OPERATIONS RESEARCH
DOI 10.1287/opre.1090.0651ec

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

E-Companion—"Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers" by Avishai Mandelbaum and Sergey Zeltyn, *Operations Research*, DOI 10.1287/opre.1090.0651.

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

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

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

March 20, 2009

Abstract

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

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

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

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

Contents

1 Structure of the Internet Appendix

In Section 2 we start with the proofs on single-interval constraint satisfaction of Theorems 4.1–4.4 from the main paper. Section 3 contains the formulation and proof of Theorem 3.1 on the global constraint satisfaction on the delay probability. Section 4 includes proofs on global constraint satisfaction from the main paper: Theorems 5.1–5.2 and Propositions 5.1–5.2. Finally, Section 5 contains an extensive numerical study.

2 Proofs for constraint satisfaction on a single interval

2.1 QED

Proof of Theorem 4.1. The first thing to verify is monotonicity in the number of servers for our performance measures and their QED approximations. These monotonicity properties are essential for the proof. First, we need to prove that the functions (11)–(13) from the main paper are strictly decreasing in β, where $P_w(\beta)$ and $W(\beta,t)$, $t>0$, decrease from one to zero and $P_a(\beta)$ decreases from infinity to zero. Second, given that λ, μ, G and $t\geq 0$ are fixed, we show that the performance measures $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{E}[W]$ and $\mathrm{P}\{W>t\}$ are decreasing in n.

Recall (see [4], for example) that the standard normal hazard rate $h_\phi(\cdot)$ is a strictly increasing convex function and $h_\phi(x)-x\downarrow 0$, as $x\to\infty$. This property implies monotonicity of (11)–(13). (Use the well-known relation $\bar{\Phi}(t)=\exp\left\{-\int_{-\infty}^{t}h_\phi(x)dx\right\}$ in order to verify the statement for (13).)

Now we check monotonicity of the performance measures. The result for $\mathrm{P}\{\mathrm{Ab}\}$ was derived in Bhattacharya and Ephmerides [1].

For $\mathrm{P}\{W>t\}$ we use M/M/n+G formulae from Zeltyn and Mandelbaum [3]. The probability not to get service immediately upon arrival is given by:

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

where

$$J \;\triangleq\; \int_0^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx\,, \tag{2.1}$$

$$\mathcal{E} \;\triangleq\; \int_0^\infty e^{-t}\left(1+\frac{t\mu}{\lambda}\right)^{n-1}dt\,, \tag{2.2}$$

and $H(x)$ was defined in (23) from the main paper. Since J decreases in n and $\mathcal{E}$ increases in n, $\mathrm{P}\{V>0\}$ decreases in n. Now note that

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

Hence, it is enough to prove monotonicity in n for

$$\mathrm{P}\{V>t|V>0\} \;=\; \frac{\int_t^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx}{\int_0^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx}\,, \tag{2.3}$$

(Formula (2.3) is derived from the list of performance measures provided in Section 9 of [3].) In other words, we must show that for $n_2>n_1$, $t>0$

$$\frac{\int_t^\infty \exp\left\{\lambda H(x)-n_2\mu x\right\}dx}{\int_t^\infty \exp\left\{\lambda H(x)-n_1\mu x\right\}dx} \;\leq\; \frac{\int_0^\infty \exp\left\{\lambda H(x)-n_2\mu x\right\}dx}{\int_0^\infty \exp\left\{\lambda H(x)-n_1\mu x\right\}dx}\,. \tag{2.4}$$

Formula (2.4) is implied by the following observation:

Lemma. Let functions f_1 and f_2 be defined on $[0,\infty)$. Assume that they are positive and f_2/f_1 decreases. Then

$$\frac{\int_t^\infty f_2(y)dy}{\int_t^\infty f_1(y)dy} \;\le\; \frac{\int_0^\infty f_2(y)dy}{\int_0^\infty f_1(y)dy}\,,$$

assuming that the right-hand side is well-defined.

The proof of the Lemma follows from the inequalities

$$\frac{\int_t^\infty f_2(y)dy}{\int_t^\infty f_1(y)dy} \;\le\; \frac{f_2(t)}{f_1(t)} \;\le\; \frac{\int_0^t f_2(y)dy}{\int_0^t f_1(y)dy}\,,$$

which are easily verified. For example,

$$\int_t^\infty f_2(y)dy \;=\; \int_t^\infty f_1(y)\cdot\frac{f_2(y)}{f_1(y)}dy \;\le\; \frac{f_2(t)}{f_1(t)}\cdot\int_t^\infty f_1(y)dy\,.$$

Finally, monotonicity in n of $\mathrm{P}\{W>t\}$ implies monotonicity of the average wait $\mathrm{E}[W]$. (The conventional stochastic order implies the same order between expectations.)

We have proved the needed monotonicity properties and now can complete the proof of Theorem 4.1. First, note that the existence of a unique solution β^* of equation (21) from the main paper follows from strict monotonicity of (11)–(13).

Then for some $\epsilon>0$ define

$$\begin{aligned} n_\lambda^1 &\triangleq \left\lceil R+(\beta^*-\epsilon)\sqrt{R}\right\rceil,\\ n_\lambda^2 &\triangleq \left\lfloor R+(\beta^*+\epsilon)\sqrt{R}\right\rfloor. \end{aligned}$$

According to relations (15)–(18) from the main paper, monotonicity of the functions (11)–(13) and the definition of β^*, we get that $U(n_\lambda^1,\lambda)\to M+\delta_1$ and $U(n_\lambda^2,\lambda)\to M-\delta_2$, for some $\delta_1,\delta_2>0$, as $\lambda\to\infty$. Therefore, for λ large enough, $U(n_\lambda^1,\lambda)>M+\delta_1/2$ and $U(n_\lambda^2,\lambda)<M-\delta_2/2$.

Since $\mathrm{P}\{\mathrm{Ab}\}$, $\mathrm{E}[W]$ and $\mathrm{P}\{W>t\}$ decrease in n, the cost function (19) decreases too. Then definition (4.2) from the main paper implies that, for λ large enough, $n_\lambda^1<n_\lambda^*\le n_\lambda^2$, which, since $\epsilon>0$ is arbitrary, proves (20) and Part **a.**

If we define asymptotically optimal staffing by

$$n_{\mathrm{QED}}^* \;\triangleq\; \left\lceil R+\beta^*\sqrt{R}\right\rceil,$$

Part **b** follows from (15)–(18) and (20) of the main paper. ■

2.2 ED

Proof of Theorem 4.2. Theorem 6.1 from Zeltyn and Mandelbaum [3] implies that in the ED operational regime:

$$n \;=\; (1-\gamma)\cdot R+o(R)\,,\qquad \gamma>0\,; \tag{2.5}$$

the probability to abandon and average wait converge to γ and $H(G^{-1}(\gamma))$, respectively. The proof of Theorem S4.2 follows from this statement in the same way that Theorem 4.1 follows from (15)–(18). ■

Remark 2.1 (On the $o(R)$ term in (2.5)) Theorem 6.1 in [3] defined the ED regime with an $o(\sqrt{R})$ term. However, the convergence proofs for $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$ remain unchanged if definition (2.5) is used. The $o(\sqrt{R})$ term is needed if one derives an exponential rate of convergence for the delay probability.

2.3 ED + QED

Proof of Theorem 4.3.
1 ⇒ 2. We shall prove this statement using a continuous indexing of M/M/$n+G$ queues by the arrival rate λ, which is more general than indexing by n. (We do this for a smoother transition into Theorem 4.4, where indexing by λ is required.) In this case, the ED + QED staffing level is given by

$$n \;=\; \bar{G}(T)\cdot\frac{\lambda}{\mu}+\delta\sqrt{\frac{\lambda}{\mu}}+f(\lambda)\,, \qquad -\infty<\delta<\infty\,,\quad f(\lambda)=o(\sqrt{\lambda})\,. \tag{2.6}$$

From Zeltyn and Mandelbaum [3],

$$\mathrm{P}\{W>T\} \;=\; \frac{\lambda\bar{G}(T)J(T)}{\mathcal{E}+\lambda J}\,. \tag{2.7}$$

where J and $\mathcal{E}$ were defined in (2.1) and (2.2), respectively, and

$$J(T) \;=\; \int_T^\infty \exp\left\{\lambda H(x)-n\mu x\right\}dx\,.$$

In order to calculate the asymptotics for $\mathrm{P}\{W>T\}$, we start with the building blocks J, $J(T)$ and $\mathcal{E}$, using the Laplace method for asymptotic calculation of integrals (see de Bruijn [2] or Sections 10 and 11 in Zeltyn and Mandelbaum [3]). We shall provide a detailed derivation of the approximation for J; the derivations for $J(T)$ and $\mathcal{E}$ are similar.

