k+2)/2}}\right) \qquad (\lambda\to\infty)$$

and, then substitute

$$J_1 \; = \; \int_0^\infty x \cdot \exp\left\{\lambda\int_0^x \bar{G}(u)du - \lambda x - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx$$

instead of (6.110).

b. $\boldsymbol{\beta = 0}$. Using Lemma 4.2 with

$$k_1 = 1, \quad l_1 = k+2, \quad k_2 = \frac{1}{k+2}, \quad l_2 = 1, \quad m = 0,$$

we get

$$J_A \; \triangleq \; \int_0^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx$$

$$= \; \frac{1}{k+2} \cdot \left(\frac{(k+2)!}{\lambda g_{0k}}\right)^{1/(k+2)} \cdot \Gamma\left(\frac{1}{k+2}\right) \; + \; o\left(\frac{1}{\lambda^{(k+1)/2}}\right), \qquad (6.111)$$

and taking $m = 1$,

$$J_{1A} \; \triangleq \; \int_0^\infty x \cdot \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx$$

$$= \; \frac{1}{k+2} \cdot \left(\frac{(k+2)!}{\lambda g_{0k}}\right)^{2/(k+2)} \cdot \Gamma\left(\frac{2}{k+2}\right) \; + \; o\left(\frac{1}{\lambda^{(k+1)/2}}\right). \qquad (6.112)$$

Then we use (6.108), the Laplace method and Lemma 4.3 in order to substitute

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

and

$$J_1 \; = \; \int_0^\infty x \cdot \exp\left\{\lambda\int_0^x \bar{G}(u)du - \lambda x - f(\lambda)\mu x\right\} dx\,,$$

into (6.111) and (6.112), respectively. Note that Lemma 4.3 cannot be applied immediately to get

$$\int_\delta^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \; = \; o\left(e^{-\lambda^\nu}\right)$$

and

$$\int_\delta^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - \lambda x - f(\lambda)\mu x\right\} dx \; = \; o\left(e^{-\lambda^\nu}\right)$$

($-f(\lambda)$ can be positive). However, this problem can be easily solved. For example,

$$\int_\delta^\infty \exp\left\{-f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \; \le \; \int_\delta^\infty \exp\left\{-\frac{1}{2}\cdot\frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx$$

for λ large enough.

c. $\beta < 0$. As in part **a**, we approximate J by

$$J_A \triangleq \int_0^\infty \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx\,, \tag{6.113}$$

and, then, apply the Laplace method to show that $J \sim J_A$. However, since $-\beta$ is a positive number, the integrand increases near zero, which requires additional work that involves somewhat cumbersome calculations. Define

$$x^* = \left(\frac{[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu]\cdot(k+1)!}{\lambda g_{0k}}\right)^{1/(k+1)},$$

to be equal to the point where the integrand of (6.113) reaches a maximum (note that x^* converges to zero at rate $\lambda^{-1/(2k+2)}$). Performing the variable change $y = x - x^*$, we get

$$\begin{aligned} J_A &= \exp\{[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu]\cdot x^*\} \\ &\cdot \int_{-x^*}^\infty \exp\left\{-\beta\sqrt{\lambda\mu}y - f(\lambda)\mu y - \frac{\lambda g_{0k}(y+x^*)^{k+2}}{(k+2)!}\right\} dy\,. \end{aligned} \tag{6.114}$$

Note that

$$\int_{-\infty}^{-x^*} \exp\{-\beta\sqrt{\lambda\mu}y\}dy = \frac{1}{-\beta\sqrt{\lambda\mu}}\exp\{\beta\sqrt{\lambda\mu}x^*\}\,.$$

Since β is negative, the integral above decreases at rate $\dfrac{\exp\{-\lambda^{k/(2k+2)}\}}{\sqrt{\lambda}}$ and we can change the integral limits in (6.114) to $\int_{-\infty}^{\infty}$. Now we expand $(y+x^*)^{k+2}$ from (6.114). The free term $(x^*)^{k+2}$ is taken out of the integral and the $(k+2)y(x^*)^{k+1}$ term is cancelled by $[-\beta\sqrt{\lambda\mu} - f(\lambda)\mu]\cdot y$. We must show now that the quadratic term in the expansion dominates the others. In other words,

$$\begin{aligned} J_A &\sim \exp\left\{\frac{k+1}{k+2}\cdot\left[\frac{(k+1)!}{g_{0k}}\right]^{1/(k+1)}\cdot(\lambda-n\mu)^{(k+2)/(k+1)}\right\}\cdot\int_{-\infty}^{\infty}\exp\left\{-\frac{\lambda g_{0k}}{2k!}(x^*)^k y^2\right\}dy \\ &= \exp\left\{\frac{k+1}{k+2}\cdot\left[\frac{(k+1)!}{\lambda g_{0k}}\right]^{1/(k+1)}\cdot(\lambda-n\mu)^{(k+2)/(k+1)}\right\} \\ &\cdot\sqrt{2\pi k!}\cdot(\lambda g_{0k})^{-1/(2k+2)}\cdot((k+1)!(\lambda-n\mu))^{-k/(2k+2)}\,. \end{aligned}$$

We shall prove that the quadratic term in the integral (6.114) dominates the cubic term, an argument that can be repeated for the terms with larger degrees of y using Remark 4.2. Ignoring cumbersome constants, we must show that

$$\int_{-\infty}^{\infty}\exp\left\{-\lambda^{(k+2)/(2k+2)}y^2 - \lambda^{(k+3)/(2k+2)}y^3\right\}dy$$

$$\sim \int_{-\infty}^{\infty} \exp\left\{-\lambda^{(k+2)/(2k+2)} y^2\right\} dy \;=\; \sqrt{2\pi}\lambda^{-(k+2)/(4k+4)}\,.$$

The equivalence above follows from Lemma 4.1 with

$$k_1 = \frac{k+2}{2k+2}\,,\quad l_1 = 2,\quad k_2 = \frac{k+3}{2k+2}\,,\quad l_2 = 3,$$

(Note that condition (4.1) prevails for $k > 0$.)

Formula (6.27) for J_1 is proved via the approximation

$$J_{1A} \;\sim\; \int_0^{\infty} x \cdot \exp\left\{-\beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_{0k} x^{k+2}}{(k+2)!}\right\} dx \;\sim\; x^* \cdot J_A\,,$$

where the second equivalence is obtained via the change of variables $y = x - x^*$.

Proof of Theorem 6.2.

a. Delay probability.

$\boldsymbol{\beta > 0}$. Recall the asymptotic expression for $\mathcal{E}$:

$$\mathcal{E} \;=\; \sqrt{\frac{\lambda}{\mu}} \cdot \frac{1}{h(-\beta)}\,,$$

which does not depend on the patience distribution G. Hence,

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

$\boldsymbol{\beta = 0}$. From Part **b** of Lemma 6.2 and taking into account that $\mathcal{E} \sim \sqrt{\frac{\lambda}{\mu}} \cdot \frac{\pi}{2}$:

$$\mathrm{P}\{W = 0\} \;=\; \frac{\mathcal{E}}{\mathcal{E} + \lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; \frac{1}{\lambda^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2\mu}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{(k+2)!}\right]^{\frac{1}{k+2}}$$

$$\sim\; \frac{1}{n^{k/(2k+4)}} \cdot \sqrt{\frac{\pi}{2}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}}\,.$$

$\boldsymbol{\beta < 0}$. Use that

$$\mathrm{P}\{W = 0\} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; \frac{1}{\sqrt{\mu\lambda} J} \cdot \frac{1}{h(-\beta)}\,.$$

b. Probability of delayed customers to abandon.

$\boldsymbol{\beta > 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V > 0\} \;=\; \frac{1 + (\lambda - n\mu)J}{\lambda J} \;\sim\; \frac{\frac{\lambda g_{0k}}{(\beta\sqrt{\lambda\mu})^{k+2}}}{\frac{\sqrt{\lambda}}{\beta\sqrt{\mu}}}$$

$$\sim \frac{g_{0k}}{\beta^{k+1}(\lambda\mu)^{(k+1)/2}} \sim \frac{1}{n^{(k+1)/2}} \cdot \frac{g_{0k}}{(\beta\mu)^{k+1}}.$$

$\boldsymbol{\beta = 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} = \frac{1-(\beta\sqrt{\lambda\mu}+f(\lambda)\mu)J}{\lambda J} \sim \frac{1}{\lambda J} \sim \frac{1}{n^{(k+1)/(k+2)}} \cdot \frac{k+2}{\Gamma\left(\frac{1}{k+2}\right)} \cdot \left[\frac{g_{0k}}{\mu^{k+1}(k+2)!}\right]^{\frac{1}{k+2}}.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{Ab}|V>0\} = \frac{1-(\beta\sqrt{\lambda\mu}+f(\lambda)\mu)J}{\lambda J} \sim -\beta\sqrt{\frac{\mu}{\lambda}} \sim \frac{-\beta}{\sqrt{n}}.$$

c. Average offered waiting time.

$\boldsymbol{\beta > 0}$.

$$\mathrm{E}[V|V>0] = \frac{J_1}{J} \sim \frac{1}{\beta\sqrt{\lambda\mu}} \sim \frac{1}{\beta\mu\sqrt{n}}.$$

$\boldsymbol{\beta = 0}$.

$$\mathrm{E}[V|V>0] = \frac{J_1}{J} \sim \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)}.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{E}[V|V>0] = \frac{J_1}{J} \sim x^* \sim \frac{1}{n^{1/(2k+2)}} \cdot \left[\frac{-\beta(k+1)!}{g_{0k}}\right]^{1/(k+1)}.$$

d. Average waiting time.

Since the survival function $\bar{G}$ is strictly decreasing near the origin, the proof of Part **d** of Theorem 6.1 can be duplicated.

Proof of Theorem 6.3.

We start with some definitions. Consider the underlying Markov process of the phase-type distribution F_T. Let S_0 denote the set of states that correspond to the positive values of the initial distribution: $i \in S_0$ iff $q_i > 0$. Then let S_1 be the set of states that can be reached by one jump from some initial state. Formally, $j \in S_1$ iff $j \notin S_0$ and there exists $i \in S_0$ such that $R_{ij} > 0$. Finally, we define recursively the set S_k which comprises states that can be reached by k jumps: $j \in S_k$ iff $j \notin S_0, \ldots, S_{k-1}$ and there exists $i \in S_{k-1}$ such that $R_{ij} > 0$. According to the definition (6.40) of L, the absorbing state $\Delta \in S_L$.

We shall number the states of the underlying Markov process in the following way: first, the states from S_0, then the states that belong to $S_1, \ldots, S_L$, etc. A relevant part of the generator

matrix is equal to:

	S_0	S_1	...	S_l	S_{l+1}	...	S_{L-1}	S_L	...	Δ
S_0	?	+	0	0	0	0	0	0	0	0
S_1	?	?	+ 0	0	0	0	0	0	0	0
...	?	?	?	? +	0	0	0	0	0	0
S_l	?	?	?	?	+	0	0	0	0	0
S_{l+1}	?	?	?	?	?	+ 0	0	0	0	0
...	?	?	?	?	?	?	? +	0	0	0
S_{L-1}	?	?	?	?	?	?	?	+	0	+

Here "?" means that the corresponding terms are irrelevant and "0" that all terms in the quadrant are zero. The sign "+" means that there are no negative terms in the quadrant and every column contains, at least, one positive term.

