Call Centers with Impatient Customers: Exact Analysis and Many-Server Asymptotics of the M/M/n+G Queue

Sergey Zeltyn

Faculty of Industrial Engineering and Management, Technion, Haifa, Israel

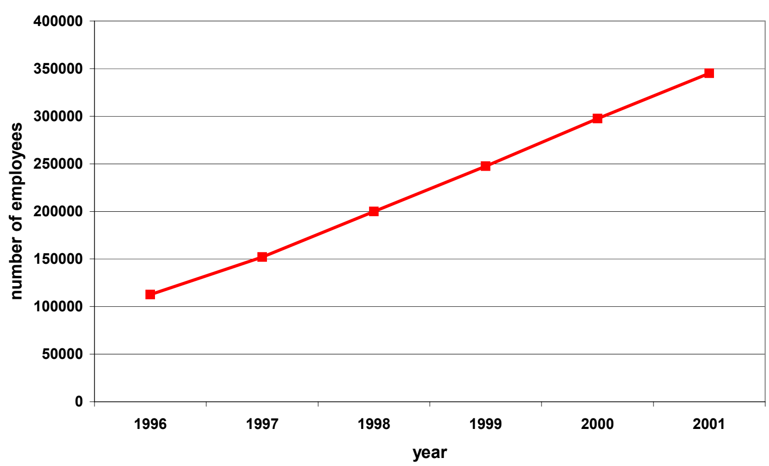
Based on:

- Zeltyn & Mandelbaum. Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n+G Queue. Submitted to *QUESTA*.
- Mandelbaum & Zeltyn. The Impact of Customers' Patience on Delay and Abandonment: Some Empirically-Driven Experiments with the M/M/n+G Queue, *OR Spectrum* (2004) 26:377-411.
- Dimensioning Call Centers with Abandonment. Research in progress (with Borst, Mandelbaum & Reiman).

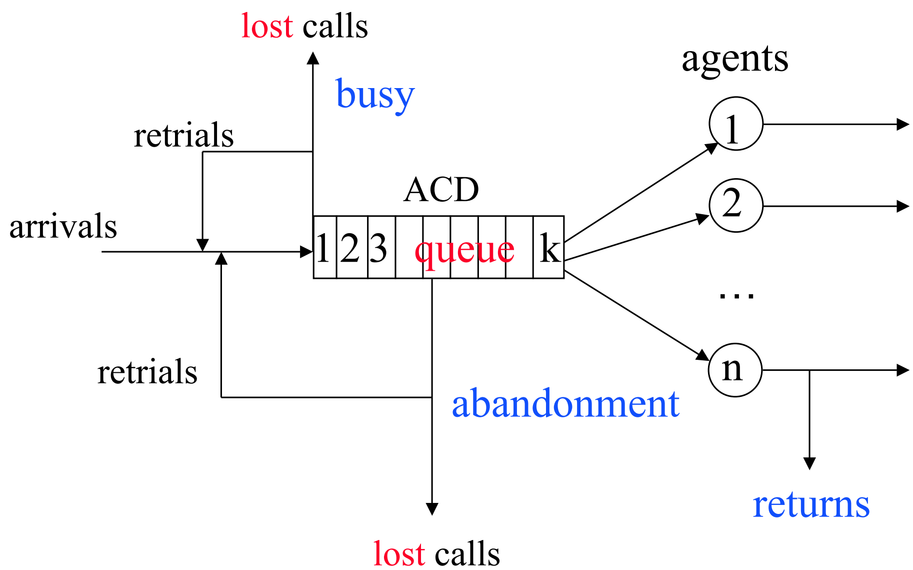
The World of Call Centers

U.S. 3% workforce (several millions);
1000's agents in a "single" call center.
Growing extensively.

Germany: number of call center employees

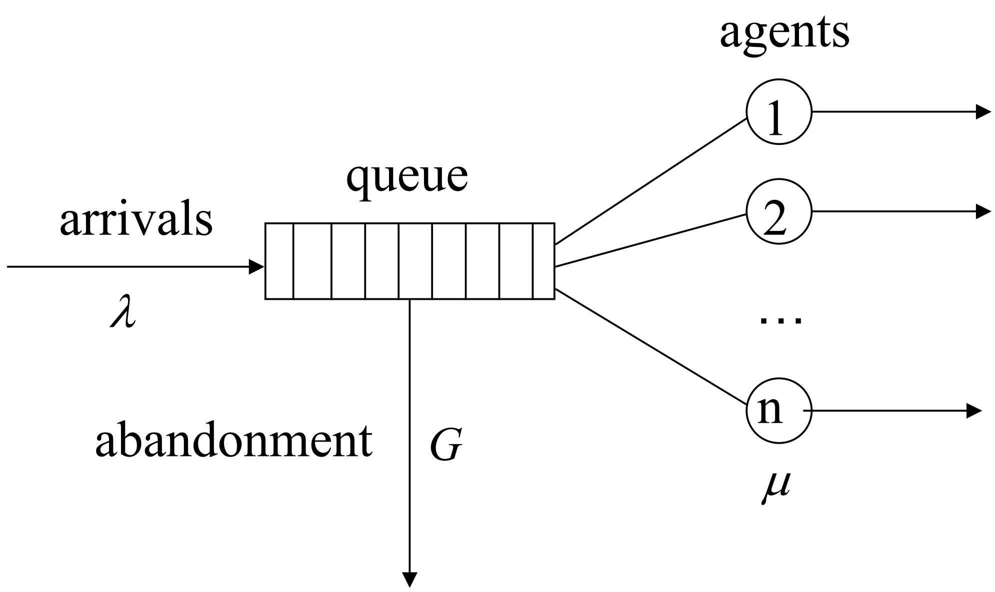


Schematic Representation of a Basic Telephone Call Center



How to model?

M/M/n+G Queue



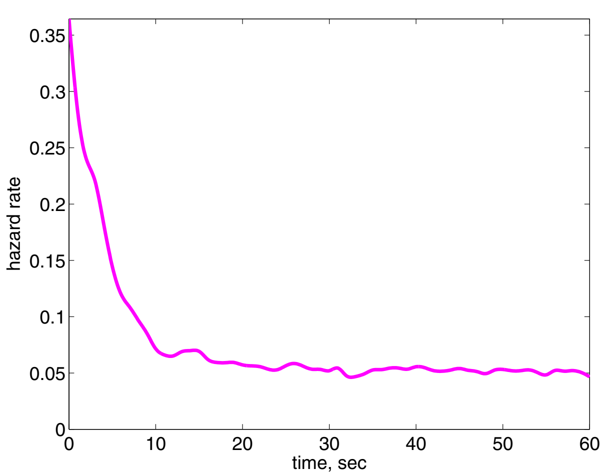
- λ – Poisson arrival rate;
- μ – Exponential service rate;
- n service agents;
- G – Patience distribution.