Lemma 2.1 (Approximation for J) Under the staffing level n given by (2.6), and $\lambda\to\infty$,

$$J \;\sim\; \exp\{\lambda H(T)-n\mu T\}\cdot\exp\left\{\frac{\delta^2\mu}{2g(T)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda g(T)}}\,. \tag{2.8}$$

Proof of Lemma 2.1. Performing the change-of-variables $y=x-T$ in (2.1), we get

$$J \;=\; \exp\{\lambda H(T)-n\mu T\}\cdot\int_{-T}^\infty \exp\left\{\lambda\int_T^{T+y}\bar{G}(u)du-n\mu y\right\}dy\,, \tag{2.9}$$

where, if $y<0$, $\int_T^{T+y}\triangleq-\int_{T+y}^T$. The main idea of the proof, via the Laplace method, is to replace the internal integral in (2.9) by the first two terms of the Taylor expansion:

$$\int_T^{T+y}\bar{G}(u)du \;\approx\; \bar{G}(T)y-\frac{1}{2}g(T)y^2+O(y^3)\,, \qquad (y\to 0) \tag{2.10}$$

and validate this substitution by showing that the value of the external integral in (2.9) depends mainly on the behavior of the integrand near the origin.

Specifically, define

$$J^-_{A,\epsilon} \;\triangleq\; \exp\{\lambda H(T)-n\mu T\}\cdot\int_{-T}^\infty \exp\left\{\lambda\bar{G}(T)y-\frac{1}{2}\lambda(g(T)+\epsilon)y^2-n\mu y\right\}dy \tag{2.11}$$

$$=\; \exp\{\lambda H(T)-n\mu T\}\cdot\int_{-T}^\infty \exp\left\{-\delta\sqrt{\lambda\mu}y-\frac{1}{2}\lambda(g(T)+\epsilon)y^2-f(\lambda)\mu y\right\}dy \tag{2.12}$$

$$\sim\; \exp\{\lambda H(T)-n\mu T\}\cdot\exp\left\{\frac{\delta^2\mu}{2(g(T)+\epsilon)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda(g(T)+\epsilon)}}\,, \tag{2.13}$$

where (2.12) follows from (2.6), and (2.13) is derived via straightforward calculations. Now define $J^+_{A,\epsilon}$, replacing $(g(T)+\epsilon)$ by $(g(T)-\epsilon)$ in (2.11). In addition, define J_ξ, $J^-_{A,\epsilon,\xi}$ and $J^+_{A,\epsilon,\xi}$ replacing $\int_{-T}^{\infty}$ by $\int_{-\xi}^{\xi}$, $0<\xi<T$, in (2.9), (2.11) and the corresponding expression for $J^+_{A,\epsilon}$.

Now we shall prove that $\forall \xi \in (0,T),\ \ \exists\, \nu>0$ such that

$$|J - J_\xi| = \exp\{\lambda H(T) - n\mu T\}\cdot o(e^{-\nu\lambda})\,, \tag{2.14}$$

and that the same exponential bound prevails for $|J^-_{A,\epsilon,\xi} - J^-_{A,\epsilon}|$ and $|J^+_{A,\epsilon,\xi} - J^+_{A,\epsilon}|$.

In order to prove (2.14), define

$$\zeta \ \triangleq\ \frac{\bar G(T) - \bar G(T+\xi/2)}{2}\,,$$

where $\zeta>0$ since G has positive density at T. Then, for large λ,

$$\begin{aligned}
&\int_\xi^\infty \exp\left\{\int_T^{T+y}\left(\lambda\bar G(u) - \lambda\bar G(T) - \delta\sqrt{\lambda\mu} - f(\lambda)\mu\right)du\right\}dy\\
&\quad= \int_\xi^\infty \exp\left\{\int_T^{T+\xi/2}\cdots du + \int_{T+\xi/2}^{T+y}\cdots du\right\}dy\\
&\quad\le \int_\xi^\infty \exp\left\{-(\xi/2)(\delta\sqrt{\lambda\mu}+f(\lambda)\mu) - \lambda\zeta(y-\xi/2)\right\}dy\\
&\quad= \frac{1}{\lambda\zeta}\cdot\exp\left\{-(\xi/2)(\delta\sqrt{\lambda\mu}+f(\lambda)\mu)\right\}\cdot\exp\left\{-\frac{\lambda\xi\zeta}{2}\right\} \ =\ o(e^{-\nu\lambda})\,.
\end{aligned}$$

Other exponential bounds are derived in a similar manner.

Now the Taylor expansion (2.10) of $\int_T^{T+y}\bar G(u)du$ implies that $\forall\epsilon>0,\ \ \exists\xi>0$ such that for $y\in[-\xi,\xi]$:

$$\bar G(T)y - \frac{1}{2}\left(g(T)+\frac{\epsilon}{4}\right)y^2 \ <\ \int_T^{T+y}\bar G(u)du \ <\ \bar G(T)y - \frac{1}{2}\left(g(T)-\frac{\epsilon}{4}\right)y^2\,,$$

which implies, combined with (2.9) and (2.11), that

$$J^-_{A,\epsilon/4,\xi} \ <\ J_\xi \ <\ J^+_{A,\epsilon/4,\xi}\,.$$

The exponential bounds (2.14) imply that, for large λ,

$$J^-_{A,\epsilon/2} \ <\ J \ <\ J^+_{A,\epsilon/2}\,,$$

and (2.13) implies that, for large λ,

$$\begin{aligned}
&\exp\{\lambda H(T) - n\mu T\}\cdot\exp\left\{\frac{\delta^2\mu}{2(g(T)+\epsilon)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda(g(T)+\epsilon)}} \le J \qquad (2.15)\\
&\le\ \exp\{\lambda H(T) - n\mu T\}\cdot\exp\left\{\frac{\delta^2\mu}{2(g(T)-\epsilon)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda(g(T)-\epsilon)}}\,.
\end{aligned}$$

Since ϵ is arbitrary, we proved Lemma 2.1. ■

In the same way, via the Laplace method,

$$\begin{aligned}
J(T) \ &\triangleq\ \int_T^\infty \exp\left\{\lambda\int_0^x \bar G(u)du - n\mu x\right\}dx\\
&\sim\ \exp\{\lambda H(T) - n\mu T\}\cdot\exp\left\{\frac{\delta^2\mu}{2g(T)}\right\}\cdot\sqrt{\frac{2\pi}{\lambda g(T)}}\cdot\bar\Phi\left(\delta\sqrt{\frac{\mu}{g(T)}}\right)\,. \qquad (2.16)
\end{aligned}$$

Finally, substituting $t = \lambda x$ in (2.2):

$$\mathcal{E} \;=\; \lambda \int_0^\infty e^{-\lambda x}(1+\mu x)^{n-1} dx$$

$$=\; \lambda \int_0^\infty \exp\left\{-\lambda x + \left[\frac{\lambda}{\mu}\bar{G}(T) + \delta\sqrt{\frac{\lambda}{\mu}} + f(\lambda) - 1\right] \cdot \log(1+\mu x)\right\} dx\,.$$

Using the Laplace method and the Taylor expansion for $\log(1+\mu x)$, we get that

$$\mathcal{E} \;\sim\; \frac{1}{G(T)}\,, \tag{2.17}$$

which is asymptotically negligible compared to the λJ term in the denominator of (2.7). Substitution of (2.8), (2.16) and (2.17) into (2.7) implies the following approximation for the tail probability under the staffing (2.6):

$$\mathrm{P}\{W > T\} \;\sim\; \bar{G}(T) \cdot \bar{\Phi}\left(\delta\sqrt{\frac{\mu}{g(T)}}\right). \tag{2.18}$$

Hence, if the QED parameter (2.6) is given by:

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

then

$$\mathrm{P}\{W > T\} \;\sim\; \alpha\,.$$

1 ⇒ 3. This result is a straightforward consequence of the ED results from [3]. Specifically, the probability to abandon of delayed customers is given by

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

where $J \to \infty$, as $\lambda \to \infty$, at an exponential rate. Moreover, $\mathrm{P}\{V > 0\}$ converges to 1 at an exponential rate. Now, the asymptotics for $\mathrm{P}\{\mathrm{Ab}\}$ can be derived by substituting the ED + QED staffing level in Part **1** of the theorem into the expression $\frac{\lambda - n\mu}{\lambda}$.

1 ⇒ 4. From [3], the average wait of the delayed customers in M/M/n+G is given by

$$\mathrm{E}[W|V > 0] \;=\; \frac{J_H}{J}\,, \tag{2.19}$$

where

$$J_H \;=\; \int_0^\infty H(x) \cdot \exp\{\lambda H(x) - n\mu x\}\, dx\,,$$

and J, H were defined in (2.1) and formula (23) from the main paper, respectively. Since the delay probability $\mathrm{P}\{V > 0\}$ converges to 1 at an exponential rate, the asymptotics for $\mathrm{E}[W]$ and $\mathrm{E}[W|V > 0]$ coincide. Now we shall calculate the asymptotics of (2.19) under the ED + QED staffing (2.6).

