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.


• 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

- \( \lambda \) – Poisson arrival rate;
- \( \mu \) – 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

![Hazard rate for US bank](image1)

<table>
<thead>
<tr>
<th>time, sec</th>
<th>hazard rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.15</td>
</tr>
<tr>
<td>20</td>
<td>0.25</td>
</tr>
<tr>
<td>30</td>
<td>0.3</td>
</tr>
<tr>
<td>40</td>
<td>0.35</td>
</tr>
<tr>
<td>50</td>
<td>0.4</td>
</tr>
<tr>
<td>60</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Israel bank

![Hazard rate for Israeli bank](image2)

<table>
<thead>
<tr>
<th>time, sec</th>
<th>hazard rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5 x 10^{-3}</td>
</tr>
<tr>
<td>50</td>
<td>4.5</td>
</tr>
<tr>
<td>100</td>
<td>3.5</td>
</tr>
<tr>
<td>150</td>
<td>2.5</td>
</tr>
<tr>
<td>200</td>
<td>1.5</td>
</tr>
</tbody>
</table>
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

average wait  probability to abandon

Conclusion: study models with general patience.
On the Relation between $P\{\text{Ab}\}$ and $E[W]$

Israeli Call Center data: linear pattern

The graphs are based on 4158 hour intervals.

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

$$P\{\text{Ab}\} = \theta \cdot E[W].$$

However, patience times are not exponential!
Research Goals

- Asymptotic analysis of moderate-to-large call centers;
- Impact of patience distribution on $P\{\text{Ab}\}/E[W]$ 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;
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.

\[
\begin{align*}
J & \triangleq \int_{0}^{\infty} \exp \{ \lambda H(x) - n \mu x \} \, dx , \\
J_1 & \triangleq \int_{0}^{\infty} x \cdot \exp \{ \lambda H(x) - n \mu x \} \, dx , \\
J_H & \triangleq \int_{0}^{\infty} H(x) \cdot \exp \{ \lambda H(x) - n \mu x \} \, dx , \\
J(t) & \triangleq \int_{t}^{\infty} \exp \{ \lambda H(x) - n \mu x \} \, dx . \\
J_1(t) & \triangleq \int_{t}^{\infty} x \cdot \exp \{ \lambda H(x) - n \mu x \} \, dx , \\
J_H(t) & \triangleq \int_{t}^{\infty} H(x) \cdot \exp \{ \lambda H(x) - n \mu x \} \, dx .
\end{align*}
\]

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

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

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

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

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

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

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

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

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

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

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

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

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

$$P\{\text{Ab} \mid W > t\} = \frac{\lambda - n\mu - G(t)}{\lambda \bar{G}(t)} + \frac{\exp\{\lambda H(t) - n\mu t\}}{\lambda \bar{G}(t) J(t)}.$$
## Asymptotic Operational Regimes

Health insurance company. ACD Report.

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

Let arrival rate \( \lambda \rightarrow \infty \) and

\[
n = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty.
\]

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);
Building Blocks

\[ J = \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mu g_0}} \cdot \frac{1}{h(\hat{\beta})} + o\left(\frac{1}{\sqrt{n}}\right), \]

\[ \mathcal{E} = \frac{\sqrt{n}}{h(-\beta)} + o(\sqrt{n}), \]

\[ J_1 = \frac{1}{n\mu g_0} \left[ 1 - \frac{\hat{\beta}}{h(\hat{\beta})} \right] + o\left(\frac{1}{n}\right), \]

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:

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

• Probability to abandon decreases at rate \(\frac{1}{\sqrt{n}}\):

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

• Average wait decreases at rate \(\frac{1}{\sqrt{n}}\):

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

• Ratio between \(P\{\text{Ab}\}\) and \(E[W]\) converges to patience density at the origin:

\[
\frac{P\{\text{Ab}\}}{E[W]} \sim g_0
\]

• Asymptotic distribution of wait:

\[
P\left\{\frac{W}{E[S]} > \frac{t}{\sqrt{n}} \mid W > 0\right\} \sim \frac{\Phi \left(\hat{\beta} + \sqrt{\frac{g_0}{\mu}} \cdot t\right)}{\Phi(\hat{\beta})}, \quad t \geq 0.
\]
QED Regime: Delay Probability

QED Regime: Probability to Abandon (n=400)

Note convergence to $-\beta/\sqrt{n}$ for large negative $\beta$. 
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 $P\{W = 0\} \approx 50\%$.

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

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. $P\{\text{Ab}\}$ decreases at $n^{-(k+1)/2}$ rate.
  - If $\beta < 0$, almost all customers delayed, $E[W] \to 0$ slowly.
    \[ P\{\text{Ab}\} \approx -\beta/\sqrt{n}. \]
  - 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 $P\{\text{Blk}\}$.
  Example. $M/M/n/n$ (Erlang-B).
  \[
  \begin{align*}
  &- P\{W > 0\} \text{ decreases at rate } 1/\sqrt{n}; \\
  &- P\{\text{Ab}|V > 0\} \approx P\{\text{Blk}\}; \\
  &- P\{\text{Ab}\} \approx h(-\beta)/\sqrt{n}, \text{ asymptotic loss probability for Erlang-B.}
  \end{align*}
  \]
QED Regime: Numerical Experiments–1

Patience distributions:

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

Service grade \(\beta = 0\).