Modelling Abandonment

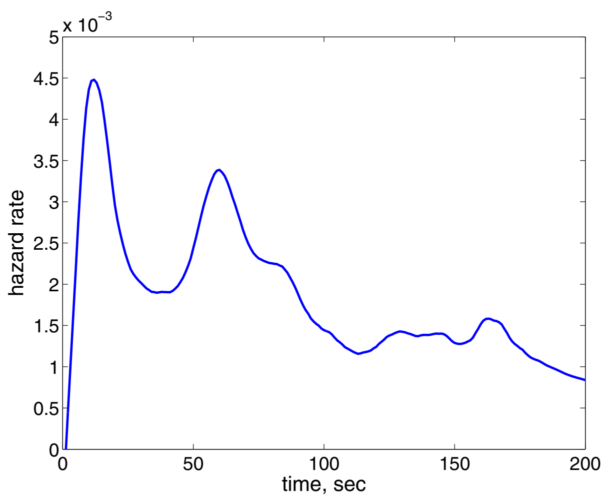
- **Patience time** $\tau \sim G$:
  time a customer is willing to wait for service;
- **Offered wait** V:
  waiting time of a customer with infinite patience;
- If $\tau \leq V$, customer abandons; otherwise, gets service;
- **Actual wait** $W = \min(\tau, V)$.

Customers' Patience: Examples of Hazard Rates

US bank

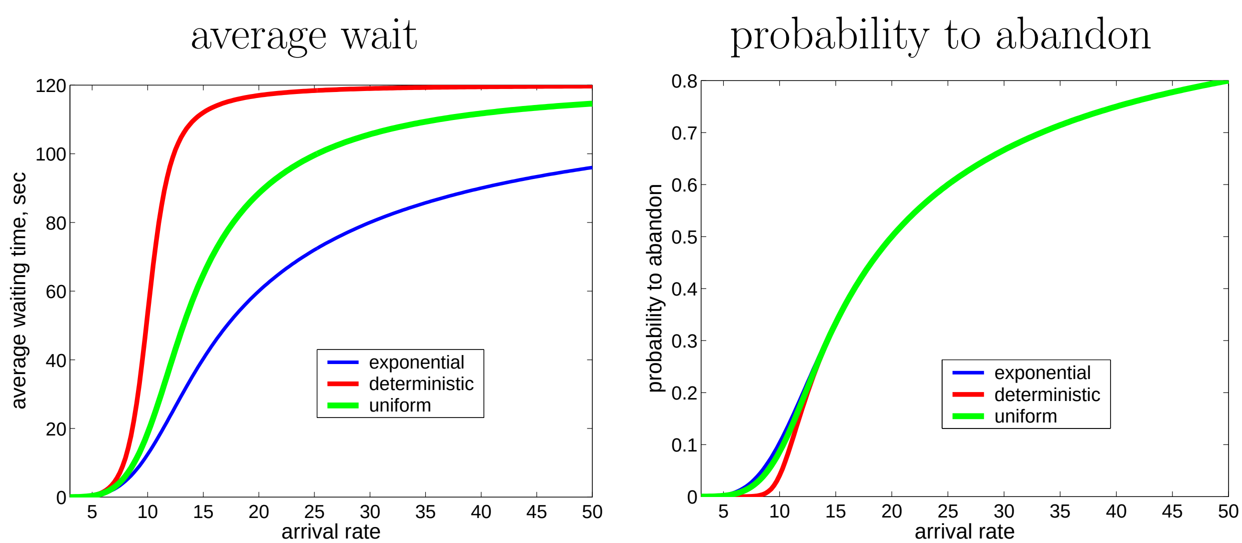


Israeli bank



Impact of Patience Distribution on System Performance

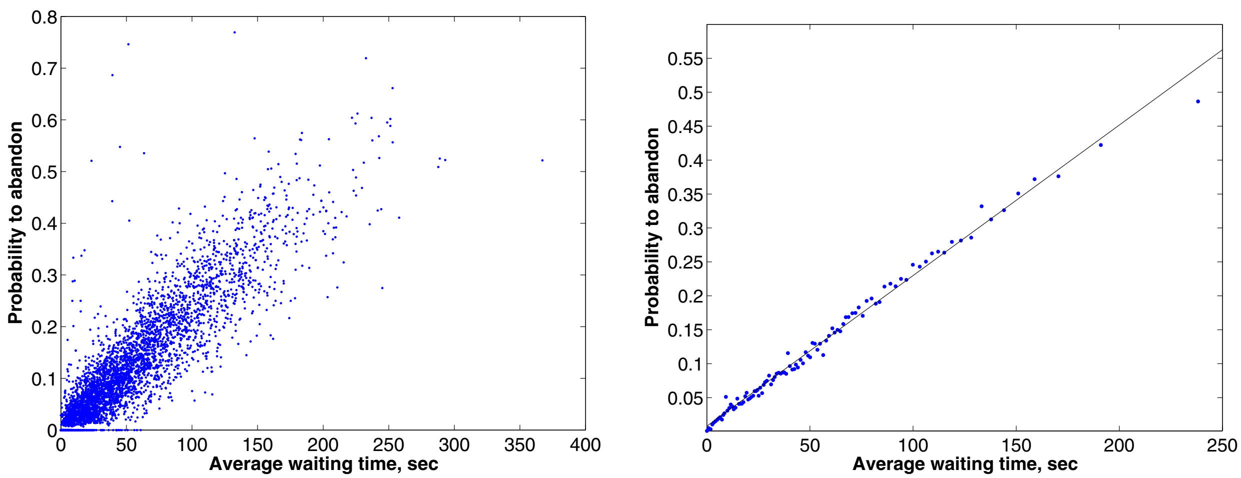
1 min average service time, 2 min average patience, 10 agents, arrival rate varies from 3 to 50 per minute



Conclusion: study models with general patience.

On the Relation between P{Ab} and E[$\boldsymbol{W}$]

Israeli Call Center data: linear pattern



The graphs are based on 4158 hour intervals.

If Patience is $\exp(\theta)$, then

$$\mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[W]\,.$$

However, patience times are not exponential!

Research Goals

- Asymptotic analysis of moderate-to-large call centers;
- Impact of patience distribution on P{Ab}/E[W] relation and performance measures;
- Quality/efficiency tradeoff.

M/M/n+G Queue: Exact Results

- Baccelli and Hebuterne (1981) – probability to abandon, distribution of offered wait;
- Brandt and Brandt (1999, 2002) – number-in-system and waiting time distributions;
- Mandelbaum, Zeltyn (2004) – extensive list of performance measures.

Calculation of Performance Measures: Building blocks

$$H(x) \triangleq \int_0^x \bar{G}(u)du \, ,$$

where $\bar{G}(\cdot)$ is survival function of patience time.

$$\begin{aligned}
J &\triangleq \int_0^\infty \exp\left\{\lambda H(x) - n\mu x\right\} dx \, , \\
J_1 &\triangleq \int_0^\infty x \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx \, , \\
J_H &\triangleq \int_0^\infty H(x) \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx \, , \\
J(t) &\triangleq \int_t^\infty \exp\left\{\lambda H(x) - n\mu x\right\} dx \, . \\
J_1(t) &\triangleq \int_t^\infty x \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx \, , \\
J_H(t) &\triangleq \int_t^\infty H(x) \cdot \exp\left\{\lambda H(x) - n\mu x\right\} dx \, .
\end{aligned}$$

Finally,

$$\mathcal{E} \triangleq \frac{\sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^{n-1}} \, .$$

Performance Measures

P{Ab} – probability to abandon, P{Sr} – probability to be served,
W – waiting time, V – offered wait,
Q – queue length.