The variable-change $y = x - T$ transforms (2.19) to

$$H(T) \;+\; \frac{\displaystyle\int_{-T}^\infty \left(\int_T^{T+y} \bar{G}(u)du\right) \cdot \exp\left\{\lambda \int_T^{T+y} \bar{G}(u)du - n\mu y\right\} dy}{\displaystyle\int_{-T}^\infty \exp\left\{\lambda \int_T^{T+y} \bar{G}(u)du - n\mu y\right\} dy}\,. \tag{2.20}$$

If we substitute into (2.20) the Taylor approximation (2.10) and replace $-T$ by $-\infty$ in the integrals, we get

$$\begin{aligned} H(T) \;+\; & \frac{\int_{-\infty}^{\infty} \bar{G}(T) y \cdot \exp\left\{-\delta\sqrt{\lambda\mu}y - \frac{1}{2}\lambda g(T) y^2\right\} dy}{\int_{-\infty}^{\infty} \exp\left\{-\delta\sqrt{\lambda\mu}y - \frac{1}{2}\lambda g(T) y^2\right\} dy} \\ = \;& H(T) \;+\; \bar{G}(T) \cdot \frac{\int_{-\infty}^{\infty} \left(\frac{z}{\sqrt{\lambda g(T)}} - \delta\sqrt{\frac{\mu}{\lambda}}\frac{1}{g(T)}\right) e^{-z^2/2} dz}{\int_{-\infty}^{\infty} e^{-z^2/2} dz} \\ \sim \;& H(T) \;-\; \bar{G}(T) \cdot \delta \cdot \sqrt{\frac{\mu}{\lambda}} \cdot \frac{1}{g(T)} \\ \sim \;& \int_0^T \bar{G}(u) du \;-\; \frac{1}{\sqrt{n}} \cdot \frac{\delta}{\sqrt{1-\gamma}} \cdot \frac{1}{h_G(T)} \,. \end{aligned} \tag{2.21}$$

For a rigorous validation of the transition from (2.20) to (2.21), asymptotic equivalence of the numerators of the second terms in these formulae must be shown. This equivalence can be checked along the same lines of Lemma 2.1, and is omitted for brevity.

2, 3, 4 $\Rightarrow$ 1. These statements can be proved in the spirit of the proof of Theorem 4.1 above, using monotonicity of the corresponding performance measures. ■

Proof of Theorem 4.4. The proof of Theorem 4.4 is similar to that of Theorem 4.1 above and uses the equivalence between Statements **1** and **2** in Theorem 4.3. ■

3 Global constraint on the delay probability

Consider the following constraint on the delay probability:

$$\mathrm{P}\{W > 0\} \;\leq\; \alpha \,. \tag{3.22}$$

which is equivalent to constraint satisfaction for the performance cost functions

$$U_i(n_i, \lambda) = C_b \cdot \mathrm{P}_i\{W > 0\};$$

here $\mathrm{P}_i\{W > 0\}$ is the steady-state delay probability at interval i and $\mathrm{P}\{W > 0\} = \sum_{i=1}^{K} r_i \mathrm{P}_i\{W > 0\}$. The relation between the performance constraint α and the cost constraint M is given via $\alpha = M/C_b$.

In order to present our asymptotically optimal solution, we must solve two optimization problems. First, introduce an optimization problem on all subsets of time intervals $H \subseteq \{1, 2, \ldots, K\}$:

$$\begin{cases} \max\limits_{H} \sum\limits_{i \in H} c_i r_i \,, \\ \\ \text{s.t.} \sum\limits_{i \in H} r_i \;\leq\; \alpha \,, \end{cases} \tag{3.23}$$

which is an example of the classical "knapsack problem".

Assume that (3.23) has a unique solution H^* and define

$$\tilde{\alpha} \;\triangleq\; \alpha - \sum_{i \in H^*} r_i \,. \tag{3.24}$$

Assume, in addition, that $\tilde{\alpha}$ is positive.

Let $\bar{H}^*$ be the complement of H^*. Introduce a second nonlinear optimization problem with respect to $\bar{\beta} = \{\beta_i,\ i \in \bar{H}^*\}$:

$$\begin{cases} \min\limits_{\bar{\beta}} \sum\limits_{i\in\bar{H}^*} c_i\beta_i\sqrt{r_i}\,, \\ \\ \text{s.t.} \sum\limits_{i\in\bar{H}^*} r_i P_w(\beta_i) \;=\; \tilde{\alpha}\,, \end{cases} \tag{3.25}$$

where the function $P_w(\cdot)$ was defined in (17) from the main paper. As discussed in the beginning of Section 7.1 from the main paper, $P_w(\cdot)$ is a continuous function that decreases from one to zero. In addition, (3.24) implies that

$$\sum_{i\in\bar{H}^*} r_i \;=\; 1 - \sum_{i\in H^*} r_i \;>\; \tilde{\alpha}\,.$$

Therefore, the problem (3.25) has at least one solution. (Example 5.3 below demonstrates that there can be several solutions of (3.25).) Denote the value of the minimum in (3.25) by δ^*.

Theorem 3.1 (Global delay probability) Under the definitions and conditions presented above:
a. The optimal staffing level with respect to (3.22) satisfies

$$n_i^* \;=\; o(\sqrt{R})\,, \qquad i \in H^*\,, \tag{3.26}$$

$$n_i^* \;=\; R_i + O(\sqrt{R})\,, \qquad i \in \bar{H}^*\,, \tag{3.27}$$

$$\sum_{i\in\bar{H}^*} c_i n_i^* \;=\; \sum_{i\in\bar{H}^*} c_i R_i + \delta^*\sqrt{R} + o(\sqrt{R})\,, \tag{3.28}$$

where $R_i = (r_i\lambda)/\mu$ and $R = \sum R_i = \lambda/\mu$.

In addition, if $\{\beta_i^*,\ i \in \bar{H}^*\}$ is the unique solution of the optimization problem (3.25), then

$$n_i^* \;=\; R_i + \beta_i^*\sqrt{R_i} + o(\sqrt{R})\,, \qquad i \in \bar{H}^*\,. \tag{3.29}$$

b. Consider the staffing level $\tilde{n}^* = (\tilde{n}_1^*, \ldots, \tilde{n}_K^*)$, given by

$$\tilde{n}_i^* \;=\; 0\,, \qquad i \in H^*\,, \tag{3.30}$$

$$\tilde{n}_i^* \;=\; \left\lceil R_i + \beta_i^* \cdot \sqrt{R_i} \right\rceil\,, \qquad i \in \bar{H}^*\,; \tag{3.31}$$

here $\beta_i^*,\ \ i \in \bar{H}^*$, is some solution of (3.25). Then the staffing level (3.30)–(3.31) is asymptotically feasible and asymptotically optimal in the sense that:

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

Remark 3.1 In words, Theorem 3.1 gives rise to the following structure of the optimal solution. First, approach the constraint (3.22) as close as possible by "closing the gate" at the intervals $i \in H^*$, according to (3.30). Then satisfy the constraint by assigning QED staffing at the other intervals $i \in \bar{H}^*$, according to (3.31). This solution, although valid theoretically, is not very appropriate as a practical recommendation. For example, customers that were turned away will possibly try to get service again at subsequent time intervals.

3.1 Proof of Theorem 3.1

The proof of **a** proceeds via the following statements.

Statement 1.

$$\limsup_{\lambda\to\infty} \frac{\sum_{i=1}^{K} c_i n_i^*(\lambda) - \sum_{i\in\bar{H}^*} c_i R_i(\lambda)}{\delta^*\sqrt{R(\lambda)}} \;\le\; 1\,. \tag{3.32}$$

Statement 2. For all $i \in H^*$, $\mathrm{P}_i^{\lambda}\{W > 0\} \to 1$, as $\lambda \to \infty$.

Statement 3.

$$\liminf_{\lambda\to\infty} \frac{\sum_{i=1}^{K} c_i n_i^*(\lambda) - \sum_{i\in\bar{H}^*} c_i R_i(\lambda)}{\delta^*\sqrt{R(\lambda)}} \;\ge\; 1\,. \tag{3.33}$$

Statement 4. For all $i \in H^*$, $n_i^*(\lambda) = o(\sqrt{\lambda})$.

Note that Statements 1 and 3 imply (3.28) and Statement 4 is equivalent to (3.26).

Proof of Statement 1. Assume that $\{\beta_i^*,\ i \in \bar{H}^*\}$ is a solution (maybe not unique) of the optimization problem (3.25). Define the staffing level:

$$\tilde{n}_i(\lambda) \;=\; 0\,, \qquad i \in H^*\,, \tag{3.34}$$

$$\tilde{n}_i(\lambda) \;=\; \left\lceil R_i(\lambda) + (\beta_i^* + \epsilon_i)\cdot\sqrt{R_i(\lambda)}\right\rceil\,, \qquad i \in \bar{H}^*\,, \tag{3.35}$$

where ϵ_i are nonnegative and, at least, one of them is positive.