Formula (6.39) implies that we must prove

$$\bar{q}R^l\bar{r} \;=\; 0\,, \qquad 0 \le l \le L-2, \tag{6.115}$$

and

$$\bar{q}R^{L-1}\bar{r} \;>\; 0\,. \tag{6.116}$$

First, we want to show by induction that, for $0 \le l \le L-1$,

$$\bar{q}R^l \;=\; (x_1, \ldots, x_{N_{l-1}}, \boldsymbol{x_{N_{l-1}+1}, \ldots, x_{N_l}}, 0, \ldots, 0)\,, \tag{6.117}$$

where all vector elements between $x_{N_{l-1}+1}$ and x_{N_l} are positive. Statement (6.117) is clearly true for $l = 0$ (if we assign $N_{-1} = 0$). Assume that it is true for some $l \ge 0$. Then

$$\bar{q}R^{l+1} \;=\; (\bar{q}R^l)\cdot R \;=\; (y_1, \ldots, y_{N_l}, \boldsymbol{y_{N_l+1}, \ldots, y_{N_{l+1}}}, 0, \ldots, 0)\,.$$

The elements $y_{N_l+1}, \ldots, y_{N_{l+1}}$ are positive since they are calculated by multiplying the positive vector $(x_{N_{l-1}+1}, \ldots, x_{N_l})$ by the columns of $S_l \times S_{l+1}$ non-negative quadrant (one positive element in the column, at least). The elements after $y_{N_{l+1}}$ are zero since the upper part of the respective generator columns (right of S_{l+1} columns) is zero. Now note that

$$\bar{r} \;=\; (0, \ldots, 0, \boldsymbol{r_{N_{l-2}+1}, \ldots, r_{N_{l-1}}}, \ldots)'\,,$$

where the vector $(r_{N_{l-2}+1}, \ldots, r_{N_{l-1}})$ contains one positive element, at least. Substituting $l = L-1$ in (6.117) and multiplying by $\bar{r}$ we get (6.116). Equality (6.115) is obvious from (6.117), $l < L-1$.

Proof of Lemma 6.3. We proceed with the Laplace method from Lemma 11.1 of the main paper, using that $\bar{G}(u) = 1,\ 0 \le u \le c$, and approximating $\bar{G}(c+\epsilon) \approx 1 - \epsilon g_c$, for small $\epsilon > 0$.

a. $\boldsymbol{\beta > 0}$.

$$J \;=\; \int_0^\infty \exp\left\{\int_0^x [\lambda\bar{G}(u) - \lambda - \beta\sqrt{\lambda\mu} - f(\lambda)\mu]du\right\} dx \tag{6.118}$$

$$= \int_0^c \exp\{-\beta\sqrt{\lambda\mu}x - \mu f(\lambda)x\}dx + \exp\{(-\beta\sqrt{\lambda\mu} - \mu f(\lambda))c\}$$
$$\cdot \left[\int_c^\infty \exp\left\{-\beta\sqrt{\lambda\mu}(x-c) - \frac{\lambda g_c(x-c)^2}{2}\right\}dx + o\left(\frac{1}{\sqrt{\lambda}}\right)\right]$$
$$= \frac{1}{n\mu-\lambda}\cdot\left[1-e^{-c(n\mu-\lambda)}\right] + e^{-c(n\mu-\lambda)}\cdot\int_c^\infty \exp\left\{-\beta\sqrt{\lambda\mu}(x-c) - \frac{\lambda g_c(x-c)^2}{2}\right\}dx$$
$$+ o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\right)$$
$$= \frac{1}{n\mu-\lambda} - \frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\cdot\left\{\frac{1}{\beta\sqrt{\mu}} - \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}\right\} + o\left(\frac{e^{-c(n\mu-\lambda)}}{\sqrt{\lambda}}\right).$$

$\boldsymbol{\beta = 0}$.

$$J = \int_0^c e^{-\mu a x}dx + \int_c^\infty \exp\left\{[\lambda\bar{G}(u) - \lambda - a\mu]du\right\}dx$$
$$\sim \int_0^c e^{-\mu a x}dx = \begin{cases} \frac{1}{\mu a}\cdot(1-e^{-\mu a c}), & a \neq 0 \\ c, & a = 0 \end{cases}.$$

$\boldsymbol{\beta < 0}$. Define the expression in the exponent of (6.118) by $h_\lambda(x)$. Then, similar to the case $\beta > 0$:

$$J = \int_0^c \exp\{h_\lambda(x)\}dx + \int_c^\infty \exp\{h_\lambda(x)\}dx$$
$$= \frac{1}{\lambda-n\mu}\cdot\left[e^{c(\lambda-n\mu)}-1\right] + e^{c(\lambda-n\mu)}\cdot\int_c^\infty \exp\left\{-\beta\sqrt{\lambda\mu}(x-c) - \frac{\lambda g_c(x-c)^2}{2}\right\}dx$$
$$+ o\left(\frac{e^{c(\lambda-n\mu)}}{\sqrt{\lambda}}\right)$$
$$= \frac{e^{c(\lambda-n\mu)}}{\sqrt{\lambda}}\cdot\left\{\frac{1}{-\beta\sqrt{\mu}} + \frac{1}{h(\hat{\beta}_c)\sqrt{g_c}}\right\} + o\left(\frac{e^{c(\lambda-n\mu)}}{\sqrt{\lambda}}\right).$$

(The term $\dfrac{1}{\lambda-n\mu}$ is negligible if $\beta < 0$.)

b. $\boldsymbol{\beta > 0}$. Here we need only the first approximation term, ignoring $\int_c^\infty$:

$$J_1 = \int_0^c x\exp\{h_\lambda(x)\}dx + \int_c^\infty x\exp\{h_\lambda(x)\}dx \sim \frac{1}{(n\mu-\lambda)^2} = \frac{1}{\beta^2\lambda\mu} + o\left(\frac{1}{\lambda}\right).$$

$\boldsymbol{\beta = 0}$.

$$J_1 \sim \int_0^c x e^{-\mu a x}dx = \begin{cases} \frac{1}{\mu^2 a^2}\cdot(1-e^{-\mu a c}) - \frac{c e^{-\mu a c}}{\mu a}, & a \neq 0 \\ \frac{c^2}{2}, & a = 0 \end{cases}.$$

$\boldsymbol{\beta < 0}$.

$$\int_0^c x \exp\{h_\lambda(x)\}dx \;=\; \int_0^c x e^{(\lambda-n\mu)x}dx \;=\; \frac{ce^{c(\lambda-n\mu)}}{\lambda-n\mu} - \frac{e^{c(\lambda-n\mu)}}{(\lambda-n\mu)^2}\,. \tag{6.119}$$

The term $\int_c^\infty x \exp\{h_\lambda(x)\}dx$ can be approximated by

$$\int_c^\infty x \cdot \exp\left\{(\lambda-n\mu)x - \frac{\lambda g_c(x-c)^2}{2}\right\}dx \;=\; e^{c(\lambda-n\mu)}\int_0^\infty (y+c)\cdot \exp\left\{-\beta\sqrt{\lambda\mu}y - \frac{\lambda g_c y^2}{2}\right\}dx$$

$$= \frac{ce^{c(\lambda-n\mu)}}{h(\hat\beta_c)\sqrt{\lambda g_c}} + o\left(\frac{e^{c(\lambda-n\mu)}}{\sqrt{\lambda}}\right)\,. \tag{6.120}$$

Now formulae (6.119) and (6.120) imply (6.50).

Proof of Theorem 6.4.

a. Delay probability.

$\boldsymbol{\beta > 0}$. The asymptotic formula and its proof are the same as in Theorem 6.2.

$\boldsymbol{\beta = 0}$. Substitute (6.44), (6.45) and formula (11.3) from the main paper into

$$\mathrm{P}\{W=0\} \;=\; \frac{\mathcal{E}}{\mathcal{E}+\lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J}\,.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{W}=0\} \;=\; \frac{\mathcal{E}}{\mathcal{E}+\lambda J} \;\sim\; \frac{\mathcal{E}}{\lambda J} \;\sim\; e^{c(\lambda-n\mu)}\cdot \frac{\frac{1}{h(-\beta)\sqrt{\mu}}}{-\frac{1}{\beta\sqrt{\mu}} + \frac{1}{h(\hat\beta_c)\sqrt{g_c}}}\,.$$

b. Probability of delayed customers to abandon.

$\boldsymbol{\beta > 0}$. Formula (6.55) is derived by substituting (6.42) into

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J}\,.$$

$\boldsymbol{\beta = 0}$. Substitute (6.44) and (6.45) into

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;=\; \frac{1-a\mu J}{\lambda J}\,.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1+(\lambda-n\mu)J}{\lambda J} \;\sim\; \frac{\lambda-n\mu}{\lambda} \;\sim\; \frac{-\beta}{\sqrt{n}}\,.$$

c. Average offered waiting time.

$\boldsymbol{\beta > 0}$.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;=\; \frac{1}{\beta\mu\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)\,.$$

$\boldsymbol{\beta = 0}$. Substitute (6.44), (6.45), (6.48) and (6.49) into

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J}\,.$$

$\boldsymbol{\beta < 0}$.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim\; c\,.$$

d. Average waiting time.
According to the formulation of the theorem, $\tau = c + Y$, where $c > 0$ is a constant and Y is a random variable with a positive density at the origin. First, we prove that for all $\delta > 0$

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V > c+\delta]}{\mathrm{E}_\lambda[V]} \;=\; 0\,.$$

(which turns out equivalent to the proof in Theorem 6.1, part **d**, after the change of variables $y = x - c$). Then we continue along the lines of the proof of Theorem 6.1 via

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V > \tau]}{\mathrm{E}_\lambda[V]} \;\le\; \mathrm{P}\{\tau < c+\delta\}\,;$$

and

$$\lim_{\lambda\to\infty} \frac{\mathrm{E}_\lambda[V;\, V \le \tau]}{\mathrm{E}_\lambda[V]} \;=\; 1\,.$$

The proof of Theorem 6.5 is very similar to the proof of Theorem 6.4.

Proof of Lemma 6.4.
a. Recall that the patience survival function at the origin is equal to $\bar{G}(0) = 1 - \mathrm{P}\{\mathrm{Blk}\}$ and $g_0 = -\bar{G}'(0)$. Then $\forall\epsilon > 0\ \exists\delta > 0$ such that, for $u \in [0, \delta]$,

$$1 - \mathrm{P}\{\mathrm{Blk}\} - (g_0+\epsilon)u \;\le\; \bar{G}(u) \;\le\; 1 - \mathrm{P}\{\mathrm{Blk}\} - (g_0-\epsilon)u\,. \tag{6.121}$$

We shall approximate J by

$$J_A \;=\; \int_0^\infty \exp\left\{-\lambda\mathrm{P}\{\mathrm{Blk}\}x - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x - \frac{\lambda g_0 x^2}{2}\right\} dx$$

Applying Lemma 4.1 with

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

we get

$$J_A \;=\; \frac{1}{\lambda\mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\mu\lambda} + \mu f(\lambda)} - \frac{g_0}{\lambda^2\mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right)$$

$$= \frac{1}{\lambda \mathrm{P}\{\mathrm{Blk}\} + (n\mu - \lambda)} - \frac{g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^3} + o\left(\frac{1}{\lambda^2}\right).$$

Now using the Laplace method and (6.121), we can prove that the same approximation is valid for

$$J = \int_0^\infty \exp\left\{\lambda \int_0^x (\bar{G}(u) - 1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx\,.$$

b. Similar to **a**. (However, here only one approximation term is needed.)

$$J_1 = \int_0^\infty x \cdot \exp\left\{\lambda \int_0^x (\bar{G}(u) - 1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\} dx \sim \int_0^\infty x \cdot \exp\left\{-\lambda \mathrm{P}\{\mathrm{Blk}\}x\right\} dx$$

$$= \frac{1}{\lambda^2 \cdot \mathrm{P}\{\mathrm{Blk}\}^2} \sim \frac{1}{n^2\mu^2\ \mathrm{P}\{\mathrm{Blk}\}^2}\,.$$

Proof of Theorem 6.6.

a. Formula (6.77) from Lemma 6.4 implies that

$$J = \frac{1}{\lambda \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{\lambda}\right). \tag{6.122}$$

Now the formula for the probability of positive offered wait follows from (3.8), (6.122) and $\lambda \sim n\mu$ $(\lambda, n \to \infty)$. Since

$$\mathrm{P}\{W > 0 | V > 0\} = 1 - \mathrm{P}\{\mathrm{Blk}\}\,,$$

formula (6.80) for the probability of actual wait prevails.