$$
\begin{aligned}
\mathrm{P}\{V > 0\} &= \frac{\lambda J}{\mathcal{E} + \lambda J}\,, \\
\mathrm{P}\{W > 0\} &= \frac{\lambda J}{\mathcal{E} + \lambda J} \cdot \bar{G}(0)\,, \\
\mathrm{P}\{\mathrm{Ab}\} &= \frac{1 + (\lambda - n\mu)J}{\mathcal{E} + \lambda J}\,, \\
\mathrm{P}\{\mathrm{Sr}\} &= \frac{\mathcal{E} + n\mu J - 1}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[V] &= \frac{\lambda J_1}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W] &= \frac{\lambda J_H}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[Q] &= \frac{\lambda^2 J_H}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W \mid \mathrm{Ab}] &= \frac{J + \lambda J_H - n\mu J_1}{(\lambda - n\mu)J + 1}\,, \\
\mathrm{E}[W \mid \mathrm{Sr}] &= \frac{n\mu J_1 - J}{\mathcal{E} + n\mu J - 1}\,, \\
\mathrm{P}\{W > t\} &= \frac{\lambda \bar{G}(t) J(t)}{\mathcal{E} + \lambda J}\,, \\
\mathrm{E}[W \mid W > t] &= \frac{J_H(t) - (H(t) - t\bar{G}(t)) \cdot J(t)}{\bar{G}(t) J(t)}\,, \\
\mathrm{P}\{\mathrm{Ab} \mid W > t\} &= \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)}\,.
\end{aligned}
$$

Asymptotic Operational Regimes

Health insurance company. ACD Report.

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

M/M/n+G: QED Operational Regime.

Main case: positive density of patience at the origin.

Density of patience time: $g = \{g(x), x \geq 0\}$, where $g(0) \triangleq g_0 > 0$.

Fix service rate μ.

Let arrival rate $\lambda \to \infty$ and

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

Square-Root Staffing Rule: Described by Erlang in 1924!
Formal analysis:

- Erlang-C: Halfin & Whitt (1981), $\beta > 0$;
- Erlang-B (M/M/n/n): Jagerman (1974);
- Erlang-A: Garnett, Mandelbaum, Reiman (2002);
- Mandelbaum & Zeltyn (2004).

Building Blocks

$$
\begin{aligned}
J &= \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \\
\mathcal{E} &= \frac{\sqrt{n}}{h(-\beta)} + o(\sqrt{n}), \\
J_1 &= \frac{1}{n\mu g_0}\left[1 - \frac{\hat{\beta}}{h(\hat{\beta})}\right] + o\left(\frac{1}{n}\right),
\end{aligned}
$$

where

$$
\hat{\beta} \triangleq \beta\sqrt{\frac{\mu}{g_0}},
$$

$h(\cdot)$ – hazard rate of standard normal distribution.

Proofs: Combine M/M/n+G formulae above and the Laplace method for asymptotic calculation of integrals.

Main Case: Performance Measures

- Probability of wait converges to constant:

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

- Probability to abandon decreases at rate $\dfrac{1}{\sqrt{n}}$:

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

- Average wait decreases at rate $\dfrac{1}{\sqrt{n}}$:

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

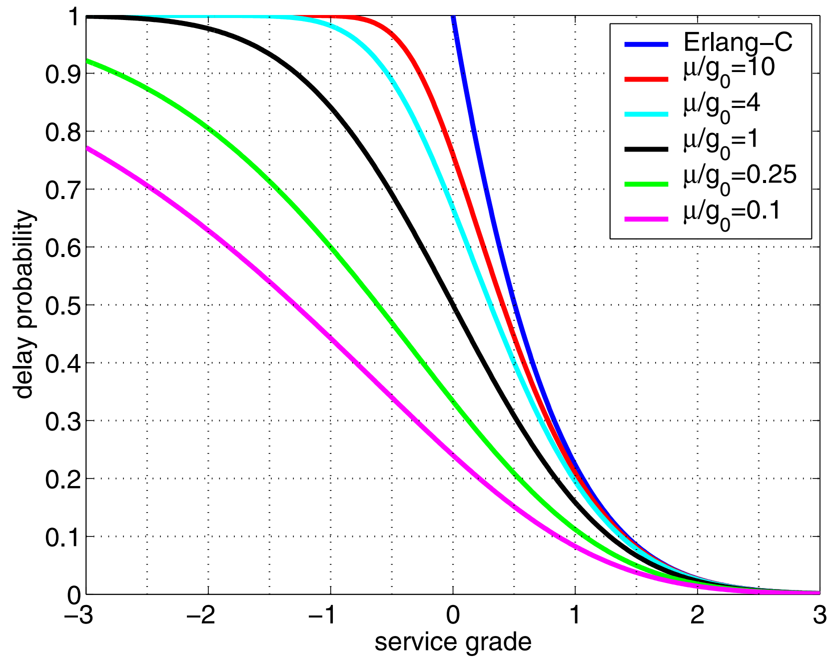
- Ratio between $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$ converges to patience density at the origin:

$$\boxed{\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;\sim\; g_0}$$

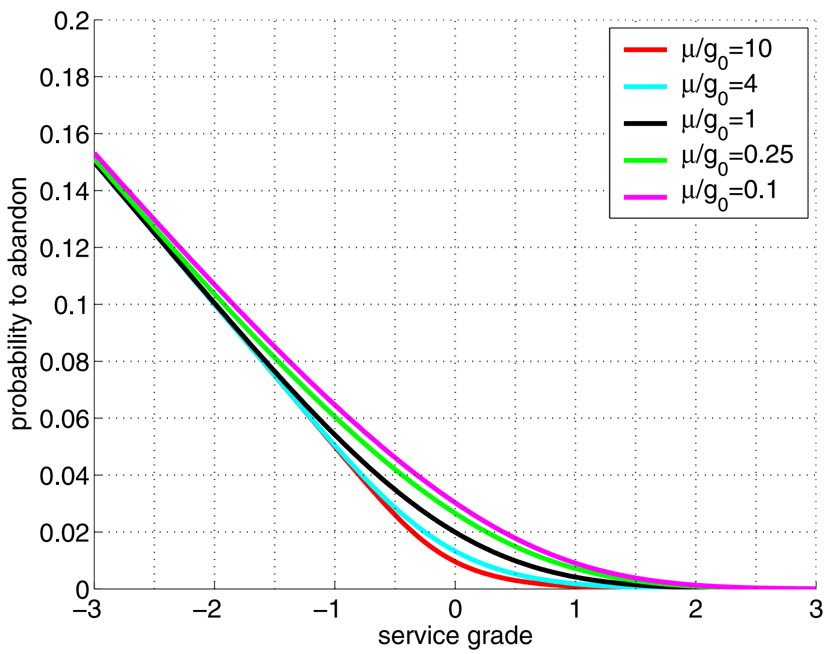
- Asymptotic distribution of wait:

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

QED Regime: Delay Probability



QED Regime: Probability to Abandon (n=400)



Note convergence to $-\beta/\sqrt{n}$ for large negative β.

QED Operational Regime: Right Answer for Wrong Reasons

If $\beta = 0$, QED staffing level:

$$n \;=\; \frac{\lambda}{\mu} \;=\; R\,.$$

Equivalent to deterministic rule: assign number of agents equal to *offered load*. (Common in stochastic-ignorant operations.)

M/M/n (Erlang-C): queue "explodes".

M/M/n+G: assume $\mu = g_0$. Then $\mathrm{P}\{W = 0\} \approx 50\%$.

If $n = 100$, $\mathrm{P}\{\mathrm{Ab}\} \approx 4\%$, and $\mathrm{E}[W] \approx 0.04 \cdot \mathrm{E}[S]$.