It follows from (15) of the main paper, (3.25) and monotonicity properties from the proof of Theorem 4.1 in the main paper that this staffing is feasible for large λ. Since ϵ_i can be arbitrarily small, feasibility of (3.34)–(3.35) implies Statement 1. ∎

Proof of Statement 2. Part a of Theorem 4.1 from Zeltyn and Mandelbaum [3] implies that if $\mathrm{P}_i\{W > 0\} \to c\,, 0 < c < 1$, then there exists finite β such that

$$n_i(\lambda) = R_i(\lambda) + \beta\sqrt{R_i(\lambda)} + o(\sqrt{\lambda})\,. \tag{3.36}$$

Since the number of time intervals is finite, there exists a sequence $\{\lambda_m\} \to \infty$, such that for $1 \le i \le K$,

$$\mathrm{P}_i^{\lambda_m}\{W > 0\} \to \xi_i\,, \qquad m \to \infty\,,$$

where $0 \le \xi_i \le 1$.

Define by F^* the set of states with $\xi_i = 1$ and by $\bar{F}^*$ its complement. Due to (3.36),

$$\liminf_{\lambda\to\infty} \frac{\sum_{i=1}^{K} n_i^*(\lambda_m)}{\sum_{i\in\bar{F}^*} R_i(\lambda_m)} \;\ge\; 1\,. \tag{3.37}$$

According to (3.32) and the optimality of $\{n_i^*\}$,

$$\sum_{i\in\bar{F}^*} c_i r_i \;\le\; \sum_{i\in\bar{H}^*} c_i r_i\,, \tag{3.38}$$

and feasibility demands

$$\sum_{i\in F^*} r_i \;\le\; \alpha\,. \tag{3.39}$$

Formulae (3.37)–(3.39), combined with the definition of $\bar{H}^*$ and uniqueness of solution of (3.23), imply $F^* = H^*$ and $\bar{F}^* = \bar{H}^*$.

Now it is straightforward to prove Statement 2. Assume that, for some $i \in H^*$, $\mathrm{P}_i^{\lambda_m}\{W > 0\} \to c \neq 1$. Define a sub-subsequence λ_{m_l} such that the delay probabilities for all intervals converge along it, define the set F^* as above and get a contradiction to $F^* = H^*$. ■

Proof of Statement 3. Assume that there exists a subsequence $\{\lambda_m\} \to \infty$ such that

$$\lim_{m\to\infty} \frac{\sum_{i=1}^K c_i n_i^*(\lambda_m) - \sum_{i\in\bar{H}^*} c_i R_i(\lambda_m)}{\delta^*\sqrt{R(\lambda_m)}} \;<\; 1\,. \tag{3.40}$$

Define the QED parameters for $i \in \bar{H}^*$ via

$$\beta_i(\lambda) \;=\; \frac{n_i^*(\lambda) - R_i(\lambda)}{\sqrt{R_i(\lambda)}}\,. \tag{3.41}$$

Then there exists a converging sub-subsequence $\{\lambda_{m_l}\}$ such that

$$\beta_i(\lambda_{m_l}) \;\to\; \tilde{\beta}_i\,, \quad \text{as } l \to \infty\,. \tag{3.42}$$

If all limits in (3.42) are finite then, combining (3.40) and (3.41),

$$\sum_{i\in\bar{H}^*} c_i\tilde{\beta}_i\sqrt{r_i} \;< \delta^*\,,$$

which combined with (3.25) contradicts feasibility.

If, for some $i \in \bar{H}^*$, $\tilde{\beta}_i = -\infty$ then $\mathrm{P}_i^{\lambda}\{W > 0\} \to 1$, which, combined with Statement 2, contradicts feasibility. If, for some $i \in \bar{H}^*$, $\tilde{\beta}_i = +\infty$ then, combining (3.40) and (3.41), there should be $\tilde{\beta}_j = -\infty$ for some $j \in \bar{H}^*$. We got contradictions in all the special cases and proved Statement 3. ■

Proof of Statement 4. Now assume that there exists $j \in H^*$, $\epsilon > 0$ and a sequence $\lambda_m \to \infty$ such that

$$n_j^*(\lambda_m) > \epsilon\sqrt{\lambda_m}\,, \qquad \epsilon > 0\,.$$

Construct converging subsequences $\beta_i(\lambda_{m_l})$ for $i \in \bar{H}^*$. Consider the staffing level

$$\begin{aligned}
\tilde{n}_i(\lambda_{m_l}) &= 0\,, \qquad i \in H^*\,,\\
\tilde{n}_i(\lambda_{m_l}) &= R_i(\lambda_{m_l}) + \beta_i\sqrt{R_i(\lambda_{m_l})} + \frac{\epsilon}{2K}\sqrt{\lambda_{m_l}}\,, \qquad i \in \bar{H}^*\,.
\end{aligned}$$

It is easy to check, via (15) from the main paper, that this staffing level is feasible for large m and

$$\sum_{i=1}^K c_i\tilde{n}_i(\lambda_{m_l}) \;<\; \sum_{i=1}^K c_i n_i^*(\lambda_{m_l})\,,$$

which contradicts the optimality of $n_i^*(\lambda_m)$. Hence, Statement 4 is proven. ■

In order to complete Part **a**, we must prove (3.27) and (3.29). The proof of (3.29) is similar to the proof of Statement 3. We should show that the QoS parameters cannot have any other limiting point, except the unique solution. In order to prove (3.27), we must show that along any subsequence of arrival rates, the QoS parameters for $i \in \bar{H}^*$ cannot converge to $\pm\infty$. The proof, again, is similar to Statement 3.

The proof of Part **b** directly follows from (15) of the main paper and Part **a**.

4 Proofs for global constraint satisfaction

4.1 Global constraint on the probability to abandon in the QED regime

Proof of Theorem 5.1. Define function $P_{ab}(\beta) = P_w(\beta)P_a(\beta), \quad -\infty < \beta < \infty$, where $P_w(\beta)$ and $P_a(\beta)$ were defined in (11) and (12) from the main paper, respectively. From the monotonicity statements in the proof of Theorem 4.1 in Section 2.1, function $P_{ab}(\beta)$ is strictly decreasing from infinity to zero. Together with continuity of $P_{ab}(\beta)$, it guarantees that the optimization problem (38) from the main paper has at least one solution.

Now we show that if $\{n_i^*(\lambda),\ 1 \le i \le K\}$ is the optimal staffing level with respect to (37) from the main paper, then

$$\limsup_{\lambda\to\infty} \frac{\sum_{i=1}^K c_i n_i^*(\lambda) - \sum_{i=1}^K c_i R_i(\lambda)}{\delta^*\sqrt{R(\lambda)}} \ \le\ 1\,. \tag{4.43}$$

Assume that $\{\beta_i^*,\ 1 \le i \le K\}$ is a solution (maybe not unique) of (38) from the main paper. Define the staffing level:

$$\tilde{n}_i(\lambda) \ =\ \left\lceil R_i(\lambda) + (\beta_i^* + \epsilon_i)\cdot\sqrt{R_i(\lambda)}\right\rceil\,, \qquad 1 \le i \le K\,, \tag{4.44}$$

where ϵ_i are nonnegative and, at least, one of them is positive.

It follows from (16), (38) and monotonicity properties in the proof of Theorem 4.1 from the main paper that this staffing is feasible for large λ. Since ϵ_i can be arbitrarily small, feasibility of (4.44), combined with definition of δ^*, imply (4.43).

Now assume that there exists a subsequence $\{\lambda_m\} \to \infty$ such that

$$\lim_{m\to\infty} \frac{\sum_{i=1}^K c_i n_i^*(\lambda_m) - \sum_{i=1}^K c_i R_i(\lambda_m)}{\delta^*\sqrt{R(\lambda_m)}} \ <\ 1\,. \tag{4.45}$$

Define the QED parameters via

$$\beta_i(\lambda) \ =\ \frac{n_i^*(\lambda) - R_i(\lambda)}{\sqrt{R_i(\lambda)}},\quad 1 \le i \le K. \tag{4.46}$$

Since the number of intervals is finite, there exists a converging sub-subsequence $\{\lambda_{m_l}\}$ such that

$$\beta_i(\lambda_{m_l}) \ \to\ \tilde{\beta}_i\,,\quad \text{as } l\to\infty,\quad 1 \le i \le K. \tag{4.47}$$

If all limits in (4.47) are finite then, combining (4.45) and (4.46), we get

$$\sum_{i=1}^K c_i\tilde{\beta}_i\sqrt{r_i} \ <\delta^*\,.$$

Now if we apply (16) from the main paper and the definition of δ^*, we get contradiction to feasibility of the staffing level $\{n_i^*(\lambda)\}$.

If, for some i, $\tilde{\beta}_i = -\infty$ then $\lim_{\lambda\to\infty} \mathrm{P}_i^\lambda\{\mathrm{Ab}\}\cdot\sqrt{\lambda} = \infty$, which, combined with (16) from the main paper, contradicts feasibility of $\{n_i^*(\lambda)\}$. If for some i, $\tilde{\beta}_i = +\infty$ then, (4.45) and (4.46) imply that there should be $\tilde{\beta}_j = -\infty$ for some interval j. So we got contradictions to (4.45) in all the special cases. Combining this conclusion with (4.43), we get (40) from the main paper.