b. The conditional probability to abandon

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

$$= \frac{1 - [\beta\sqrt{\lambda\mu} + \mu f(\lambda)] \cdot \left[\frac{1}{\lambda \mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\lambda\mu} + \mu f(\lambda)} - \frac{g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^2}\right] + o\left(\frac{1}{\lambda^2}\right)}{\frac{\lambda}{\lambda \mathrm{P}\{\mathrm{Blk}\} + \beta\sqrt{\lambda\mu} + \mu f(\lambda)} - \frac{\lambda g_0}{\lambda^2 \mathrm{P}\{\mathrm{Blk}\}^2} + o\left(\frac{1}{\lambda}\right)}$$

$$= \mathrm{P}\{\mathrm{Blk}\} + \frac{1}{n} \cdot \frac{g_0}{\mu \cdot \mathrm{P}\{\mathrm{Blk}\}} + o\left(\frac{1}{n}\right).$$

Note that

$$\mathrm{P}\{\mathrm{Ab}\} = \mathrm{P}\{\mathrm{Ab}; W = 0\} + \mathrm{P}\{\mathrm{Ab}; W > 0\} = \mathrm{P}\{\mathrm{Blk}\} \cdot \mathrm{P}\{V > 0\} + \mathrm{P}\{\mathrm{Ab}|W > 0\} \cdot \mathrm{P}\{W > 0\}$$

$$= \mathrm{P}\{V > 0\} \cdot (\mathrm{P}\{\mathrm{Blk}\} + (1 - \mathrm{P}\{\mathrm{Blk}\}) \cdot \mathrm{P}\{\mathrm{Ab}|W > 0\})\,.$$

Hence,

$$\mathrm{P}\{\mathrm{Ab}|W > 0\} = (\mathrm{P}\{\mathrm{Ab}|V > 0\} - \mathrm{P}\{\mathrm{Blk}\}) \cdot \frac{1}{1 - \mathrm{P}\{\mathrm{Blk}\}}\,,$$

which implies formula (6.82). Finally, (6.83) follows from formulae (6.79) and (6.81).

c. Statement (6.84) is a consequence of (3.14). Then formula (6.79) implies (6.85).

d. First we derive (6.87).

$$\mathrm{E}[W] \;=\; \mathrm{E}[\min(V,\tau)] \;=\; \mathrm{E}[\min(V,\tau)|\tau>0]\cdot(1-\mathrm{P}\{\mathrm{Blk}\}) \;\sim\; \mathrm{E}[V]\cdot(1-\mathrm{P}\{\mathrm{Blk}\})\,,$$

where the last equivalence can be proved using the methods of Theorem 6.1, part **d**. Now formulae (6.85) and (6.80) imply (6.86) and (6.87), respectively.

Proof of Lemma 6.5.

a. The proof is similar to Part **a** of Lemma 11.1 from the main paper, and is implied by

$$J \;=\; \int_0^\infty \exp\left\{\lambda\int_0^x \bar{G}(u)du - n\mu x\right\}dx \;=\; \int_0^\infty \exp\left\{\lambda\int_0^x (\bar{G}(u)-1)du - \beta\sqrt{\lambda\mu}x - f(\lambda)\mu x\right\}dx$$

$$\sim\; \int_0^\infty \exp\left\{-(\beta+p_b)\sqrt{\lambda\mu}x - \frac{\lambda g_0 x^2}{2}\right\}dx\,, \qquad (6.123)$$

where (6.123) follows from

$$\mathrm{P}_n\{\mathrm{Blk}\} \;=\; \frac{p_b}{\sqrt{n}} \;\sim\; p_b\sqrt{\frac{\mu}{\lambda}}\,,$$

and the Laplace method.

The proof of Part **b** is identical to the proof of Lemma 6.1 (Part **b**), where β is replaced by $(\beta+p_b)$.

Proof of Theorem 6.7.

a. Direct consequence of Lemma 6.5 (parts **a** and **b**):

$$\mathrm{P}\{V>0\} \;=\; \frac{\lambda J}{\mathcal{E}+\lambda J} \;\sim\; \left[1+\sqrt{\frac{g_0}{\mu}}\right]^{-1}.$$

Since the fraction of balking customers cannot exceed the order $O\left(\frac{1}{\sqrt{n}}\right)$ we get $\mathrm{P}\{V>0\}\sim\mathrm{P}\{W>0\}$.

b. From Lemma 6.5,

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

Now note (see the proof of Theorem 6.6, Part **b**) that

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; (\mathrm{P}\{\mathrm{Ab}|\mathrm{V}>0\}-\mathrm{P}\{\mathrm{Blk}\})\cdot\frac{1}{1-\mathrm{P}\{\mathrm{Blk}\}}\,.$$

Since $1 - \mathrm{P}\{\mathrm{Blk}\} \sim 1$, the last expression is equivalent to

$$\frac{1}{\sqrt{n}} \cdot \left[\sqrt{\frac{g_0}{\mu}} \cdot h(\hat{\beta}) - \beta - p_b \right] + o\left(\frac{1}{\sqrt{n}}\right) \;=\; \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}] + o\left(\frac{1}{\sqrt{n}}\right).$$

c.

$$\mathrm{E}[V|V>0] \;=\; \frac{J_1}{J} \;\sim \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot [h(\hat{\beta}) - \hat{\beta}]\,.$$

d. As in the proof of Theorem 6.6, part **d**,

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V] \cdot (1 - \mathrm{P}\{\mathrm{Blk}\}) \;\sim\; \mathrm{E}[V]\,.$$

The equivalence between $\mathrm{E}[W|W>0]$ and $\mathrm{E}[V|V>0]$ follows from part **a**.

7 Quality-Driven operational regime

7.1 Formulation of results

Recall Theorem 7.1 from the main paper.

Theorem 7.1 (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,

$$\mathrm{P}\{\mathrm{W}>0\} \;\sim\; \frac{1}{\sqrt{2\pi n}} \cdot \frac{1}{\gamma} \cdot \left(\frac{1}{1+\gamma}\right)^{n-1} \cdot \exp\left\{\frac{\lambda\gamma}{\mu}\right\}. \tag{7.1}$$

b. Probability to abandon given wait:

$$\mathrm{P}\{\mathrm{Ab}|W>0\} \;=\; \frac{1}{n} \cdot \frac{1+\gamma}{\gamma} \cdot \frac{g_0}{\mu} + o\left(\frac{1}{n}\right) \;=\; \frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{g_0}{\mu} + o\left(\frac{1}{n}\right). \tag{7.2}$$

c. Average offered waiting time:

$$\mathrm{E}[V \mid V>0] \;=\; \frac{1}{n} \cdot \frac{1+\gamma}{\gamma} \cdot \frac{1}{\mu} + o\left(\frac{1}{n}\right) \;=\; \frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{1}{\mu} + o\left(\frac{1}{n}\right). \tag{7.3}$$

d. Average waiting time:

$$\mathrm{E}[W] \;\sim\; \mathrm{E}[V]; \quad \mathrm{E}[W \mid W>0] \;\sim\; \mathrm{E}[V \mid V>0]\,. \tag{7.4}$$

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

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;\sim\; g_0\,. \tag{7.5}$$

f. Total Service Factor:

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{t}{n} \;\middle|\; W>0 \right\} \;\sim\; e^{-(1-\rho)t}\,. \tag{7.6}$$

7.2 Proof of Theorem 7.1

In the main paper, we proved all parts of the theorem, except the last one.

f. The proof can be given along the lines of Theorem 6.1, part **g**. Sketch of the calculations is given by:

$$\frac{\mathrm{P}\left\{W > \dfrac{t}{n\mu}\right\}}{\mathrm{P}\{W>0\}} \sim \frac{\int_{t/(n\mu)}^{\infty} \exp\{-\lambda\gamma x\}dx}{\int_0^{\infty} \exp\{-\lambda\gamma x\}dx} \sim \exp\left\{\frac{-\lambda\gamma t}{n\mu}\right\} \sim e^{-\rho\gamma t} \sim e^{-(1-\rho)t}.$$

7.3 Numerical experiments

Two distributions that were already used in Subsection 6.2 are considered here again: uniform with support $[0,4]$ and hyperexponential (mixture of two exponentials with means 1 and 3). We do not use the other two distributions from Subsection 6.2, since in the QD regime our approximations are established for $g_0 > 0$ only. (Since the QD operational regime is often less important than the QED regime, we did not try to duplicate the extensive set of special cases, analyzed in Theorems 6.1-6.7.)

The experiments are performed according to guidelines in Subsection 6.2. The arrival rate λ changes from 20 to 2000 (it has been from 20 to 1000 in the QED case). The quality-driven staffing rule is

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

Four values of ρ: 0.8, 0.9, 0.95 and 0.98 were chosen. In addition, we calculate the QED regime approximation using

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

(The service grade β increases with λ and n.)

Example 1 (Figures 6,7): $\rho = 0.8$.

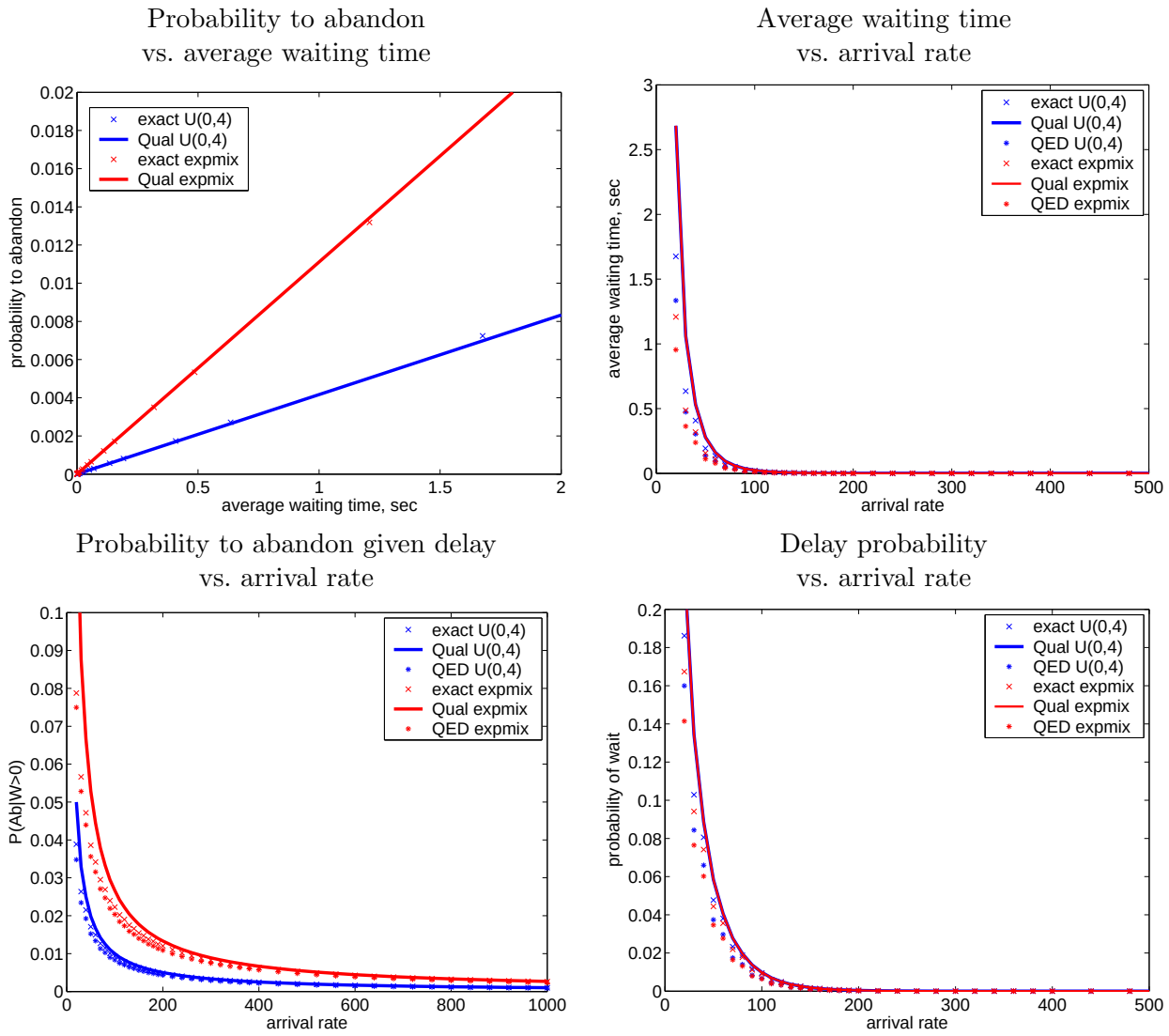
Figure 6 presents evolution of several performance measures in the format of Subsection 6.2. Figure 7 studies our approximations when performance measures take very small values.