Overall, good service level.

QED Operational Regime: Special Cases

- **Patience density vanishing near the origin.**
  (k-1) derivatives at the origin are zero, the k-th derivative is positive.
  Examples: Erlang, Phase-type.
  - If $\beta > 0$, wait similar to Erlang-C. P$\{$Ab$\}$ decreases at $n^{-(k+1)/2}$ rate.
  - If $\beta < 0$, almost all customers delayed, E$[W] \to 0$ slowly. P$\{$Ab$\} \approx -\beta/\sqrt{n}$.
  - If $\beta = 0$, intermediate behavior.
- **Delayed distribution of patience.**
  Customers do not abandon till $c > 0$.
  Examples: Delayed exponential, deterministic.
  Similar to the previous case. For $\beta < 0$, wait converges to c.
- **Balking.**
  Customer, not served immediately, balks with probability P$\{$Blk$\}$.
  Example. M/M/n/n (Erlang-B).
  - P$\{W > 0\}$ decreases at rate $1/\sqrt{n}$;
  - P$\{$Ab$|V > 0\} \approx$ P$\{$Blk$\}$;
  - P$\{$Ab$\} \approx h(-\beta)/\sqrt{n}$, asymptotic loss probability for Erlang-B.

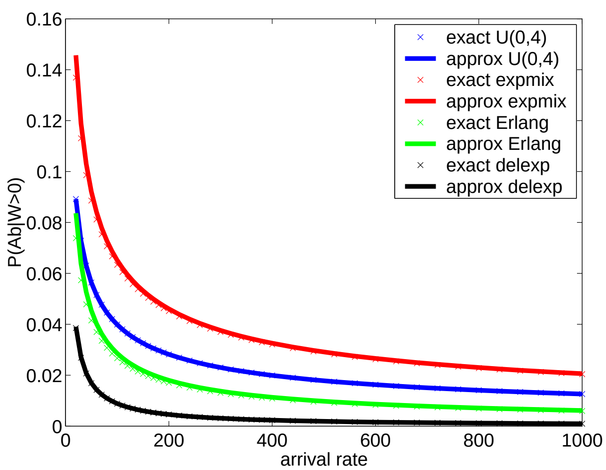
QED Regime: Numerical Experiments–1

Patience distributions:

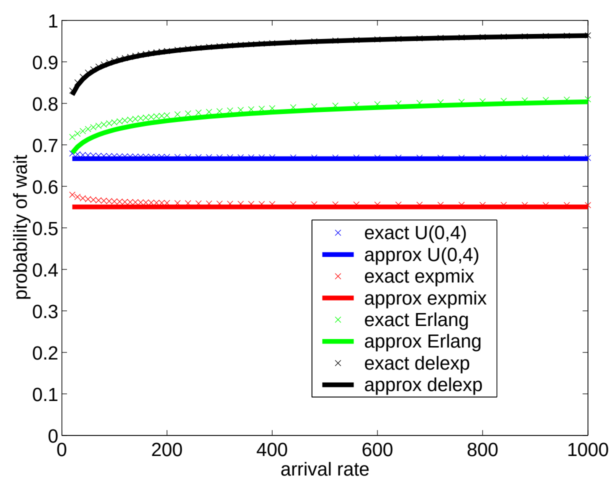
- *Uniform* on [0,4], $g_0 = 0.25$;
- *Hyperexponential*, 50-50% mixture of exp(mean=1) and exp(mean=1/3), $g_0 = 2/3$;
- *Erlang*, two exp(mean=1) phases, $g_0 = 0$;
- *Delayed exponential*, 1 + exp(mean=1), $g_0 = 0$.

Service grade $\beta = 0$.

Probability to abandon given delay vs. arrival rate



Probability of wait vs. arrival rate

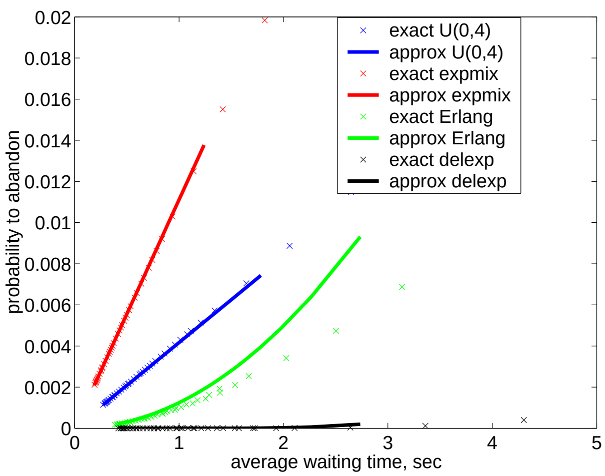


P{Ab} convergence rates: $1/\sqrt{n}$, $1/\sqrt{n}$, $n^{-2/3}$, exp, respectively.

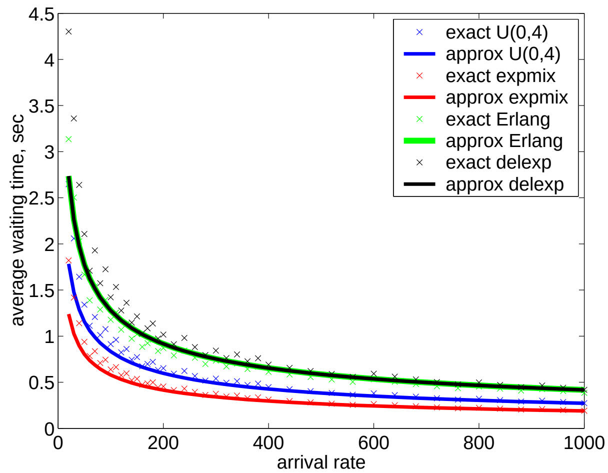
QED Regime: Numerical Experiments–2

Service grade $\beta = 1$.

Probability to abandon
vs. average waiting time



Average waiting time
vs. arrival rate



Note linear patterns in the first plot.

M/M/n+G: QD Operational Regime.

Density of patience time at the origin $g_0 > 0$.
Staffing level

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

Performance Measures

- $\mathrm{P}\{W > 0\}$ decreases exponentially in n.
- Probability to abandon of delayed customers:

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

- Average wait of delayed customers:

$$\mathrm{E}[W \mid W > 0] \;=\; \frac{1}{n} \cdot \frac{1+\gamma}{\gamma} \cdot \frac{1}{\mu} + o\left(\frac{1}{n}\right) \,.$$

- Linear relation between $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$.

$$\boxed{\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W]} \;\sim\; g_0}$$

Numerical experiments: QED approximations are better, except very high-performance systems.

M/M/n+G: ED Operational Regime.

Assume $G(x) = \gamma$ has a unique solution x^* and $g(x^*) > 0$.