In order to complete Part **a**, we must prove (39) and (41) from the main paper. Both proofs are very similar to the proof of (40) above. In order to prove (39), we must show that, along any subsequence of arrival rates, the QED parameters for all intervals i, $1 \le i \le K$, cannot converge to $\pm\infty$. In order to prove (41), we must show that the QED parameters $\beta_i(\lambda)$, $1 \le i \le K$, can have no other limiting point but $\{\beta_i^*,\ 1 \le i \le K\}$, the unique solution of (38) from the main paper.

The proof of Part **b** directly follows from (16) of the main paper and Part **a**. ■

4.2 Global constraint on average wait in the ED regime

Proof of Theorem 5.2. The proof is based on approximation (24) from the main paper and proceeds along the same lines as the proof of Theorem 5.1 above. Existence of solution of optimization problem (46) from the main paper follows from condition (45) in the main paper and assumptions on the distribution function G. ■

Proof of Proposition 5.1. Define function

$$u(\gamma) \triangleq \int_0^{G^{-1}(\gamma)} \bar{G}(x)dx = \int_0^{\gamma} \frac{dx}{h_G(G^{-1}(x))}, \quad 0 \le \gamma \le 1. \tag{4.48}$$

where h_G is the hazard rate of the patience distribution. (Formula (4.48) can be derived via differentiating $u(\gamma)$.) (4.48) implies that if G is DHR distribution, then u is a strictly convex function. Now we must prove that for strictly convex u, the unique solution of the following optimization problem with respect to $\bar{\gamma} = \{\gamma_i,\ 1 \le i \le K\}$

$$\begin{cases} \max\limits_{\bar{\gamma}} \sum\limits_{i=1}^{K} r_i\gamma_i\,, \\ \text{s.t.} \sum\limits_{i=1}^{K} r_i u(\gamma_i) = T\,, \\ \text{s.t. } 0 \le \gamma_i \le 1,\ 1 \le i \le K. \end{cases} \tag{4.49}$$

is given by $\gamma_i = \gamma^*$, $1 \le i \le K$, where γ^* solves $u(\gamma) = T$ with respect to γ.

Define $\gamma_1^* = \gamma_2^* = \cdots = \gamma_K^* = \gamma^*$ and assume that there exists a vector $(\gamma_1, \gamma_2, \ldots, \gamma_K)$ with $\gamma_i \ne \gamma_j$ for some $i \ne j$, that satisfies the two constraints in (4.49) and inequality

$$\sum_{i=1}^{K} r_i\gamma_i \ge \sum_{i=1}^{K} r_i\gamma_i^*\,.$$

Define $\tilde{\gamma} = \sum_{i=1}^{K} r_i\gamma_i$. Then

$$\tilde{\gamma} = \sum_{i=1}^{K} r_i\gamma_i \ge \sum_{i=1}^{K} r_i\gamma_i^* = \gamma^*\,. \tag{4.50}$$

Due to strict convexity of $u(\gamma)$ and the equality $\sum_{i=1}^{K} r_i = 1$,

$$u(\tilde{\gamma}) = u\left(\sum_{i=1}^{K} r_i\gamma_i\right) < \sum_{i=1}^{K} r_i u(\gamma_i) = T\,. \tag{4.51}$$

Since $u(\cdot)$ is an increasing function, inequality $u(\tilde{\gamma}) < T$ in (4.51) contradicts $\tilde{\gamma} \ge \gamma^*$ in (4.50). Hence, $(\gamma_1^*, \gamma_2^*, \ldots, \gamma_K^*)$ is the unique solution of (4.49). ■

Proof of Proposition 5.2. Similarly to Proposition 5.1, the IHR condition implies that the function $u(\cdot)$, defined in (4.48), is strictly concave. Now we have to prove that for strictly concave increasing $u(\gamma)$, $0 \le \gamma \le 1$, the solution of (4.49) has five properties, summarized in the formulation of Proposition 5.2 in the main paper. Note that by definition, $u(0) = 0$ and $u(1) = \bar{\tau}$. Then $u^{-1}(x)$, $0 \le x \le \bar{\tau}$ is a strictly convex increasing function with $u^{-1}(0) = 0$ and $u^{-1}(\bar{\tau}) = 1$.

Property 1. Without loss of generality, assume that we have an optimal solution $(\gamma_1^*, \ldots, \gamma_K^*)$ of (4.49) with $0 < \gamma_i^* < 1$, $1 \le i \le 2$. Define $\tilde{T} = T - \sum_{k=3}^{K} r_k u(\gamma_k^*)$. In fact, now it is enough to prove

that (γ_1^*, γ_2^*) cannot be an optimal solution of

$$\begin{cases} \max\limits_{\gamma_1,\gamma_2} \; r_1\gamma_1 + r_2\gamma_2\,, \\ \text{s.t. } r_1 u(\gamma_1) + r_2 u(\gamma_2) \;=\; \tilde{T}\,, \\ \\ \text{s.t. } 0 \le \gamma_i \le 1,\; 1 \le i \le 2. \end{cases} \tag{4.52}$$

The Lagrangian of this optimization problem is equal to

$$L(\gamma_1, \gamma_2, \xi) = \xi \cdot \left(r_1 u(\gamma_1) + r_2 u(\gamma_2) - \tilde{T} \right) - r_1\gamma_1 - r_2\gamma_2\,. \tag{4.53}$$

Differentiating (4.53), we observe that the sufficient condition for the internal extremum point $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ implies

$$\xi = 1/u'(\tilde{\gamma}_1) = 1/u'(\tilde{\gamma}_2)\,.$$

Since derivative of u is strictly monotone, $\tilde{\gamma}_1 = \tilde{\gamma}_2$, and the only "candidate" for the optimal solution in the internal point is

$$\tilde{\gamma}_1 = \tilde{\gamma}_2 = u^{-1}\left(\frac{\tilde{T}}{r_1 + r_2} \right) \tag{4.54}$$

However, it can be shown in the same way as in the Proof of Proposition 5.1 above that $(\tilde{\gamma}_1, \tilde{\gamma}_2)$ brings the objective function in (4.52) to minimum. Hence, the maximum cannot be reached in the internal point and we proved Property 1. (An alternative proof can be derived directly via definition of a concave function, similarly to the proof of Property 4 below.) ■

Property 2. follows from the equality

$$\bar{\tau} \cdot \sum_{k \in \mathcal{K}_1} r_k + r_i u(\gamma_i) = T$$

and $0 < u(\gamma_i) < \bar{\tau}$. ■

Property 3. If the solution is defined by $\mathcal{K}_1 = \mathcal{K}$ and $\mathcal{K}_2 = \{\emptyset\}$ then the value of the objective function is equal to $T/\bar{\tau}$. Moreover, any feasible solution with $\mathcal{K}_2 = \{\emptyset\}$ has the same value of the objective function. Now consider a solution with nonempty $\mathcal{K}_2 = \{i\}$. Then

$$\gamma_i \;=\; u^{-1}\left(\frac{T - \bar{\tau} \cdot \sum_{k \in \mathcal{K}_1} r_k}{r_i} \right).$$

Since $u^{-1}(\cdot)$ is strictly convex, $u^{-1}(0) = 0$ and $u^{-1}(\bar{\tau}) = 1$,

$$\gamma_i \;<\; \frac{T}{\bar{\tau} r_i} - \frac{\sum_{k \in \mathcal{K}_1} r_k}{r_i}\,. \tag{4.55}$$

Inequality (4.55) implies that $\sum_{l=1}^{K} r_l \gamma_l < T/\bar{\tau}$. Hence, there is no optimal solution with nonempty $\mathcal{K}_2$. ■

Property 4. Without loss of generality, let $\mathcal{K}_1 = \{1, \ldots, m-1\}$ and $\mathcal{K}_2 = \{m\}$, $m \le K$. We use a contradiction argument in the proof. Assume $r_{m-1} < r_m$ and

$$\bar{\tau} \cdot \left(\sum_{k=1}^{m-2} r_k + r_m \right) \;<\; T. \tag{4.56}$$

Define

$$x_m^* \triangleq \frac{T - \bar{\tau} \cdot \sum_{k=1}^{m-1} r_k}{r_m}, \quad x_{m-1}^* \triangleq \frac{T - \bar{\tau} \cdot \left(\sum_{k=1}^{m-2} r_k + r_m\right)}{r_{m-1}}. \tag{4.57}$$

Denote *Solution 1* (the optimal one) by $\gamma_1 = \cdots = \gamma_{m-1} = 1, \gamma_m = u^{-1}(x_m^*)$ and *Solution 2* by $\gamma_1 = \cdots = \gamma_{m-2} = 1, \gamma_{m-1} = u^{-1}(x_{m-1}^*), \gamma_m = 1$. Formulae (4.56) and (4.57) imply that Solution 2 satisfies the constraint conditions in the optimization problem (4.49).