- Here and in all other special cases of Subsection 7.3, we observe an excellent linear fit between the average wait and the probability to abandon. In general, if the offered wait is small (QD and QED regimes) and patience density at the origin is positive, a linear pattern prevails.

- The average wait and the delay probability decrease exponentially on λ. The approximation does not depend on the specific distribution. The conditional probability to abandon decreases at rate $1/n$ or, the same, $1/\lambda$. For small values of λ the exact values are somewhere between the QD and QED approximations. Figure 7 demonstrates that for large

values of λ the quality-driven approximations are excellent (and much better than the QED approximations).

Figure 6: **Offered load per server $\rho = 0.8$, performance measures and approximations**

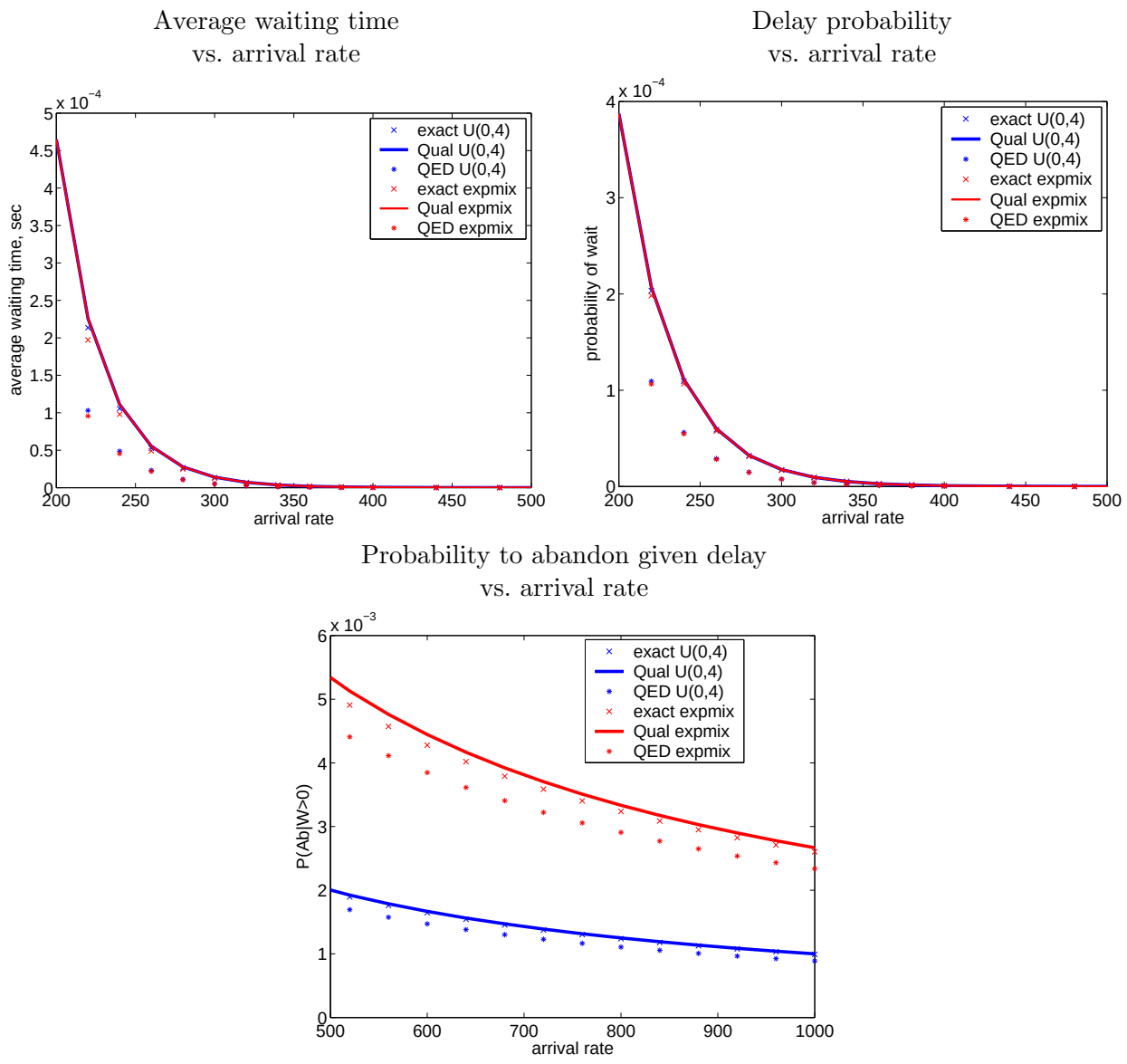


Example 2 (Figure 8): $\rho = 0.9$.

For small values of λ the QED approximations are better than the quality-driven. For larger values both types of approximations are good (and, again, one can check that the quality-driven approximations are excellent for small values).

If we consider probability-to-abandon separately, the QD approximation is better for the uniform distribution.

Figure 7: **Offered load per server $\rho = 0.8$, performance measures and approximations. Large values of arrival rate**



Example 3 (Figure 9): $\rho = 0.95$.

The quality-driven approximations are good only for $n \geq 500$ (uniform distribution) or $n \geq 1000$ (hyperexponential distribution). The QED approximations are excellent.

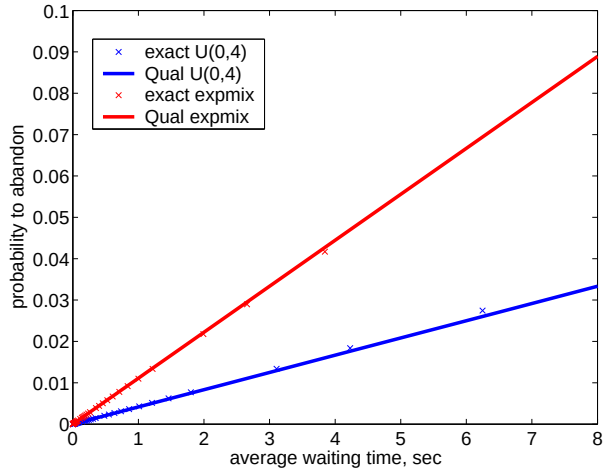
Conclusions.

It is reasonable to use the QD approximations, instead of QED, if the values of the performance measures (probabilities of wait and abandonment, average wait) are small. For a "rule of thumb",

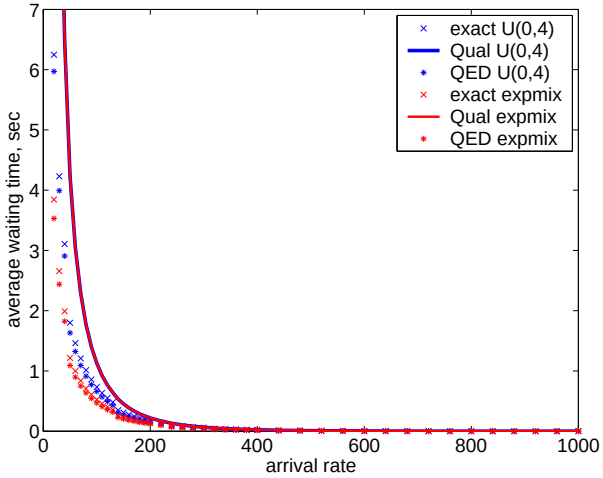
one can take the delay probability to be less than 0.1 (or even 0.05). The linear relation $\mathrm{P}\{\mathrm{Ab}\} = g_0 \cdot \mathrm{E}[W]$ prevails for all the special cases considered here. The reason is that this relation is asymptotically true both in the QD and the QED regimes, where the waiting time converges to zero.

Figure 8: **Offered load per server $\rho = 0.9$, performance measures and approximations.**

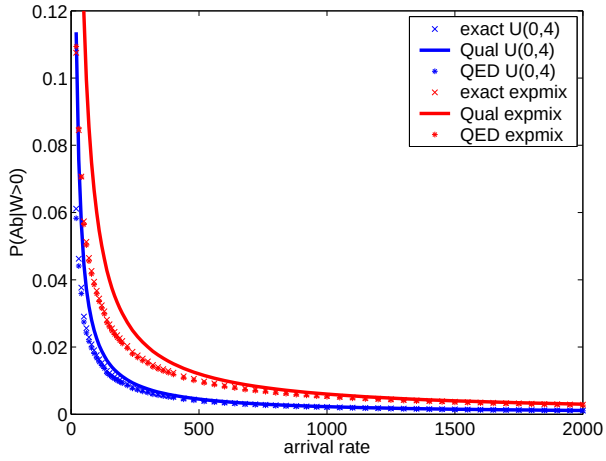
Probability to abandon
vs. average waiting time



Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate



Delay probability
vs. arrival rate

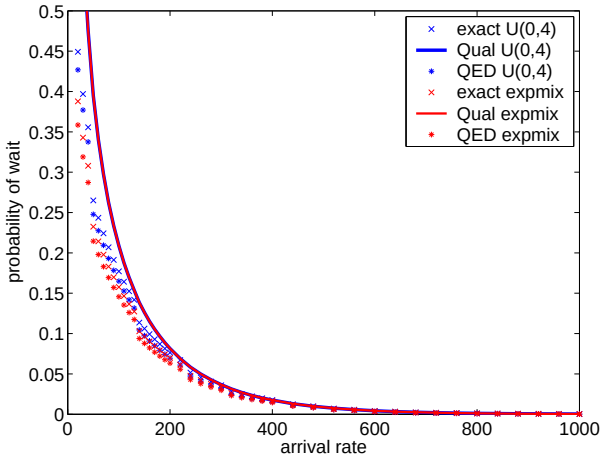
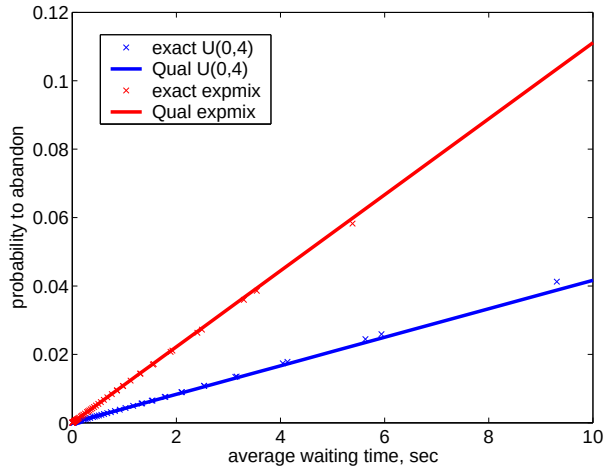
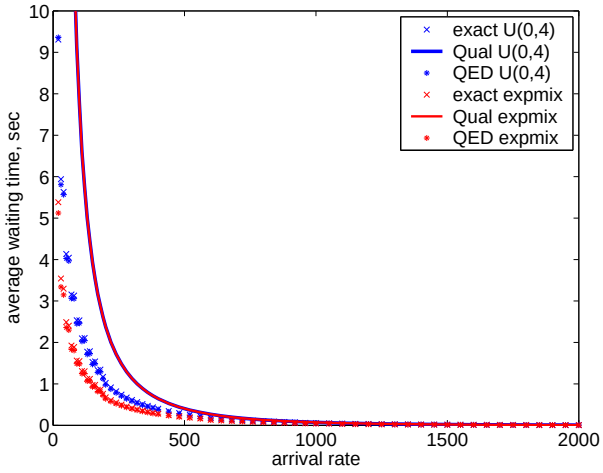