Staffing level

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

Performance Measures

- $\mathrm{P}\{W = 0\}$ decreases exponentially in n.
- Probability to abandon converges to:

$$\mathrm{P}\{\mathrm{Ab}\} \;\sim\; \gamma \;\approx\; 1 - \frac{1}{\rho}\,.$$

- Offered wait converges to x^*:

$$\mathrm{E}[V] \;\sim\; x^*\,, \qquad V \;\xrightarrow{p}\; x^*\,.$$

- Distribution G^* of $\min(x^*, \tau)$

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

Asymptotic distribution of wait:

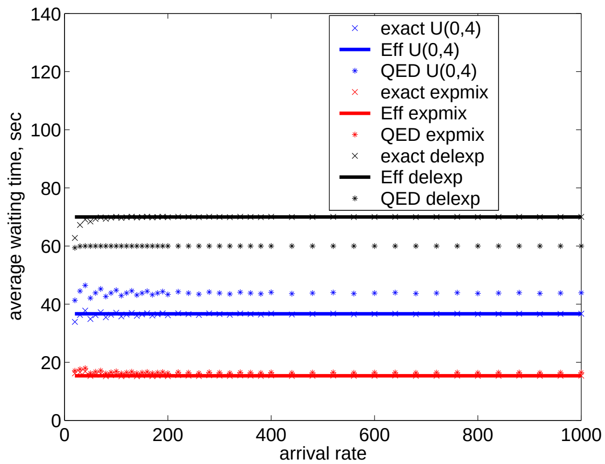
$$W \;\xrightarrow{w}\; G^*\,, \qquad \mathrm{E}[W] \;\to\; \mathrm{E}[\min(x^*, \tau)]\,.$$

ED Regime: Numerical Experiments

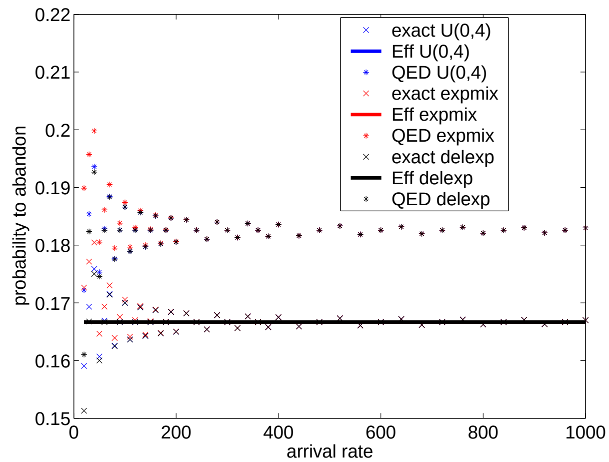
Patience distributions: Uniform, hyperexponential, delayed exponential. Compared with exact and QED.

Service grade $\gamma = 1/6$, $\rho = 1.2$.

Average waiting time



Probability to abandon



For heavy-loaded systems, ED approximations for $P\{Ab\}$ and $E[W]$ can be better than QED.

Current research on the ED regime: Whitt (2004), Bassamboo, Harrison and Zeevi (2004).

Impact of Customers' Patience: Theoretical Results

Lemma. Consider M/M/n+G; λ, μ, and n fixed.
Assume that for two patience distributions G_1 and G_2:

$$\int_0^x \bar{G}_1(\eta)d\eta \ \geq \ \int_0^x \bar{G}_2(\eta)d\eta \,, \qquad x > 0 \,.$$

Then,

a. $\mathrm{P}^1\{V > 0\} \ \geq \ \mathrm{P}^2\{V > 0\}; \quad \mathrm{P}^1\{W > 0\} \ \geq \ \mathrm{P}^2\{W > 0\}.$

b. $\mathrm{P}^1\{\mathrm{Ab}\} \ \leq \ \mathrm{P}^2\{\mathrm{Ab}\}; \quad \mathrm{P}^1\{\mathrm{Ab}|V > 0\} \ \leq \ \mathrm{P}^2\{\mathrm{Ab}|V > 0\}.$

Proof. Follows from formulae for performance measures.

Theorem. In addition, fix average patience $\bar{\tau}$.
Let G_d be the deterministic patience distribution. Then

a. G_d maximizes the probabilities of wait $\mathrm{P}\{W > 0\}$ and $\mathrm{P}\{V > 0\}$.
b. G_d minimizes the probabilities to abandon $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{P}\{\mathrm{Ab}|V > 0\}$.
c. G_d maximizes the average wait $\mathrm{E}[W]$.
d. G_d maximizes the average queue length $\mathrm{E}[Q]$.

Proof. a+b. Follow from Lemma.
c. Functional maximization. Variation calculus.
d. Follows from Little's formula.

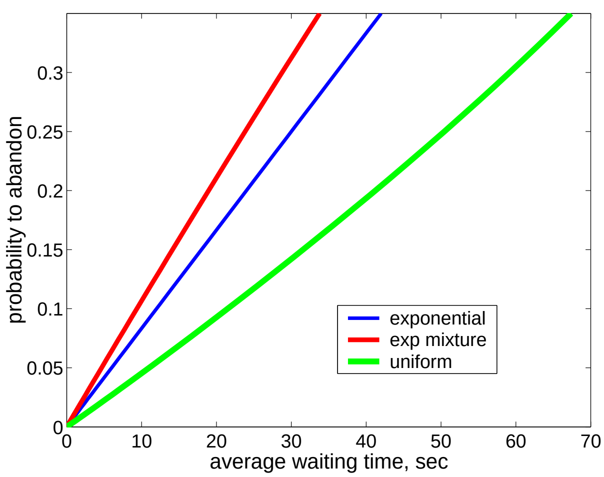
Impact of Customers' Patience: Numerical Results

Linear relations (empirically): Exp(mean=2), Uniform(0,4), Hyperexponential.

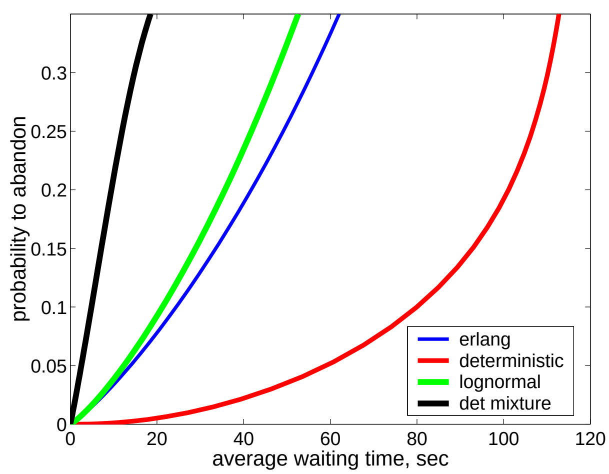
Non-linear relations: Deterministic(2), Erlang, Lognormal(2,2), mixture of two constants (0.2,3.8).

1 min average service time, 2 min average patience,
10 agents, arrival rate increases