Let V_1 and V_2 denote the values of the objective function from (4.49) for Solutions 1 and 2, respectively. Then

$$V_2 - V_1 = (r_m - r_{m-1}) + r_{m-1} u^{-1}(x_{m-1}^*) - r_m u^{-1}(x_m^*). \tag{4.58}$$

Note that

$$x_m^* - x_{m-1}^* = \frac{(r_m - r_{m-1}) \cdot (\bar{\tau} \cdot \sum_{k=1}^{m} r_k - T)}{r_{m-1} r_m} > 0\,,$$

and

$$\frac{x_m^* - x_{m-1}^*}{\bar{\tau} - x_{m-1}^*} = \frac{r_m - r_{m-1}}{r_m}\,.$$

Due to strict convexity of u^{-1},

$$\begin{aligned} u^{-1}(x_m^*) &< u^{-1}(x_{m-1}^*) + \frac{x_m^* - x_{m-1}^*}{\bar{\tau} - x_{m-1}^*} \cdot (u^{-1}(\bar{\tau}) - u^{-1}(x_{m-1}^*)) \\ &= u^{-1}(x_{m-1}^*) + \frac{r_m - r_{m-1}}{r_m} \cdot (1 - u^{-1}(x_{m-1}^*)). \end{aligned} \tag{4.59}$$

Combining (4.58) and (4.59), we get

$$V_2 - V_1 > 0,$$

which contradicts the optimality of Solution 1. ■

Property 5. Assume that the set $\mathcal{K}_1$ is fixed, $r_i < r_j$, and the following two relations prevail:

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

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

The values of the objective functions under $\mathcal{K}_2 = \{i\}$ and $\mathcal{K}_2 = \{j\}$ are equal to

$$V_i \triangleq \sum_{k \in \mathcal{K}_1} r_k + r_i u^{-1}\left(\frac{T - \bar{\tau} \cdot \sum_{k \in \mathcal{K}_1} r_k}{r_i}\right), \qquad V_j \triangleq \sum_{k \in \mathcal{K}_1} r_k + r_j u^{-1}\left(\frac{T - \bar{\tau} \cdot \sum_{k \in \mathcal{K}_1} r_k}{r_j}\right),$$

respectively. Now convexity of $u^{-1}(\cdot)$, $u^{-1}(0) = 0$ and $r_i < r_j$ imply $V_i > V_j$ and, hence, Property 5. ■

5 Numerical experiments

In this section, we validate the quality of our asymptotic solutions developed in Sections 4 and 5 from the main paper. We cover the single-interval QED, ED and ED + QED regimes in Sections 5.1–5.3, respectively. Our general approach is to compare the asymptotic staffing levels, derived in Theorems 4.1, 4.2 and 4.4 from the main paper, with the exact optima calculated via formulae in [3]. Finally, in Section 5.4 we provide some experiments on the global constraints.

5.1 QED approximations

In the QED regime, we perform an extensive numerical experiment within the following framework. Let the service rate μ=1. (In other words, the average service time is our time unit.) We consider arrival rates that vary from 10 to 100 by step 10, from 100 to 400 by step 20 and from 400 to 1,000 by step 40, for a total of 40 values of λ. In addition, six patience distributions are chosen:

- Two exponential distributions with means 2 and 0.5;
- Two uniform distributions on [0,4] and [0,1];
- Two hyperexponential distributions, both being a 50–50% mixture of two exponentials. The exponential means are 1 and 3 in the first case (mean patience equals to 2), and 1/4 and 3/4 in the second case (mean patience 1/2).

Note that we consider three types of distributions and, for each type, choose two representatives: the first one with average patience longer than the average service time, and the second one with shorter patience.

For each combination of λ, patience distribution and a specific form of disutility function, we compare between the exact optimal and asymptotically optimal staffing levels n^* and $n^*_{\rm QED}$.

Constraint on the delay probability.
In Theorem 4.1 from the main paper assume the cost coefficients $C_{ab} = 0$, $C_w = 0$, $C_b = 1$ and the deadline $t = 0$. In this case, the constraint on the performance cost function is equivalent to

$$\mathrm{P}\{W > 0\} \;\le\; M\,.$$

We take values of M equal to 0.1–0.9 by step 0.1. Combining these values with the above-mentioned arrival rates and patience distributions, we get $40 \times 6 \times 9 = 2160$ special cases. The optimal staffing levels n^* vary from 5 to 1,045. For all these values, the asymptotically optimal $n^*_{\rm QED}$ deviates from n^* by no more than 2. Specifically, $n^*_{\rm QED}$ matches n^* exactly in 57% of the cases; it deviates from n^* by 1 in 40% of the cases; and by 2 in the remaining 3%.

The exact optimal staffing levels are always larger or equal to the approximate ones and the differences become smaller as M decreases. The fit for longer patience is, overall, better.

Constraint on the probability to abandon.
In Theorem 4.1 from the main paper assume $C_{ab} = 1$, $C_w = 0$, $C_b = 0$. Then the constraint is equivalent to

$$\mathrm{P}\{\mathrm{Ab}\} \;\le\; M/\sqrt{\lambda}\,.$$

We consider values of M that vary from $0.03 \cdot \sqrt{10}$ to $0.3 \cdot \sqrt{10}$ by step $0.03 \cdot \sqrt{10}$. (For example, $M = 0.03 \cdot \sqrt{10}$ corresponds to 3% abandonment for $\lambda = 10$ and 0.3% for $\lambda = 1000$.)

Out of 2,400 experiments, we observe a perfect fit in 2,242 (93%). In all other cases, except for one, the difference is 1 (either negative or positive). For large negative QoS parameters (that correspond to large values of M), the staffing levels for the different distributions are very close to each other. (See Section 5.2 for an explanation of this phenomenon.)

Constraint on average wait.
In Theorem 4.1 from the main paper assume $C_{ab} = 0$, $C_w = 1$, $C_b = 0$. Then the constraint condition is given by

$$\mathrm{E}_{n,\lambda}[W] \;\le\; M/\sqrt{\lambda}\,.$$

For the distributions with mean patience equal to 2, values of M are chosen equal to $(1/12) \cdot \sqrt{10}, \ldots, \sqrt{10}$ by step $(1/12) \cdot \sqrt{10}$. (For example, $M = (1/3) \cdot \sqrt{10}$ implies maximal average wait of 1/3 average service times for $\lambda = 10$ and 1/30 average service times for $\lambda = 1000$.) For the three distributions with smaller patience (1/2), M varies from $(1/48) \cdot \sqrt{10}$ to $(1/4) \cdot \sqrt{10}$ by step $(1/48) \cdot \sqrt{10}$.

The fit is excellent again. Out of 2,880 experiments, a perfect fit in observed in 2,111 (73%) cases. Otherwise, the difference is equal to 1, except for a single case when it equals 2. We observe considerable staffing differences between distributions if relatively large wait is acceptable (large M).

Conclusions. The QED approximations are superb for an extensive set of parameters and patience distributions. We now proceed to check how these approximations work for very extreme values that arise if staffing is performed according to the other two operational regimes.

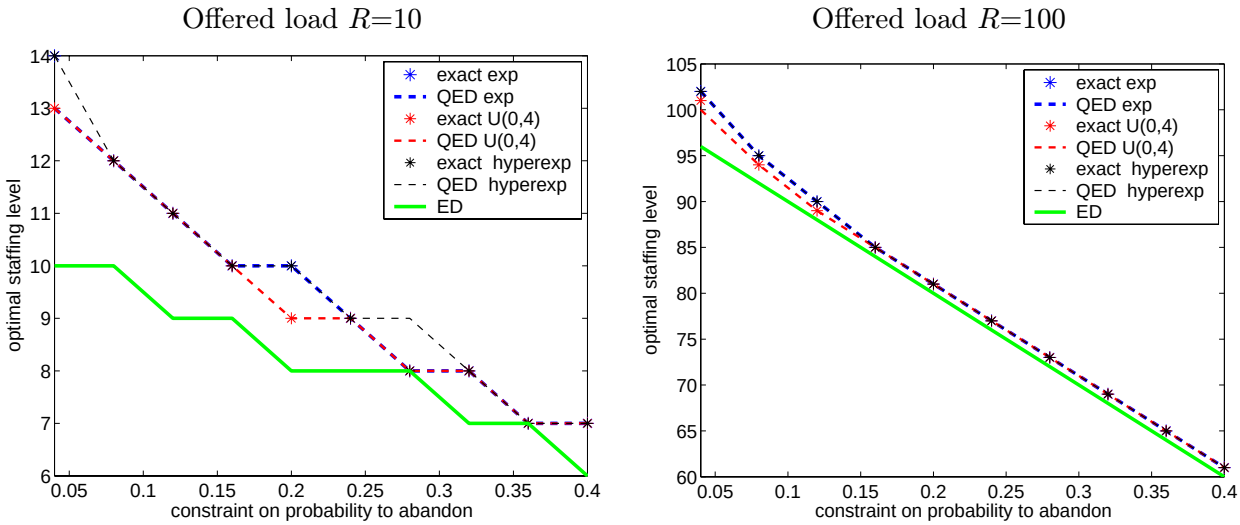
5.2 ED approximations

Here we check the quality of the approximations in Theorem 4.2 from the main paper, comparing n^* and $n^*_{\rm ED}$. In addition, for each numerical experiment we calculated an asymptotically optimal QED staffing level. Specifically, we assigned to the cost coefficients in Theorem 4.1 from the main paper the values $C_{ab} = 1/\sqrt{\lambda}$, $C_w = 1/\sqrt{\lambda}$ and $C_b = 0$.