Figure 9: **Offered load per server $\rho = 0.95$, performance measures and approximations**

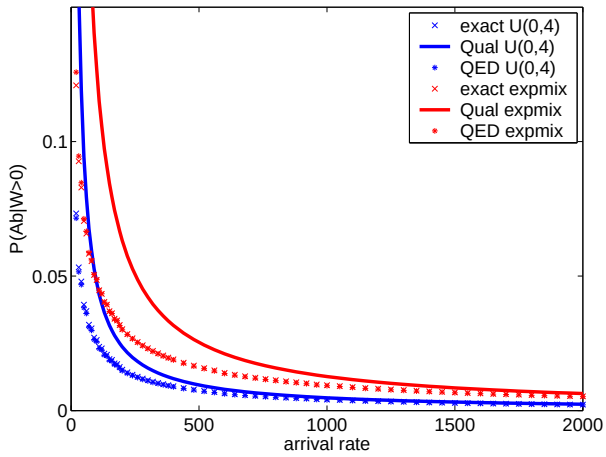
Probability to abandon
vs. average waiting time



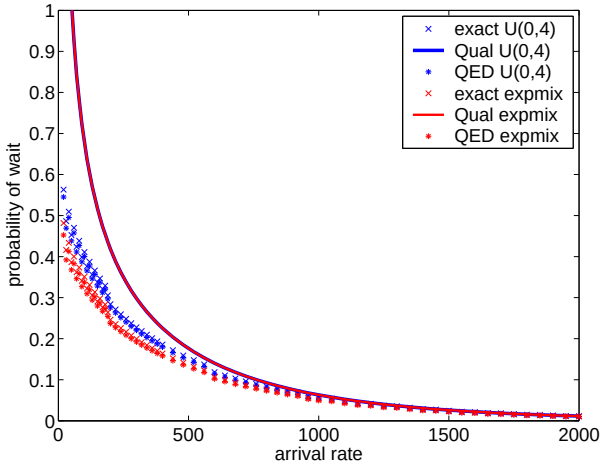
Average waiting time
vs. arrival rate



Probability to abandon given delay
vs. arrival rate



Delay probability
vs. arrival rate



8 Efficiency-Driven operational regime

8.1 Formulation of results

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

$$n = \frac{\lambda}{\mu} \cdot (1 - \gamma) + o(\sqrt{\lambda}) \,, \qquad \gamma > 0 \,. \tag{8.1}$$

The offered load per agent

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

If $o(\sqrt{\lambda}) \equiv 0$ in (8.1), the relation between ρ and γ is given by

$$\rho = \frac{1}{1-\gamma} \quad \text{and} \quad \gamma = \frac{\rho - 1}{\rho}\,. \tag{8.2}$$

Lemma 8.1 (Building blocks) Assume that the equation

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

has a unique solution x^* and that the patience density at x^* is positive: $g(x^*) > 0$. Then

a.

$$J \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}} \cdot \exp\{\lambda k(\gamma)\}\,, \tag{8.3}$$

where

$$k(\gamma) \;\stackrel{\Delta}{=}\; x^* \cdot \left(1 - \frac{n\mu}{\lambda}\right) - \int_0^{x^*} G(u)du\,. \tag{8.4}$$

(Note that the definition of x^* implies that $k(\gamma) > 0$ for λ large enough.)

b.

$$\mathcal{E} \;\sim\; \frac{1}{\gamma}\,. \tag{8.5}$$

c.

$$J_1 \;\sim\; x^* \cdot J \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}} \cdot x^* \cdot \exp\{\lambda k(\gamma)\}\,. \tag{8.6}$$

Theorem 8.1 (Performance measures) Under the assumptions of Lemma 8.1, the performance measures of the M/M/n+G queue in the efficiency-driven operational regime can be approximated by:

a. Probability to get service immediately decreases exponentially:

$$\mathrm{P}\{\mathrm{W} = 0\} \;\sim\; \frac{1}{\gamma} \cdot \sqrt{\frac{g(x^*)}{2\pi\lambda}} \cdot \exp\{-\lambda k(\gamma)\}\,. \tag{8.7}$$

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

$$\mathrm{P}\{\mathrm{Ab}\} \;\sim\; \gamma\,. \tag{8.8}$$

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

$$\mathrm{E}[V] \;\sim\; x^*\,. \tag{8.9}$$

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

$$V \;\stackrel{p}{\to}\; x^*\,. \tag{8.10}$$

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^*:

$$W \ \stackrel{w}{\to} \ G^* \, . \tag{8.11}$$

In addition,

$$\mathrm{E}[W] \ \to \ \mathrm{E}[\min(x^*, \tau)] \ = \ \int_0^{x^*} \bar{G}(u) du \, . \tag{8.12}$$

e. Total Service Factor.

The distribution of wait is given by:

$$\mathrm{P}\{V > t\} \ \sim \ \begin{cases} 1, & t < x^* \\ 0, & t > x^* \end{cases} \, . \tag{8.13}$$

$$\mathrm{P}\{W > t\} \ \sim \ \begin{cases} \bar{G}(t), & t < x^* \\ 0, & t > x^* \end{cases} \, . \tag{8.14}$$

The distribution of wait around x^* can be approximated in the following way.

Let $-\infty < t < \infty$. Then

$$\mathrm{P}\left\{ \frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \ \middle| \ V > 0 \right\} \ \sim \ \bar{\Phi}\left(t \sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) \, . \tag{8.15}$$

$$\mathrm{P}\left\{ \frac{W}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \ \middle| \ V > 0 \right\} \ \sim \ (1-\gamma) \cdot \bar{\Phi}\left(t \sqrt{\frac{g(x^*)}{\mu(1-\gamma)}} \right) \, . \tag{8.16}$$

8.2 Numerical Experiments

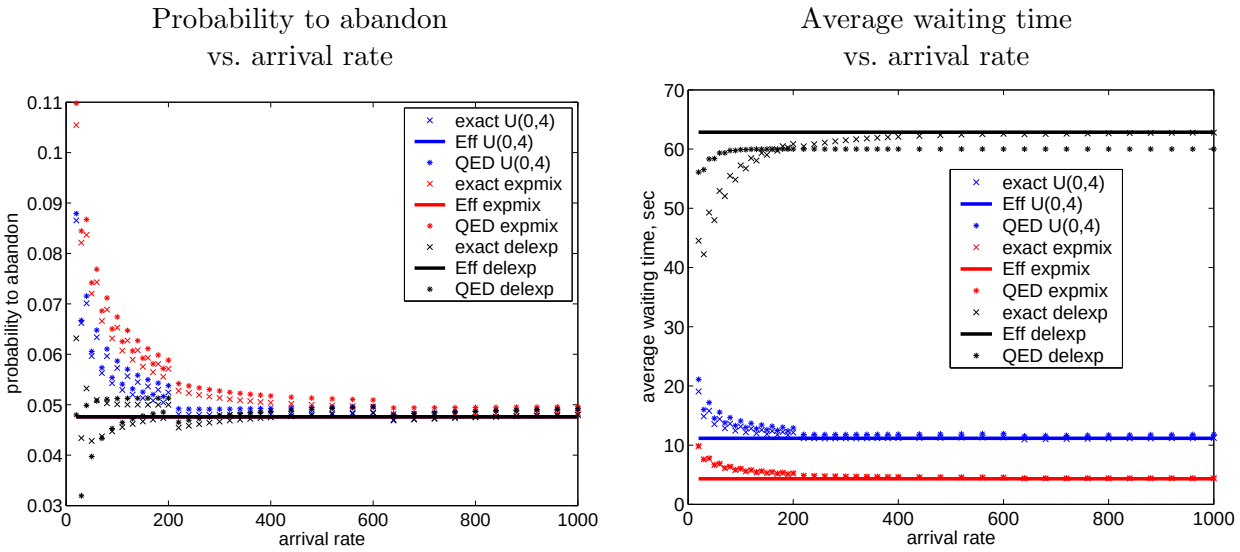
Three distributions that were considered above in Subsection 6.2 are used in our experiments: uniform, hyperexponential and delayed exponential. Instead of the conditional probability $\mathrm{P}\{\mathrm{Ab}|W > 0\}$, we plot $\mathrm{P}\{\mathrm{Ab}\}$ (the delay probability is close to one and there 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. (Although the rate of convergence of the approximations can be very different for the two types of distributions.) As in Subsection 7.3, we compare between the ED and QED approximations.

The ED staffing rule is

$$n = \left[\frac{\lambda}{\rho\mu} \right] , \qquad \rho > 1 \, . \tag{8.17}$$

Four values of ρ: 1.05, 1.1, 1.2 and 1.5 were chosen. Other assumptions are the same as in Subsection 6.2.

Figure 10: **Offered load per server $\rho = 1.05$, performance measures and approximations**



Example 1 (Figure 10): $\rho = 1.05$.

- We observe that the probability to abandon and the average wait converge to fluid limits. The limit for the probability to abandon is $\gamma = 1 - \dfrac{1}{\rho}$, independently of the patience distribution. The limit for the average wait (8.12) depends on the specific patience-time distribution.

- The QED approximations are better than the ED for small values of λ. However, we observe that QED approximations for P{Ab} and E[W] do not always converge to fluid limits (see the delayed exponential distribution for average wait).

Example 2 (Figure 11): $\rho = 1.1$.

In this case, almost all QED approximations for P{Ab} and E[W] do not converge to fluid limits (although, for very small λ they can be superior to ED.)

8.3 Proofs of the ED results

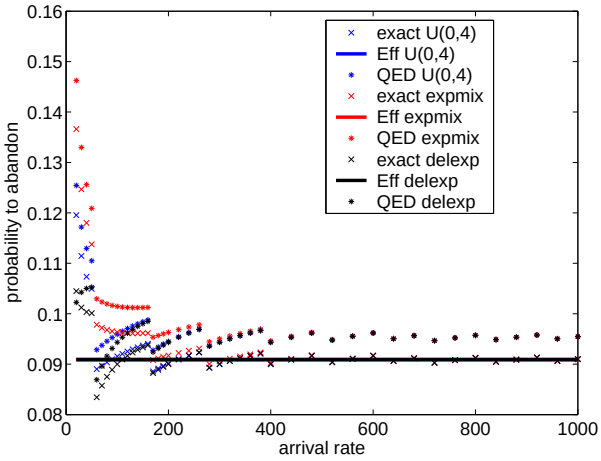
Proof of Lemma 8.1.

a. In the ED operational regime,

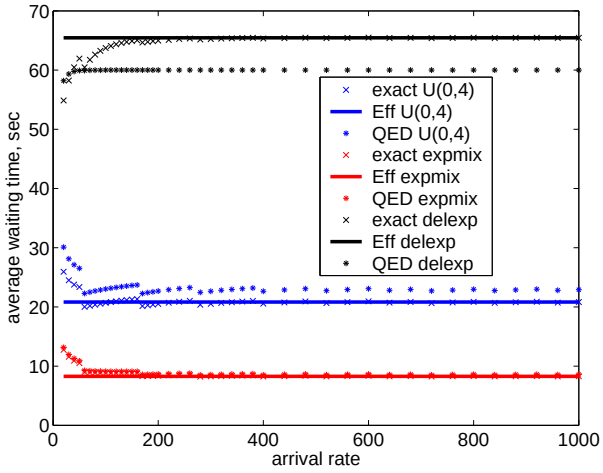
$$J \;=\; \int_0^\infty \exp\left\{\lambda\gamma x - f(\lambda)\mu x + \lambda\int_0^x [\bar{G}(u)-1]du\right\} dx\,.$$

Figure 11: **Offered load per server $\rho = 1.1$, performance measures and approximations**