linear relations



non-linear relations

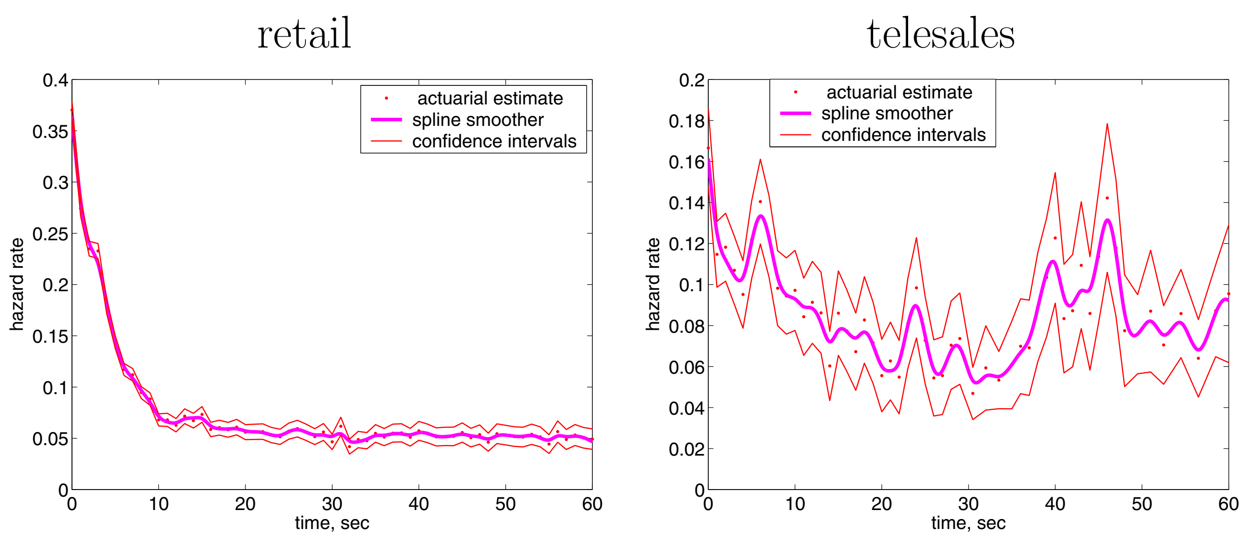


Some Applications to Call Centers

Large US bank.
Daily volume 70,000 calls; 900-1200 agents positions on weekdays.
Two service types analyzed for 5 months.

	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

Estimates of hazard rate



Problems/Challenges:

- Reliable data for number of agents n unavailable;
- Work-conservation does not always prevail;
- Significant variability of hazard/density near the origin.

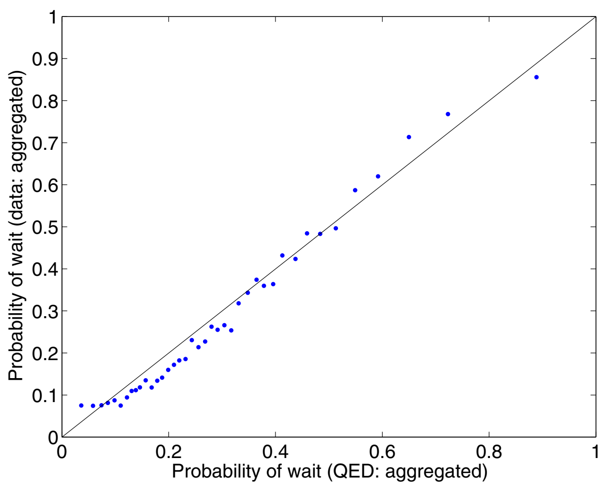
Fitting QED Approximations

Estimate n via some performance measure (P{Ab}). Fit other performance measure(s).

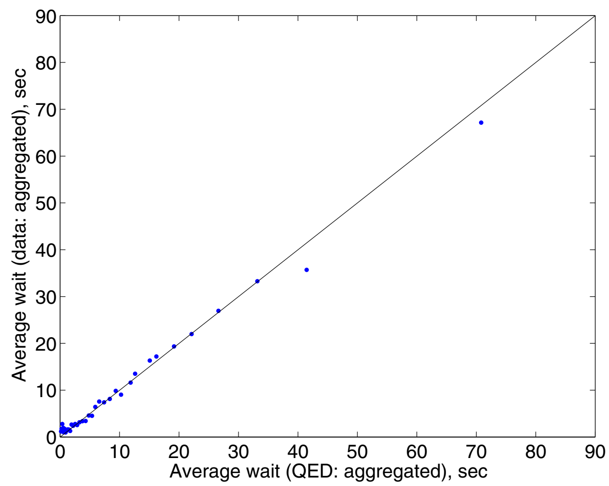
Substitute g_0 := estimate of $h(0) \Rightarrow$ unsatisfactory fit.

Solution: Substitute $g_0 :=$ overall P{Ab}/E[W] to QED formulae.

Retail. $\mathrm{P}\{W > 0\}$

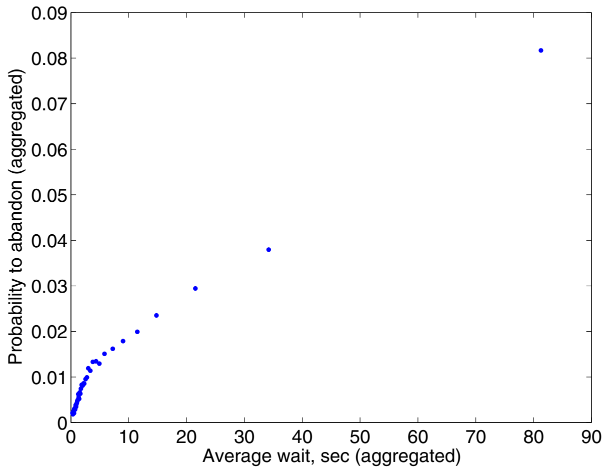


Telesales. $\mathrm{E}[W]$

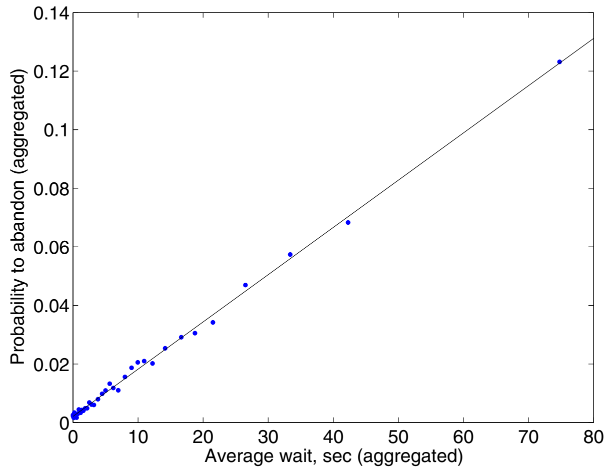


P{Ab}/E[W] Relation

Retail



Telesales



For telesales, hazard variability near the origin much smaller. Hence, pattern much closer to straight line.

Dimensioning and QED Regime

Erlang-C: Borst, Mandelbaum & Reiman, 2004.

Erlang-A, M/M/n+G with Zeltyn, in progress.

$$\text{Cost} = c \cdot n + d \cdot \lambda \mathrm{E}[W]\,,$$

c – cost of staffing;
d – cost of delay (cost of abandonment can be considered too).

Erlang-C. Optimal staffing level:

$$n^* \;\approx\; R + y^*(r)\sqrt{R}, \qquad r = \text{delay cost/staffing cost}\,.$$