Constraint on the probability to abandon. The constraints on P{Ab} are varied from 0.04 to 0.4 by step 0.04. We consider the 3 patience distributions from Section 5.1 with mean 1/2. (Results for the distributions with mean 2 are similar, but the differences between n^*, $n^*_{\rm ED}$ and $n^*_{\rm QED}$ are somewhat smaller.)

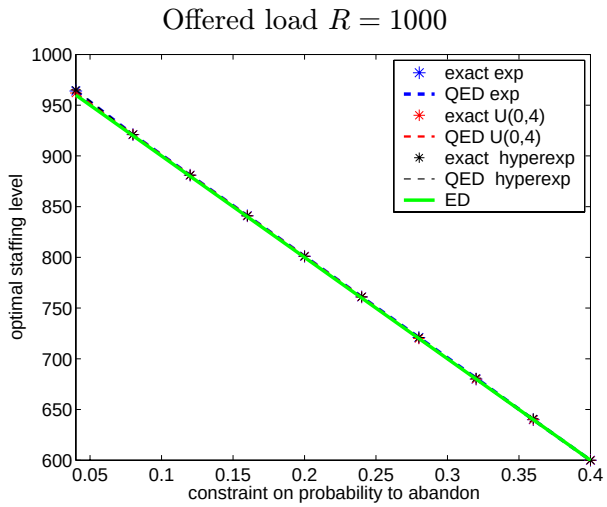
We consider three values of the offered load $R = \lambda/\mu$: 10 (small), 100 (moderate) and 1,000 (large). The results are displayed in Figures 1 and 2. Stars are for the exact optima, dashed lines are for the QED approximations and the solid line is for the ED approximation. (The ED approximation for P{Ab} does not depend on the patience distribution.)

Figure 1: **Constraint on the probability to abandon**



Even for relatively small values of R (10 or 100), we observe an almost perfect fit between the exact optimal staffing and its QED approximation. For small values of the constraint (0.05–0.2), ED estimates imply significant understaffing.

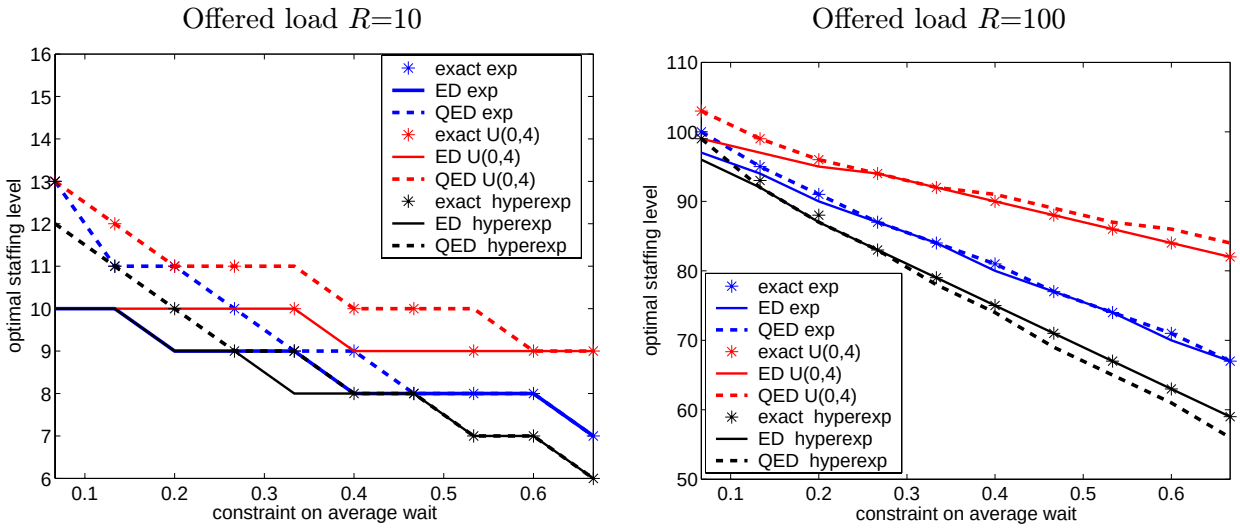
Figure 2: **Constraint on the probability to abandon**



In the case $R = 1000$, all the three types of lines merge into a single line. Both approximations are excellent.

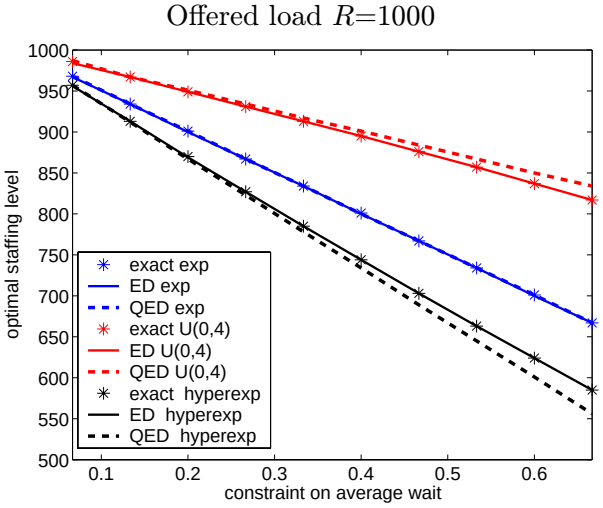
Constraint on average wait. Constraints on E[W] are varied from 1/15 to 2/3 by step 1/15. The three patience distribution from Section 5.1, with mean 2, are considered; offered load is again chosen to be R=10, 100, and 1000. (The three distributions with mean 1/2 provide very similar results.)

Figure 3: **Constraint on average wait**



If $R = 10$ (Figure 3) the fit of the QED approximations is excellent; ED is a bit worse for small values of constraint. However, for $R = 100$ we observe a bias in the QED estimates for the case of nonexponential patience distributions (2–3 at the largest values of constraint).

Figure 4: **Constraint on average wait**



Finally, for $R = 1000$ the QED bias can be very significant (20–30 for the largest value of constraint). The ED approximation, on the other hand, is almost perfect.

Constraint on the probability to abandon: conclusions and discussion. The results for the ED approximations are to be expected: the fit is not that good for small (strict) values of the constraints. It improves for looser (larger) constraints and for larger arrival rates. The results for the QED approximations are astonishing. In addition to the experiments presented in Figures 1 and 2, we performed an extensive numerical study with the 6 patience distributions and 40 values of the offered load from Section 5.1, and 10 values of constraints. Exact fit has been observed in 2,346/2,400 (98%) cases, otherwise the difference is 1. Hence, QED estimates turn out to be excellent for all the special cases considered in Sections 5.1 and 5.2.

Why do we have a perfect fit for P{Ab}? Recall the QED approximation of the probability to abandon:

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

For large negative β, the normal hazard rate $h_\phi(\cdot)$ is negligible. Using the definition of $\hat{\beta}$ (formula (14) from the main paper), we can easily deduce that the QED approximation is then close to $(R-n)/R$, namely the ED approximation.

Constraint on average wait: conclusions and discussion. The conclusions for the ED approximations of $\mathrm{E}[W]$ are similar to the conclusions for P{Ab}: they work very well for large arrival rates (hundreds and more) and relatively loose values of constraints. However, the situation with the QED approximations is different.

The QED approximations provide an excellent fit for the exponential distribution. This can be explained by the high quality of P{Ab} approximations and the relation $P\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W]$, which prevails for both exact values and QED approximations. (Note that the exponential parameter $\theta = g_0$.)

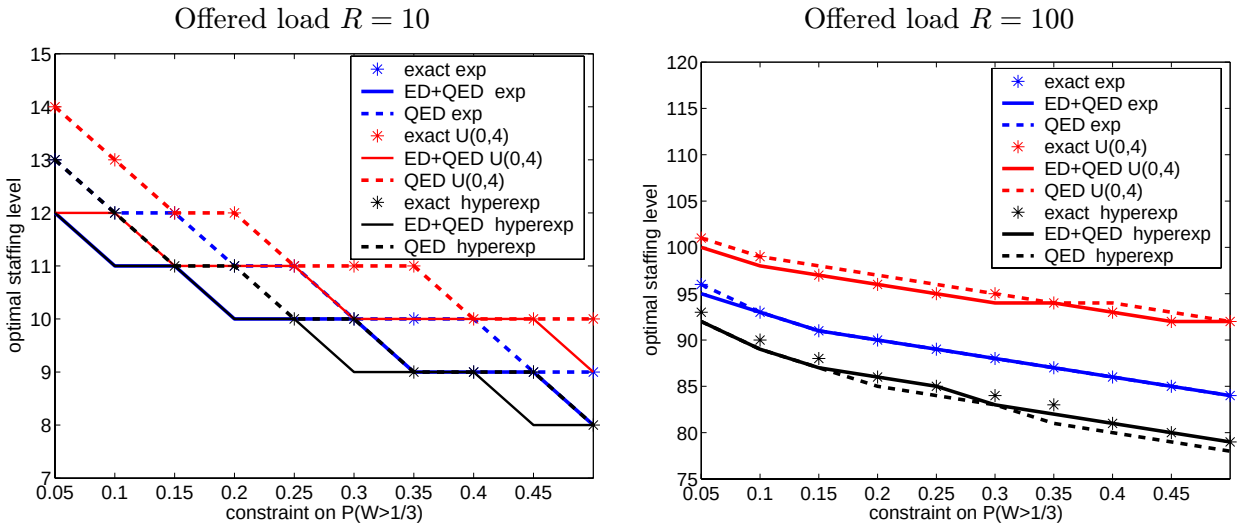
However, if λ and the constraint values are large, we observe a strong bias of the QED estimates for nonexponential distributions. Note that the relation $P\{\text{Ab}\} = g_0 \cdot \text{E}[W]$ prevails for the QED approximations but does not for ED, and the latter is very close to the exact values in this case. Therefore, the excellent QED fit for P{Ab} is not always retained for E[W].