Probability to abandon
vs. arrival rate



Average waiting time
vs. arrival rate



It is straightforward to verify that the function

$$h_\lambda(x) \;=\; \lambda\gamma x + \lambda\int_0^x [\bar{G}(u)-1]du$$

reaches a maximum at x^*. Changing variables: $y = x - x^*$, we get

$$J \;=\; \exp\{x^*\cdot(\lambda - n\mu)\}\cdot\int_{-x^*}^{\infty}\exp\left\{\lambda\gamma y - f(\lambda)\mu y + \lambda\int_0^{y+x^*}[\bar{G}(u)-1]du\right\}dx\,. \qquad (8.18)$$

The three leading terms of the Taylor expansion for $\int_0^{y+x^*}[\bar{G}(u)-1]du$ at $y=0$ are given by

$$\int_0^{y+x^*}[\bar{G}(u)-1]du \;=\; \int_0^{x^*}[\bar{G}(u)-1]du - \gamma y - \frac{1}{2}g(x^*)y^2 + O(y^3)\,.$$

Hence, $\forall\epsilon>0$ $\;\exists\delta>0$ such that for $|y|<\delta$

$$\int_0^{x^*}[\bar{G}(u)-1]du - \frac{1}{2}g(x^*+\epsilon)y^2 \;\le\; \gamma y + \int_0^{y+x^*}[\bar{G}(u)-1]du$$

$$\le\; \int_0^{x^*}[\bar{G}(u)-1]du - \frac{1}{2}g(x^*-\epsilon)y^2\,. \qquad (8.19)$$

Define

$$J_A \;=\; \exp\{\lambda k(\gamma)\}\cdot\int_{-\infty}^{\infty}\exp\left\{-f(\lambda)\mu y - \frac{\lambda g(x^*)y^2}{2}\right\}dy\,.$$

For $f(\lambda) = o(\sqrt{\lambda})$,

$$J_A \;\sim\; \sqrt{\frac{2\pi}{\lambda g(x^*)}}\cdot\exp\{\lambda k(\gamma)\}\,.$$

Now we perform the Laplace argument, based on inequalities (8.19), obtaining

$$J \;\sim\; J_A\,.$$

The equivalence of the $\int_{-\delta}^{\delta}$ integrals is derived via the Taylor expansion (recall Lemma 11.1 from the main paper, Part **a**). In addition, we must construct "exponential bounds" like in formula (11.10) of the main paper for

$$\int_{\delta}^{\infty} \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_{x^*}^{x^*+y} G(u)du\right\} dy \tag{8.20}$$

and the corresponding integral $\int_{-\infty}^{-\delta}$.
Define

$$\alpha \;=\; G\left(x^* + \frac{\delta}{2}\right) - \gamma \;>\; 0\,.$$

(The value of α is positive since G is strictly increasing at x^*.) Then the integral in (8.20) is less or equal to

$$\int_{\delta}^{\infty} \exp\left\{-f(\lambda)\mu y - \lambda\alpha\left(y - \frac{\delta}{2}\right)\right\} dy \;\le\; \exp\left\{\frac{\lambda\alpha\delta}{2}\right\} \cdot \int_{\delta}^{\infty} \exp\left\{-\frac{3}{4}\lambda\alpha y\right\} dy \;=$$

$$= \;\frac{\exp\{-(\lambda\alpha\delta)/4\}}{(3/4)\lambda\alpha} \;=\; o(e^{-\nu\lambda})\,, \qquad \nu > 0\,.$$

The bound for $\int_{-\infty}^{-\delta}$ is constructed in a similar way.

b. First we present an approximation

$$\mathcal{E}_A \;=\; \lambda\int_0^{\infty} \exp\left\{-\lambda\gamma x + \mu(f(\lambda)-1)x\right\} dx \;\sim\; \frac{1}{\gamma}\,. \tag{8.21}$$

Then

$$\begin{aligned}\mathcal{E} \;&=\; \lambda\int_0^{\infty} e^{-\lambda x}(1+\mu x)^{\frac{\lambda}{\mu}(1-\gamma)+f(\lambda)-1} dx \\ &=\; \lambda\int_0^{\infty} \exp\left\{-\lambda x + \left[\frac{\lambda}{\mu}(1-\gamma) + f(\lambda) - 1\right]\cdot \ln(1+\mu x)\right\} dx\end{aligned}$$

Now using $\ln(1+\mu x) = \mu x + o(x)$ and the Laplace method, we get that

$$\mathcal{E} \;\sim\; \mathcal{E}_A\,.$$

c. In the ED regime,

$$J_1 \;=\; \int_0^{\infty} x\cdot\exp\left\{\lambda\gamma x - f(\lambda)\mu x + \lambda\int_0^{x}[\bar{G}(u)-1]du\right\} dx$$

$$=\; e^{x^*\cdot(\lambda-n\mu)}\cdot\int_{-x^*}^{\infty}(y+x^*)\cdot\exp\left\{\lambda\gamma y - f(\lambda)\mu y + \lambda\int_0^{y+x^*}[\bar{G}(u)-1]du\right\} dx$$

$$= \; x^* \cdot J + e^{x^* \cdot (\lambda - n\mu)} \cdot \int_{-x^*}^{\infty} x^* \cdot \exp\left\{ \lambda\gamma y - f(\lambda)\mu y + \lambda \int_0^{y+x^*} [\bar{G}(u) - 1] du \right\} dx \,. \qquad (8.22)$$

Using Taylor expansion of $\int_0^{y+x^*} [\bar{G}(u) - 1] du$ (see part **a** of the proof), we get that the second term of (8.22) has a smaller order than the first one, given $\lambda \to \infty$.

Proof of Theorem 8.1.

a. Probability to get service immediately.

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

b. Probability to abandon.

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \frac{1 + (\lambda - n\mu)J}{\lambda J} \;\sim\; \frac{\lambda - n\mu}{\lambda} \;\sim\; \gamma \,.$$

c. Offered waiting time.

$$E[V] \;=\; \frac{J_1}{J} \;\sim\; x^* \,.$$

In order to prove $V \xrightarrow{p} x^*$, we must derive

$$\int_{x^*+\delta}^{\infty} v_\lambda(x) dx \;\to\; 0 \quad \text{and} \quad \int_0^{x^*-\delta} v_\lambda(x) dx \;\to\; 0 \,.$$

Both statements can be proved using "exponential bounds" (see the proof of (8.20)).

d. We have shown that

$$V_\lambda \xrightarrow{w} x^* \,.$$

Hence, the pair (V_λ, τ) converges weakly to (x^*, τ), as a two-dimensional random vector. Since the minimum function is continuous, the virtual waiting time

$$W_\lambda \;=\; \min(x^*, \tau) \;\xrightarrow{w}\; \min(x^*, \tau) \,.$$

In order to prove convergence of expectations, it is sufficient to demonstrate uniform integrability of $\{W_\lambda,\ \lambda \geq 0\}$. Since $W_\lambda \leq V_\lambda$, the uniform integrability can be shown for $\{V_\lambda,\ \lambda \geq 0\}$. The proof follows the pattern of proving (8.20). For example,

$$\lim_{\lambda \to \infty} \int_{x^*+\delta}^{\infty} x \tilde{v}_\lambda(x) dx \;=\; 0 \,.$$

e. Formulae (8.13) and (8.14) follow from parts **c** and **d**. The proof of (8.15) proceeds via

$$P\left\{ V > x^* + \frac{t}{\sqrt{n}} \,\middle|\, V > 0 \right\} \;=\; \frac{\int_{x^*+t/\sqrt{n}}^{\infty} \exp\left\{ \lambda\gamma x - f(\lambda)\mu x - \lambda \int_0^x G(u) du \right\} dx}{\int_0^{\infty} \exp\left\{ \lambda\gamma x - f(\lambda)\mu x - \lambda \int_0^x G(u) du \right\} dx}$$

$$= \frac{\int_{t/\sqrt{n}}^{\infty} \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_0^{y+x^*} G(u)du\right\} dy}{\int_{-x^*}^{\infty} \exp\left\{\lambda\gamma y - f(\lambda)\mu y - \lambda\int_0^{y+x^*} G(u)du\right\} dx} \sim \frac{\int_{t/\sqrt{n}}^{\infty} \exp\left\{-\frac{\lambda g(x^*)y^2}{2}\right\}}{\int_{-\infty}^{\infty} \exp\left\{-\frac{\lambda g(x^*)y^2}{2}\right\} dy} \tag{8.23}$$

$$= \bar{\Phi}\left(t\sqrt{\frac{\lambda g(x^*)}{n}}\right) = \bar{\Phi}\left(t\sqrt{\frac{g(x^*)\mu}{(1-\gamma)}}\right).$$

(The equivalence in (8.23) can be proved using the methods from Lemma 8.1, part **a**.) Then

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

Analyzing the actual waiting time,

$$\mathrm{P}\left\{\frac{W}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \,\middle|\, V > 0\right\} = \bar{G}\left(x^* + \frac{t}{\mu\sqrt{n}}\right) \cdot \mathrm{P}\left\{\frac{V}{\mathrm{E}(S)} > \frac{x^*}{\mathrm{E}(S)} + \frac{t}{\sqrt{n}} \,\middle|\, V > 0\right\}$$

$$\sim (1-\gamma) \cdot \bar{\Phi}\left(t\sqrt{\frac{g(x^*)}{\mu(1-\gamma)}}\right).$$

9 Economies of scale in the M/M/n+G queue

Consider m iid call centers that are pooled into a single operation. Each call center can be modelled by an M/M/n+G queue with the same characteristics: arrival rate λ, service rate μ, patience distribution G, and n servers, where n is determined by the manager of the call center.

Assume that all these call centers were run in one of the operational regimes studied in Sections 6-8. If we sustain that regime in the pooled call center, how will performance change? Will Economies Of Scale (EOS) drive improvements in service level? Tables 2-4 summarize answers to these questions.

In Gans et al. [5] an EOS framework for the Erlang-C queue was developed. In this section, we compare between Erlang-C and M/M/n+G, observing many similar EOS effects. That is somewhat surprising, taking into account significant differences between the two models.

9.1 QED regime

Recall that in the QED operational regime, staffing level is determined by

$$n = [R + \beta\sqrt{R}],$$

where R is the offered load. In Erlang-C , β must be positive, but in M/M/n+G, $-\infty < \beta < \infty$.

Define the *safety staffing* Δ as the difference between the staffing level n and the offered load $R = \lambda/\mu$. Again, for the Erlang-C queue we need $\Delta > 0$ to ensure system stability. In models with abandonment, Δ can become negative.

Table 2: **Economies of scale. QED regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \frac{\lambda}{\mu}$	mR	$R = \frac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	$\sqrt{m}\Delta$	$-\infty < \Delta < \infty$	$\sqrt{m}\Delta$
Number of agents	$R + \Delta$	$mR + \sqrt{m}\Delta$	$R + \Delta$	$mR + \sqrt{m}\Delta$
Service grade	$\beta = \frac{\Delta}{\sqrt{R}}$	$\boxed{\beta}$	$\beta = \frac{\Delta}{\sqrt{R}}$	$\boxed{\beta}$
P{W>0}	$\left[1 + \frac{\beta}{h(-\beta)}\right]^{-1}$	$\boxed{\text{P\{W>0\}}}$	$\left[1 + \frac{h(r\beta)}{rh(-\beta)}\right]^{-1}$	$\boxed{\text{P\{W>0\}}}$
Occupancy	$\frac{R}{R+\Delta}$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}}$	$\frac{R}{R+\Delta} \cdot (1 - \text{P\{Ab\}})$	$\frac{R}{R+\frac{\Delta}{\sqrt{m}}} \cdot \left(1 - \frac{\text{P\{Ab\}}}{\sqrt{m}}\right)$
P{Ab|W> 0}	—	—	$\frac{\beta}{\Delta r} \cdot (h(r\beta) - r\beta)$	$\frac{1}{\sqrt{m}} \cdot \text{P\{Ab\|W> 0\}}$
ASA	$\frac{1}{\Delta}$	$\frac{1}{\sqrt{m}} \cdot \text{ASA}$	$\frac{r\beta}{\Delta} \cdot (h(r\beta) - r\beta)$	$\frac{1}{\sqrt{m}} \cdot \text{ASA}$
TSF	$e^{-\beta t}$	$(\text{TSF})^{\sqrt{m}}$	$\frac{\bar{\Phi}\left(r\beta + \frac{t}{r}\right)}{\bar{\Phi}(r\beta)}$	$\text{TSF} \cdot \frac{\bar{\Phi}\left(r\beta + \frac{t}{r}\sqrt{m}\right)}{\bar{\Phi}\left(r\beta + \frac{t}{r}\right)}$

In order to get simple expressions that are straightforward to compare across regimes, we modify the definitions of average wait (ASA) and Total Service Factor (TSF). Both performance measures will be calculated only for delayed customers and they are measured in units of the average service time.