Erlang-A. Optimal staffing level (conjecture):

$$n^* \;\approx\; R + y^*(r;s)\sqrt{R}, \qquad s = \sqrt{\mu/\theta}\,,$$

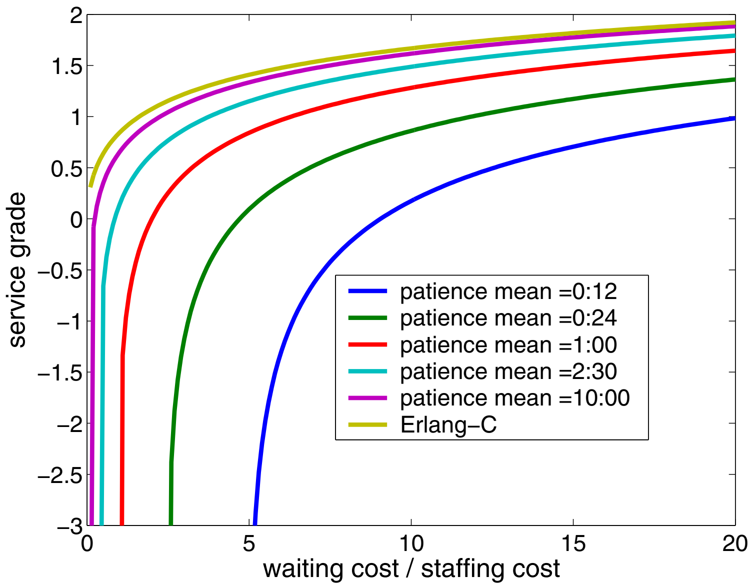
$$y^*(r;s) \;=\; \arg \min_{-\infty \le y < \infty} \{y + r \cdot P_w(y;s) \cdot s \cdot [h(ys) - ys]\}\;,$$

where

$$P_w(y;s) \;=\; \left[1 + \frac{h(ys)}{sh(-y)}\right]^{-1}\;.$$

Optimal Service Grade

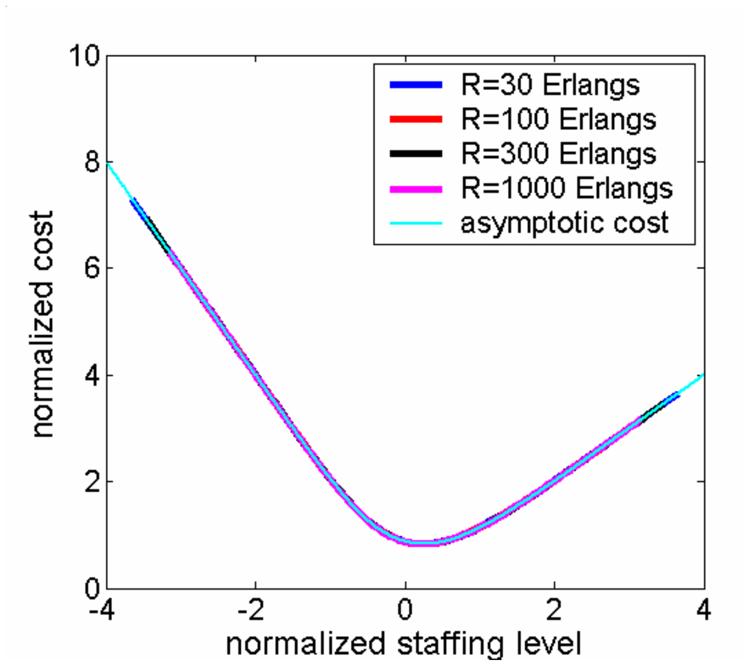
1 min average service time



- $r < \theta/\mu$ implies that "no service" is optimal.
- $r \leq 20 \;\Rightarrow y^* < 2; \quad r \leq 500 \;\Rightarrow\; y^* < 3!$
- Numerical tests exhibit **remarkable** accuracy.

Actual Cost vs. Asymptotic Cost

$\boldsymbol{\mu = 1, \theta = 1/3}$



Normalized staffing level $= (n - R)/\sqrt{R}$;

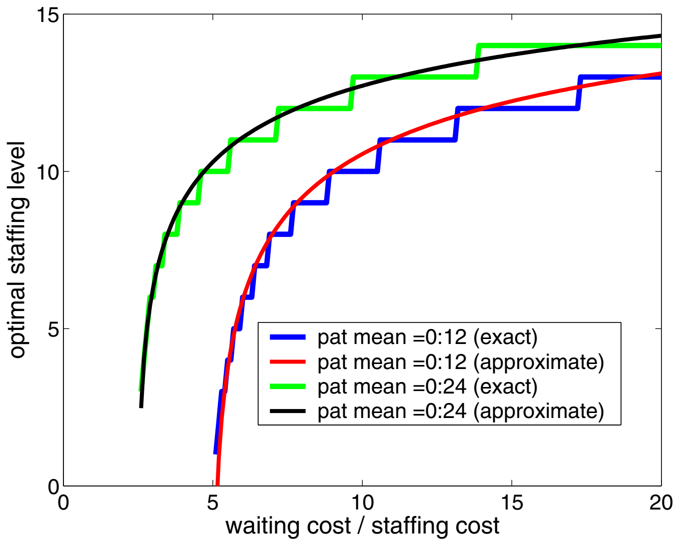
Normalized cost $= (\text{cost} - cR)/\sqrt{R}$;

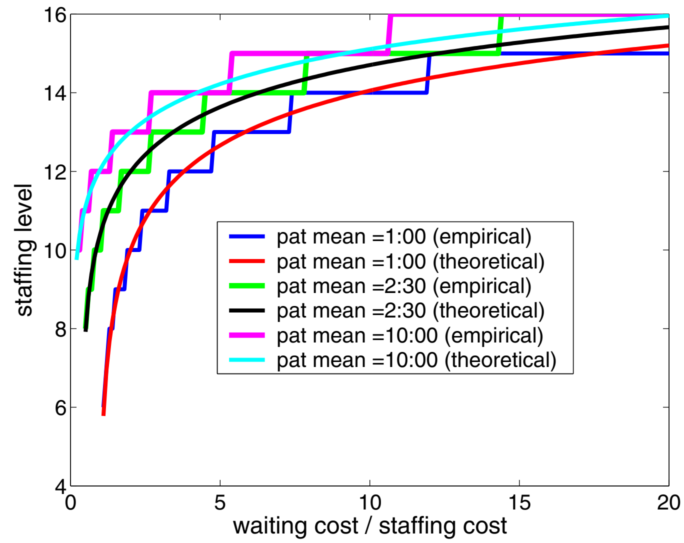
Asymptotic cost $= c \cdot y + d \cdot P_w(y; s) \cdot s \cdot [h(ys) - ys]$,

where $y =$ QED service grade.