5.3 ED + QED approximation for the tail probability

Now we check the quality of the ED + QED approximation from Theorem 4.4 of the main paper. Consider the three distributions from Section 5.2 with patience mean 2 and take the deadline equal to 1/3. Constraints on the probability to exceed this deadline are varied from 0.05 to 0.5 by step 0.05. Finally, three values of the arrival rate (offered load) are considered: 10, 100, and 1,000.

We compare, in Figures 5 and 6, these approximation with exact optimal staffing and the QED approximation derived via Theorem 4.1 from the main paper.

Figure 5: **Constraint on the tail probability**



From (33) in the main paper, the approximate ED + QED staffing level is equal to

$$n^*_{\text{ED}+\text{QED}} \;=\; \left\lceil (1-\gamma^*)\cdot R + \delta^*\sqrt{R} \right\rceil ,$$

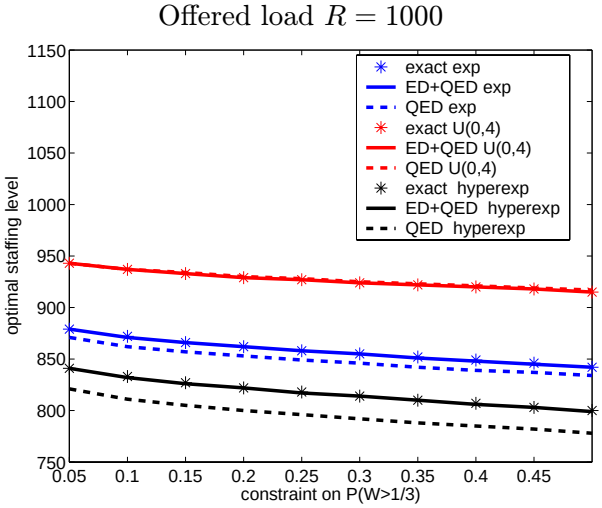
where $\gamma^* = G(T)$ and δ^* is defined by (4.28).

Note that for every distribution under consideration, the ED coefficient γ^* is constant for all experiments. Specifically, it is 0.154 for the exponential distribution, 0.083 for uniform and 0.194 for hyperexponential. (That explains why the uniform distribution requires a larger staffing level.) The QED coefficient δ^* provides fine tuning for different values of constraints.

In the case $R = 10$ (Figure 5) we observe a nearly perfect fit for QED staffing, which is preferable over ED + QED. If $R = 100$ (moderate number of agents) the fit for ED + QED is already slightly better. Finally, if $R = 1000$ (Figure 6) the fit for ED+QED is fine, while the QED estimate is strongly biased for exponential and hyperexponential distributions (6–10 and 15–23 servers, respectively). The bias takes place for very large negative QoS parameters ($-5 \cdots -7$). For the uniform distribution the QoS parameters are smaller ($-2 \cdots -3$), so the fit is considerably better.

The main conclusion, therefore, is that the ED + QED approximation is preferable over QED for moderate-to-large call centers and moderate-to-loose constraints on the service level.

Figure 6: **Constraint on the tail probability**



5.4 Experiments on global constraint

Global constraint on the delay probability. We present several examples where the QED regime is asymptotically optimal at all intervals. The Lagrange multipliers are used in order to solve the optimization problem (3.25). Fix the service rate $\mu = 1$.

Example 5.1 Consider two time-intervals, assume overall arrival rate $\lambda = 100$, $r_1 = 0.7$, $r_2 = 0.3$, the constraint $\alpha = 0.2$ (80% of customers get service immediately) and the six patience distributions considered in Section 5.1. Assume that the staffing costs in both intervals are equal: $\bar{c} = (1, 1)$. Below we compare exact optimal staffing $\bar{n}^*$ and the approximation $\bar{n}^*_{\rm QED}$ derived via Theorem 3.1. We observe an excellent fit between $\bar{n}^*$ and $\bar{n}^*_{\rm QED}$.

Patience distribution	I	II	III	IV	V	VI
$\bar{n}^*$	(80,34)	(80,35)	(79,34)	(79,32)	(79,34)	(78,32)
$\bar{n}^*_{\rm QED}$	(80,34)	(80,34)	(79,34)	(79,32)	(79,33)	(78,31)

Example 5.2 Now assume staffing costs $\bar{c} = (1, 1.8)$ and retain the other settings from Example 5.1. Below are the results of this numerical experiment.

Patience distribution	I	II	III	IV	V	VI
$\bar{n}^*$	(82,32)	(82,33)	(80,33)	(81,30)	(82,31)	(80,30)
$\bar{n}^*_{\rm QED}$	(82,32)	(82,33)	(82,31)	(83,29)	(82,31)	(83,28)

The two optima are close again. Some differences can be due to the fact that $\bar{n}^*_{\rm QED}$ is derived by rounding a continuous solution of (3.25) that can be somewhat different from the optimal solution in integers. Comparing with Example 5.1, there are more agents in the first interval and less in the second one due to the change in staffing costs.

Is the solution of (3.25) always unique? In fact, the uniqueness depends on the convexity of the curve $r_1 P_w(\beta_1) + r_2 P_w(\beta_2) = \tilde{\alpha}$. From [3] we deduce that a counterexample can be found if we

consider large values of $\tilde{\alpha}$ and impatient customers (large g_0). The following example illustrates this point.

Example 5.3 Let $g_0 = 10, k = 2, r_1 = r_2 = 0.5$ and $\tilde{\alpha} = 0.45$. Then $\beta_1^* = -3.6671, \beta_2^* = 1.2389$ and $\beta_1^* = 1.2389, \beta_2^* = -3.6671$ are two global minimums of (3.25). The point $\beta_1^* = \beta_2^* = -1.0381$ is a local maximum.

Overall, our numerical experiments show that, unless $\tilde{\alpha}$ is large and customers are very impatient (as in Example 5.3), the solution of (3.25) is unique and larger QoS parameters correspond to larger arrival rates r_i. The following example illustrates this phenomenon.

Example 5.4 Let $g_0 = 0.5, k = 5, \bar{r} = (0.25, 0.15, 0.25, 0.15, 0.2)$ and $\tilde{\alpha} = 0.1$. The the optimal solution to (3.25) is $\bar{\beta}^* = (1.4132, 1.2306, 1.4132, 1.2306, 1.3363)$.

Global constraint on the probability to abandon. We have already presented a numerical experiment in Section 2.5 from the main paper. The following two small experiments study the fit between approximate and exact optimal staffing levels. Consider the constraint "P{Ab} $\leq 3\%$" and retain other system parameters from Examples 5.1 and 5.2.

Example 5.5 Assume equal staffing costs: $\bar{c} = (1, 1)$.

Patience distribution	I	II	III	IV	V	VI
$\bar{n}^*$	(73,31)	(71,31)	(73,32)	(75,34)	(74,33)	(76,34)
$\bar{n}^*_{\text{QED}}$	(73,32)	(71,31)	(73,32)	(75,34)	(74,33)	(76,34)

Example 5.6 Assume staffing costs $\bar{c} = (1, 1.8)$.

Patience distribution	I	II	III	IV	V	VI
$\bar{n}^*$	(73,31)	(72,30)	(74,31)	(77,32)	(76,31)	(78,32)
$\bar{n}^*_{\text{QED}}$	(75,31)	(73,30)	(75,31)	(78,32)	(76,31)	(78,32)

In both experiments we observe a very good fit between $\bar{n}^*$ and $\bar{n}^*_{\text{QED}}$.

References

[1] Bhattacharya P.P. and Ephremides A. (1991) Stochastic monotonicity properties of multiserver queues with impatient customers. *Journal of Applied Probability*, 28, 673–682.

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

[3] Zeltyn S. and Mandelbaum A. (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. *Queueing Systems: Theory and Applications (QUESTA)*, 51, 361–402.

[4] Zeltyn S. and Mandelbaum A. (2005) Call centers with impatient customers: many-server asymptotics of the M/M/n+G queue. Internet Supplement. Available at http://iew3.technion.ac.il/serveng/References/references.html.