Formally,

$$\text{ASA} \triangleq \text{E}\left[\frac{W}{\text{E}(S)} \,\middle|\, W > 0\right], \tag{9.1}$$

and

$$\text{TSF} \triangleq \text{P}\left\{\frac{W}{\text{E}(S)} > \frac{t}{\sqrt{n}} \,\middle|\, W > 0\right\}$$

Definition (9.1) will be the same for the three regimes. In contrast, the definition of TSF will be modified for each special case.

Table 2 illustrates Economies of Scale for the main case of the QED regime (positive patience density at the origin). We make some changes in notation, in comparison to Theorem 6.1,

defining

$$r \stackrel{\Delta}{=} \sqrt{\frac{\mu}{g_0}}\,.$$

It will turn out that in each of the three regimes, one or several performance measures are held constant after pooling. The boxed entries in Table 2, as well as in later tables, highlight those performance measures. Indeed, in the QED case, the delay probability remains fixed under pooling. In addition, we observe that in both queues the agents' occupancy converges to 100% and ASA decreases to ASA/$\sqrt{m}$. Finally, the probability to abandon decreases at rate $1/\sqrt{m}$.

9.2 QD regime

Recall that the Quality-Driven operational regime of M/M/n+G is characterized by:

$$n \;=\; R\cdot(1+\gamma)\,,$$

where the service grade γ is positive.

The definition of TSF in the QD regime is taken to be

$$\text{TSF} \;\stackrel{\Delta}{=}\; \text{P}\left\{\frac{W}{\text{E}(S)} > \frac{t}{n}\;\middle|\; W>0\right\}\,.$$

Since abandonment is exponentially negligible in the QD regime, the M/M/n+G performance measures, presented in Table 3, are identical to the Erlang-C case: ASA decreases to ASA/m, TSF decreases to TSF^m and the delay probability converges to zero exponentially. In addition, the conditional probability to abandon in M/M/n+G decreases at rate $1/n$.

9.3 ED regime

Recall that the definitions of the ED regime for Erlang-C and M/M/n+G are different:

$$n \;=\; R+\gamma$$

for Erlang-C (see Appendix of Gurvich [8], assume that n is integer), and

$$n \;=\; R\cdot(1-\gamma)\,, \qquad \gamma>0\,,$$

for M/M/n+G.

Let

$$\text{TSF} \;\stackrel{\Delta}{=}\; \text{P}\left\{\frac{W}{\text{E}(S)} > t\;\middle|\; W>0\right\}\,.$$

The ED results are summarized in Table 4. In both queues, essentially all customers are delayed and agents are nearly 100% utilized. The waiting time remains asymptotically the same after pooling (both the mean and distribution). In addition, from Section 8 the probability to abandon in the M/M/n+G queue converges to the fluid limit $1-1/\rho$.

Table 3: **Economies of scale. QD regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \frac{\lambda}{\mu}$	mR	$R = \frac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	$m\Delta$	$\Delta > 0$	$m\Delta$
Number of agents	$n = R + \Delta$	$mR + m\Delta$	$n = R + \Delta$	$mR + m\Delta$
Service grade	$\gamma = \frac{\Delta}{R}$	$\boxed{\gamma}$	$\gamma = \frac{\Delta}{R}$	$\boxed{\gamma}$
P$\{W > 0\}$	$\frac{1}{\sqrt{2\pi n}} \cdot \frac{(\rho e^{1-\rho})^n}{1-\rho}$	$\frac{1}{\sqrt{m}} \cdot (\text{P}\{W > 0\})^m$	$\frac{1}{\sqrt{2\pi n}} \cdot \frac{(\rho e^{1-\rho})^n}{1-\rho}$	$\frac{1}{\sqrt{m}} \cdot (\text{P}\{W > 0\})^m$
Occupancy	$\frac{1}{1+\gamma}$	$\boxed{\frac{1}{1+\gamma}}$	$\frac{1}{1+\gamma}$	$\boxed{\frac{1}{1+\gamma}}$
P$\{$Ab$|$W$> 0\}$	—	—	$\frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{g_0}{\mu}$	$\frac{1}{m} \cdot \text{P}\{\text{Ab}|W > 0\}$
ASA	$\frac{1}{n} \cdot \frac{1}{1-\rho}$	$\frac{1}{m} \cdot \text{ASA}$	$\frac{1}{n} \cdot \frac{1}{1-\rho}$	$\frac{1}{m} \cdot \text{ASA}$
TSF	$e^{-(1-\rho)t}$	$(\text{TSF})^m$	$e^{-(1-\rho)t}$	$(\text{TSF})^m$

9.4 Economies of Scale: conclusions

Each operational regime corresponds to one or several performance measures that are held constant under pooling. Therefore, the following rules for the M/M/n+G queue can be deduced:

- If a call center manager would like, after pooling, to maintain agents utilization at a constant level, smaller than 100%, the QD operational regime is appropriate. It implies very high performance level, and essentially all customers get service immediately.
- If the objective is to fix the delay probability, the QED operational regime should be used. It will combine high performance level (ASA, TSF, probability to abandon) and agents' utilization that is not far from 100%.
- If it is enough to sustain the probability to abandon or/and waiting time, the understaffed ED operational regime enables this goal.

Finally, we observed close relations between EOS effects in the simple Erlang-C system and the much more complicated M/M/n+G.

Table 4: **Economies of scale. ED regime.**

	Erlang-C Queue		M/M/n+G Queue	
	Base Case	Pooled	Base Case	Pooled
Offered load	$R = \frac{\lambda}{\mu}$	mR	$R = \frac{\lambda}{\mu}$	mR
Safety staffing	$\Delta > 0$	Δ	$\Delta < 0$	$m\Delta$
Number of agents	$n = R + \Delta$	$mR + \Delta$	$n = R + \Delta$	$mR + m\Delta$
Service grade	$\gamma = \Delta$	$\boxed{\gamma}$	$\gamma = -\frac{\Delta}{R}$	$\boxed{\gamma}$
P$\{W > 0\}$	1	1	1	1
Occupancy	1	1	1	1
P{Ab}	—	—	$1 - \frac{1}{\rho}$	$\boxed{1 - \frac{1}{\rho}}$
ASA	$\frac{1}{\Delta}$	$\boxed{\text{ASA}}$	$\frac{\text{E}[\min(R, x^*)]}{\text{E}(S)}$	$\boxed{\text{ASA}}$
TSF	$e^{-t\Delta}$	$\boxed{\text{TSF}}$	$\begin{cases} \bar{G}(t/\text{E}(S)), & t < x^*/\text{E}(S) \\ 0, & t > x^*/\text{E}(S) \end{cases}$	$\boxed{\text{TSF}}$

10 Some statistical applications to call centers

10.1 General description of the data set

The source of our data is a large multi-site call center of a US bank. It has sites in New York, Pennsylvania, Rhode Island, and Massachusetts. The daily volume on a regular day is up to 300,000 calls overall. The majority of these calls end at the VRU, but up to 70,000 are seeking to reach agents. Only the latter will be considered here.

The number of agent positions at peak hours varies from 900-1200 on weekdays to 200-500 on weekends. Working hours are 24 hours a day, 7 days a week.

The call center provides many service types. In our research, we consider two of them. The first one, **Retail**, is by far the most common. The second, **Telesales**, is the most common after Retail, together with Business and Consumer Loans.

Call-by-call data was collected from March 2001 to October 2003. Our sample is taken from the five-month period between September 2002 and January 2003. Since service patterns during weekends are different, we analyze regular days only (Monday-Friday), considering calls that arrive between 7am to 24pm.

Below in Table 5 we provide overall descriptive statistics for the two service types under consideration:

Table 5: **Retail and Telesales service types. Descriptive statistics**

September 2002 - January 2003

	Calls	E[S]	P{$W > 0$}	P{Ab}	E[W]
Retail	3,451,743	224.6 sec	30.6%	1.16%	6.33 sec
Telesales	349,371	453.9 sec	24.3%	1.76%	9.66 sec

We observe that, overall, the system seems to work in the QED regime: the delay probability is neither close to zero not to one, the probability to abandon and average wait are small. However, there are huge difference between the M/M/n+G model and a large multi-site call center.

One of the most important differences lies in the protocol of customers' service. When a call arrives to the call center, it is sent to agents of a specific site. Only if a call is not served within a deadline (around 10 seconds for Retail, different numbers for other types), it can be sent to agents from other sites. This protocol violates work-conservation assumption: waiting customers and available agents can easily co-exist.

In our work with this data, we used the Data-Mocca software [13], developed in the Statistics Laboratory at the Technion.

10.2 Model primitives

In order to apply the M/M/n+G model, reliable data for its parameters should be obtained. First, we need an hourly data for λ, μ and n. Second, we must calculate estimates of patience distributions that are appropriate for our methods.

Hourly arrival and service rates are calculated from our call-by-call data. Unfortunately, detailed data on agents was not available in our database.

Patience times.

Figure 12: **Hazard rates of patience.**

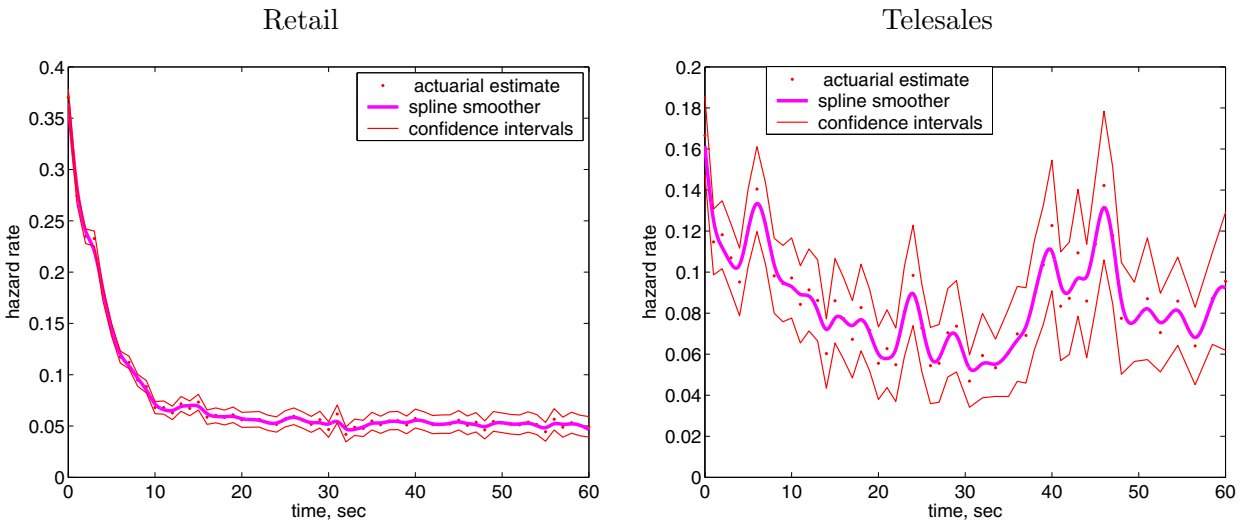
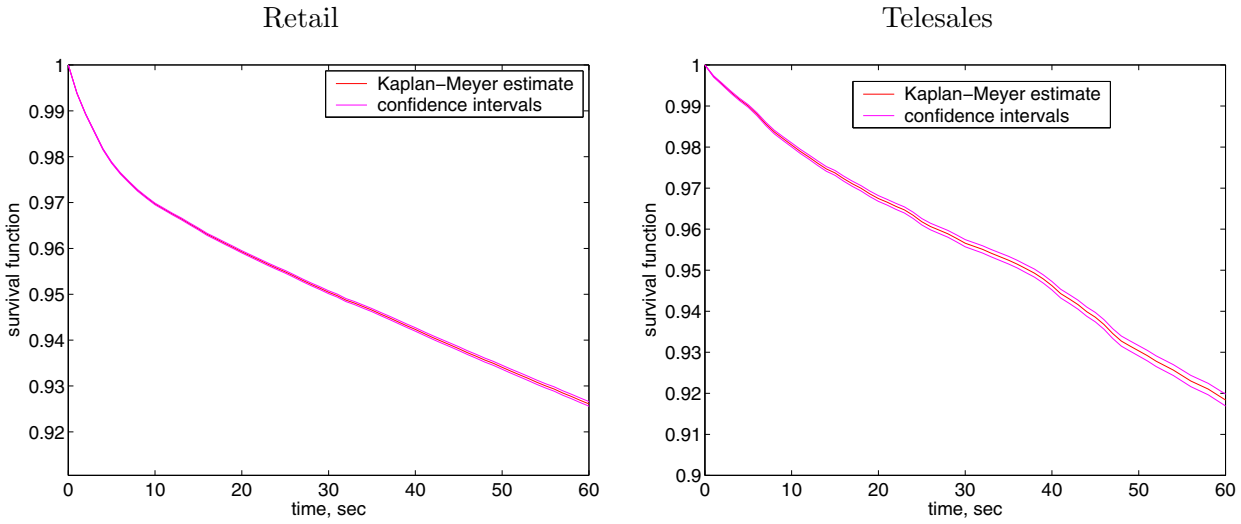