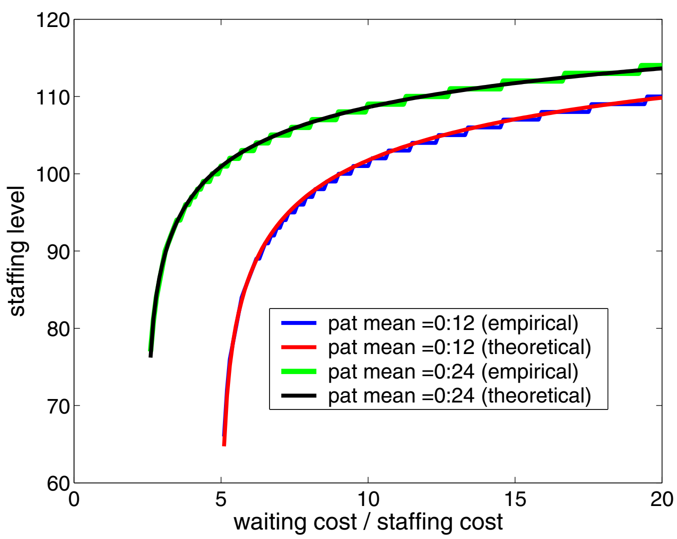
Erlang-A: Optimal Staffing

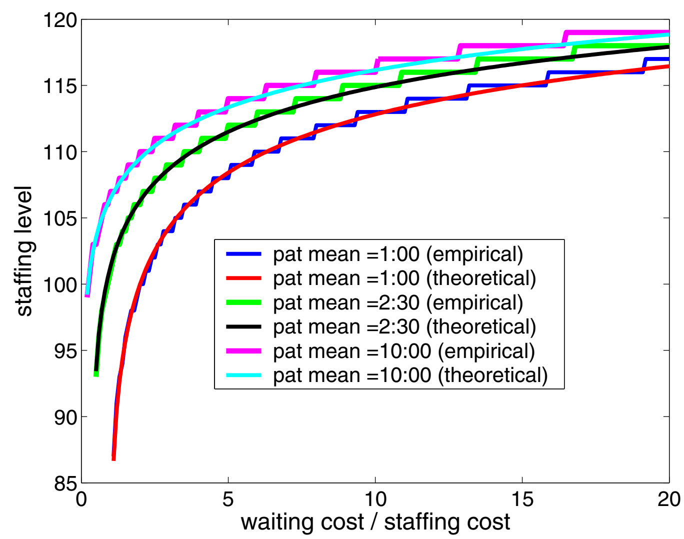
$\lambda = 10, \ \mu = 1$





$\lambda = 100, \ \mu = 1$

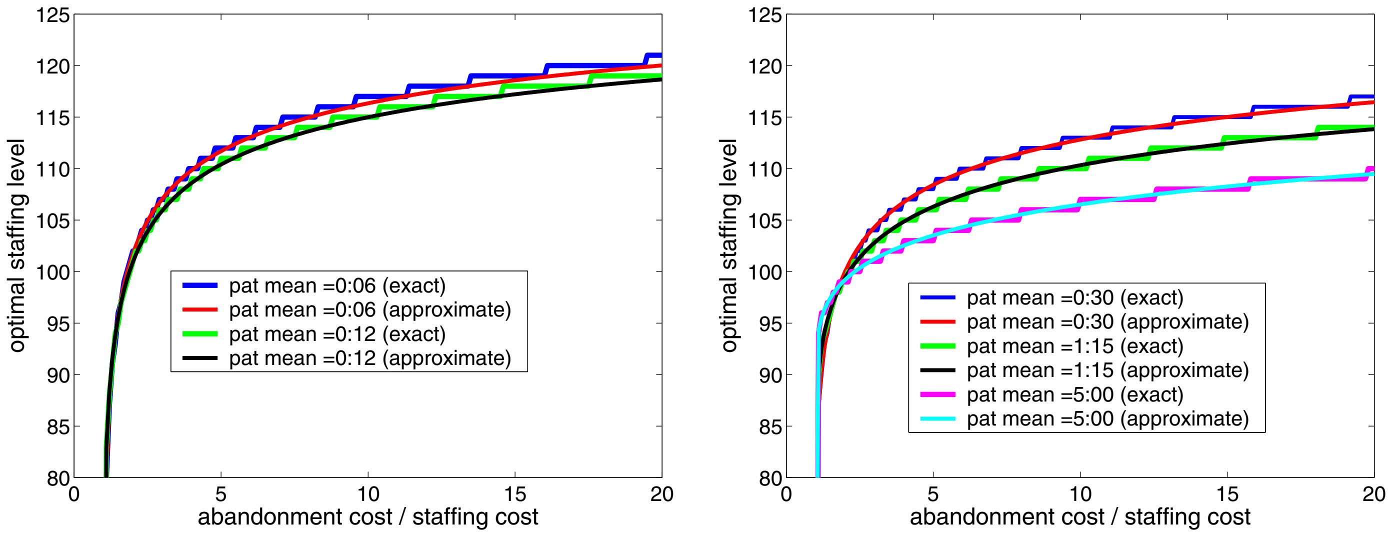




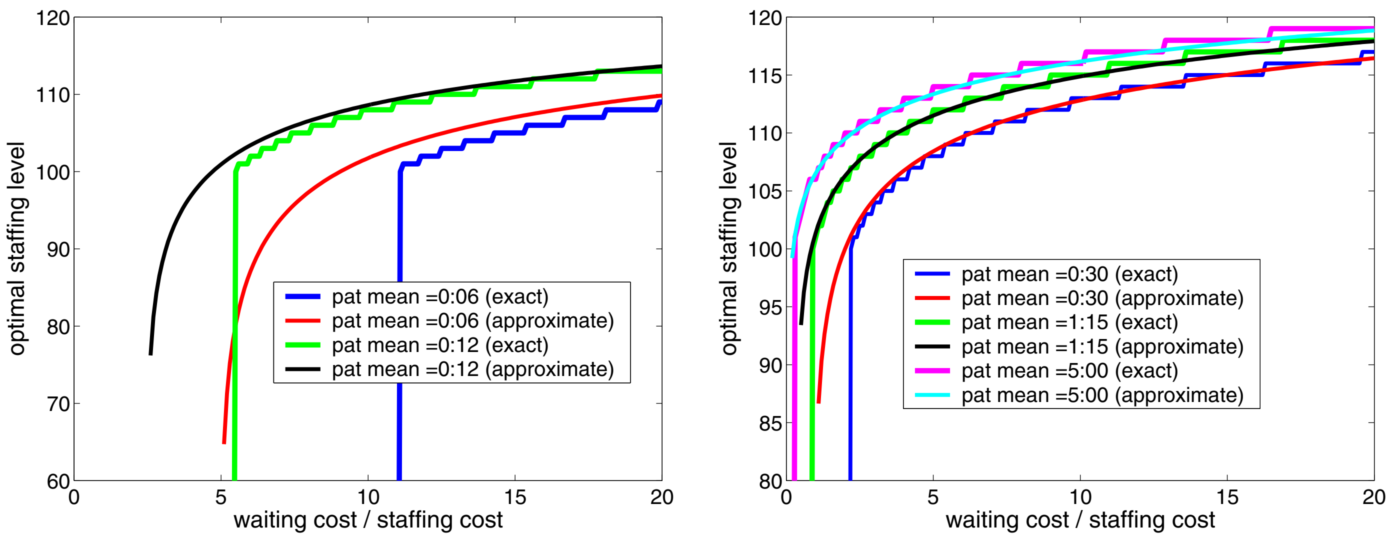
M/M/n+G: Optimal Staffing

Uniformly Distributed Patience

$$\text{Cost} = c \cdot n + a \cdot \lambda \mathrm{P}\{\mathrm{Ab}\}$$



$$\text{Cost} = c \cdot n + d \cdot \lambda \mathrm{E}[W]$$



Conclusions

QED approximation: Careful balance of quality and efficiency. Optimal staffing for linear staffing/waiting costs.

Can be performed using any software that provides the standard normal distribution (e.g. Excel). Works well for

- Number of servers n from 10's to 1000's;
- Agents highly utilized but not overloaded (∼90-98%);
- Probability of delay 10-90%;
- Probability to abandon: 3-7% for small n, 1-4% for large n.

ED approximation: Useful for overloaded call centers.

Requires solving equation $G(x) = \gamma$, and integration (calculating $H(x^*)$). Works well for

- Number of servers $n \geq 100$.
- Agents very highly utilized (close to 100%);
- Probability of delay: more than 85%;
- Probability to abandon: more than 5%.

QD approximation: preferable only for very high-performance systems.

Additional Research Directions

- Queues with uncertainty about the arrival rate.
- Queues with time-inhomogeneous arrival rate (Feldman, Mandelbaum, Massey, Whitt).
- More data analysis (Israeli cellular-phone company).
- Generally distributed service times: M/G/n+G (recent papers of Whitt).