Probability to abandon given delay vs. arrival rate

![Graph of Probability to abandon given delay vs. arrival rate]

Probability of wait vs. arrival rate

![Graph of Probability of wait vs. arrival rate]

\(P\{\text{Ab}\}\) 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

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

Performance Measures

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

\[
P\{\text{Ab}|W > 0\} = \frac{1}{n} \cdot \frac{1 + \gamma}{\gamma} \cdot \frac{g_0}{\mu} + o\left(\frac{1}{n}\right).
\]

- Average wait of delayed customers:

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

- Linear relation between \( P\{\text{Ab}\} \) and \( E[W] \).

\[
\frac{P\{\text{Ab}\}}{E[W]} \sim g_0
\]

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

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

Performance Measures

- $P\{W = 0\}$ decreases exponentially in $n$.
- Probability to abandon converges to:
  $$P\{\text{Ab}\} \sim \gamma \approx 1 - \frac{1}{\rho}.$$  
- Offered wait converges to $x^*$:
  $$E[V] \sim x^*, \quad V \overset{p}{\to} x^*.$$  
- Distribution $G^*$ of $\min(x^*, \tau)$
  $$G^*(x) = \begin{cases} 
G(x)/\gamma, & x \leq x^* \\
1, & x > x^*
\end{cases}$$  
  Asymptotic distribution of wait:
  $$W \overset{w}{\to} G^*, \quad E[W] \to E[\min(x^*, \tau)].$$
ED Regime: Numerical Experiments

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

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

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

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 $G_2$:

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

Then,

a. $P_1\{V > 0\} \geq P_2\{V > 0\}$; $P_1\{W > 0\} \geq P_2\{W > 0\}$.

b. $P_1\{\text{Ab}\} \leq P_2\{\text{Ab}\}$; $P_1\{\text{Ab}\mid V > 0\} \leq P_2\{\text{Ab}\mid V > 0\}$.

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 $P\{W > 0\}$ and $P\{V > 0\}$.

b. $G_d$ minimizes the probabilities to abandon $P\{\text{Ab}\}$ and $P\{\text{Ab}\mid V > 0\}$.

c. $G_d$ maximizes the average wait $E[W]$.

d. $G_d$ maximizes the average queue length $E[Q]$.

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
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.

<table>
<thead>
<tr>
<th></th>
<th>Calls</th>
<th>E[S]</th>
<th>P{W &gt; 0}</th>
<th>P{Ab}</th>
<th>E[W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retail</td>
<td>3,451,743</td>
<td>224.6 sec</td>
<td>30.6%</td>
<td>1.16%</td>
<td>6.33 sec</td>
</tr>
<tr>
<td>Telesales</td>
<td>349,371</td>
<td>453.9 sec</td>
<td>24.3%</td>
<td>1.76%</td>
<td>9.66 sec</td>
</tr>
</tbody>
</table>

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 $g_0 := \text{estimate of } h(0) \Rightarrow \text{unsatisfactory fit}$.

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

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

![Graph of retail probability of wait compared to QED model](image)

Telesales. $E[W]$

![Graph of telesales average wait compared to QED model](image)
For telesales, hazard variability near the origin much smaller. Hence, pattern much closer to straight line.
Dimensioning and QED Regime


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

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

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

Erlang-C. Optimal staffing level:

\[
n^* \approx R + y^*(r)\sqrt{R}, \quad r = \text{delay cost/staffing cost}.
\]

Erlang-A. Optimal staffing level (conjecture):

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

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

where

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

1 min average service time

• $r < \frac{\theta}{\mu}$ implies that “no service” is optimal.
• $r \leq 20 \Rightarrow y^* < 2$; $r \leq 500 \Rightarrow y^* < 3$!
• Numerical tests exhibit remarkable accuracy.
Actual Cost vs. Asymptotic Cost

\[ \mu = 1, \, \theta = 1/3 \]

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 \(y = \text{QED service grade}\).
Erlang-A: Optimal Staffing

$\lambda = 10, \mu = 1$

$\lambda = 100, \mu = 1$
M/M/n+G: Optimal Staffing

Uniformly Distributed Patience

Cost = \( c \cdot n + a \cdot \lambda P\{\text{Ab}\} \)

![Graph showing the relationship between abandonment cost/staffing cost, optimal staffing level, and patience mean for various scenarios.]

Cost = \( c \cdot n + d \cdot \lambda E[W] \)

![Graph showing the relationship between waiting cost/staffing cost, optimal staffing level, and patience mean for various scenarios.]
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 (\( \sim 90-98\% \));
- 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).