Figure 13: **Survival functions of patience. Kaplan-Meier estimate.**



We produced estimates of the patience hazard and survival-function, based on overall call-by-call data from 5 months, as mentioned before; actuarial estimator (see [3], for example), was used.

Figure 12 demonstrates very unstable hazard pattern near the origin, especially for the Retail service type. Figure 13 shows that customers are, overall, very patient: over 90% are willing to wait more than one minute.

We checked that the monthly patience patterns are indeed stable over the five-month period under consideration. However, for months out of this period, the patience hazard function can be very different. A probable reason could be changes in the contents of announcements and their timing. (Recall the second plot of Figure 2 from the main paper, where announcements took place at 15 and 60 seconds of customers' wait, implying peaks of abandonment.)

10.3 Performance measures

Three basic performance measures are considered here: $\mathrm{P}\{W > 0\}$, $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$. Below we discuss several issues related to their measurement.

Delay probability. It turns out that the database contains a very large fraction of waiting times that equal one second. (Around half of the observations! In addition, about 20% of waiting times equal zero.) Since it is unreasonable to assume that 50% of customers experienced *actual positive* wait of one second, the event $\{W = 0\}$ was defined to be equivalent to a wait of 0 or 1 second in the database.

Probability to abandon. Abandonments that took place at 0 or 1 second were discarded. Their meaning is unclear; probably they correspond to customers that decided to leave even before they were sent to queue or service (e.g. at the VRU stage).

Waiting times. Since the event "a customer was served immediately" is equivalent to the wait of 0 or 1 seconds in the database, all waiting times exceeding zero are reduced by 1 second.

10.4 Fitting QED approximations

Our main approach is the following. First, we estimate the number of agents n via fitting one of our basic performance measures (the probability to abandon was chosen since it performed the best). Specifically, we numerically solve the following equations, based on the hourly data:

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; f(\lambda, \mu, n, g_0)\,,$$

where n is unknown, f is the formula for the QED estimate from Theorem 6.1, λ and μ are hourly arrival and service rates, respectively, and g_0 is the patience density at the origin.

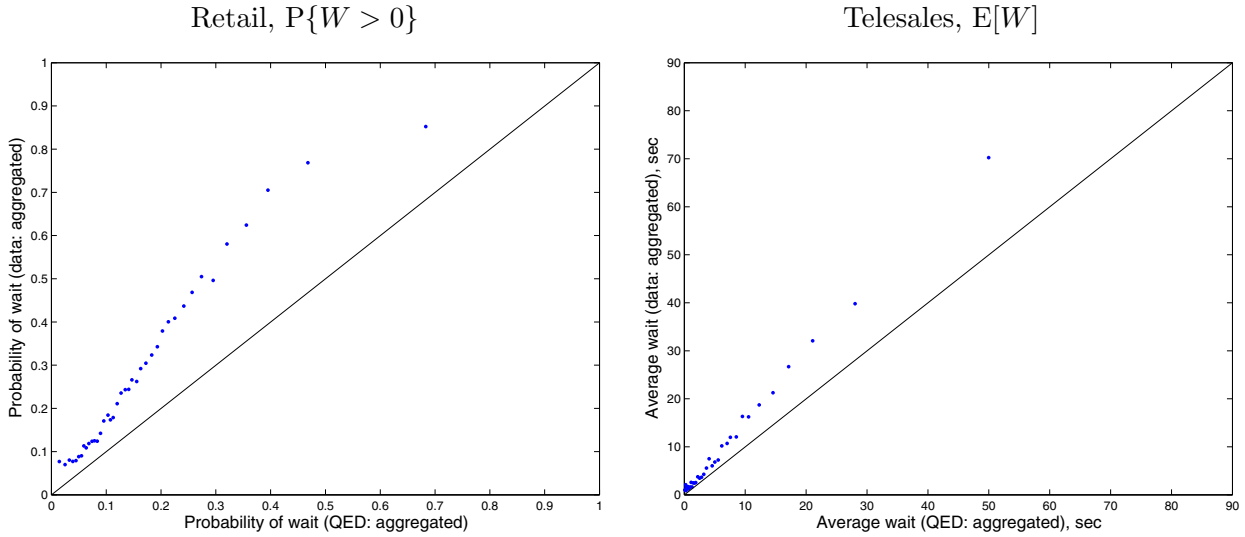
Then, using this estimate of n, we try to fit other performance measures. Since the estimate of n depends on QED formulae, such an experiment cannot "prove" that QED approximations fit the data. However, a negative result would show that some problems exist in our approach.

An additional important question arises: which value should be substituted into the QED formulae for g_0? The most straightforward way is to substitute the hazard estimate at the origin from Figure 12. Figure 14 compares the resulting QED approximations and the real-data hourly values. The results are aggregated, as in Figure 4 from the main paper.

We observe a very strong bias between data values and model values. In our opinion, the reason for the bias is instability of the two hazard estimates from Figure 12 near the origin.

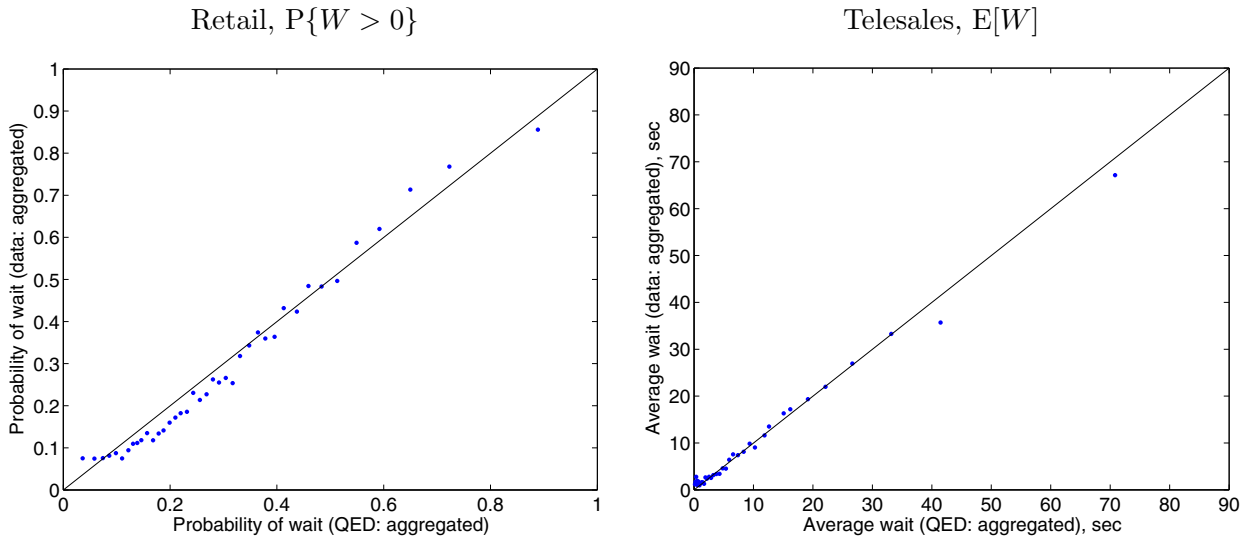
Specifically, the limit statements from Theorem 6.1 prevail in practice, if the patience density (or hazard rate) is more or less stable for typical values of waiting times. In our case, a typical wait is equal to several seconds. (In the QED limit, the wait converges to zero.) However, since the patience density oscillates significantly even within the range of several seconds, the limit QED statements do not apply directly.

Figure 14: **Fitting performance measures, g_0:=hazard at zero**



As an answer to this challenge, we suggest to substitute for g_0 the value of the ratio P{Ab}/E[W] into the QED formulae. (See Figure 15.) Now the fit for some performance measures is good. It seems that the value of the ratio gives an appropriate weighted average of the hazard rate near the origin.

Figure 15: **Fitting performance measures, g_0:= P{Ab}/E[W]**



10.5 Summary of our data analysis

In some of our experiments, we observed a good fit to the theoretical models. For example, recall Figure 4 from the main paper and Figure 15. However, several problems and challenges arise that do not enable us to characterize this data research as a definite success. The problems that deserve further attention are as follows:

- During data collection, the detailed profiles of agents should be added to call-by-call data. This will enable reliable estimates for the number of agents.
- The influence of the volatile customer behavior during the first few seconds of their wait should be explored. Probably, in addition to the approach from Subsection 10.4, one could try models with balking.
- The reality of a large modern call center is much more complicated than the M/M/n+G model. For example, due to the service protocol described in Subsection 10.2, the FCFS service discipline or work-conservation do not, in general, prevail. Hence, sometimes more complicated models should be applied.

References

[1] Asmussen S. (1987) *Applied Probability and Queues*, Wiley. 6.1.2

[2] Baccelli F. and Hebuterne G. (1981) On queues with impatient customers. In: Kylstra F.J. (Ed.), *Performance '81*. North-Holland Publishing Company, 159-179. 1, 2, 6.1.2

[3] Cox D.R. and Oakes D. (1984) Analysis of Survival Data, Chapman and Hall. 10.2

[4] Durrett R. (1991) *Probability: Theory and Examples*, Wadsworth. 5

[5] 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. 9

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

[7] Gupta P.L. and Gupta R.C. (1997) On the multivariate normal hazard. *Journal of Multivariate Analysis*, 62(1), 64-73. 5

[8] 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`. 9.3

[9] Halfin S. and Whitt W. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29, 567-588. 6.2, 6.2, 6.4

[10] Ishay E. (2003) Fitting Phase-Type Distributions to Data from a Telephone Call Center. M.Sc. Thesis, Technion. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 6.1.2, 6.1.2

[11] Jagerman D.L. (1974) Some properties of the Erlang loss function. *Bell Systems Technical Journal*, 53, 525-551. 5

[12] Jagers A.A. and Van Doorn E.A. (1986) On the continued Erlang loss function. *Operations Research Letters*, 5, 43-46. 5

[13] Trofimov V., Feigin P., Mandelbaum A. and Ishay E. (2004) DATA-MOCCA: Data Model for Call Center Analysis. Technical Report, Technion, Israel. Available at `http://iew3.technion.ac.il/serveng/References/references.html`. 10.1