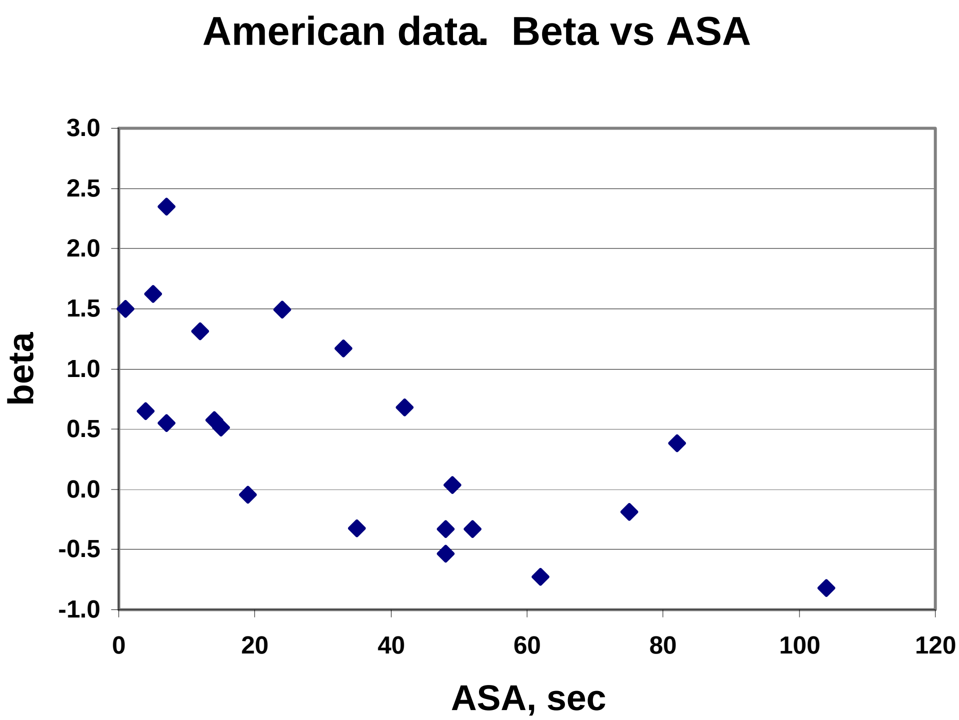
Service Engineering

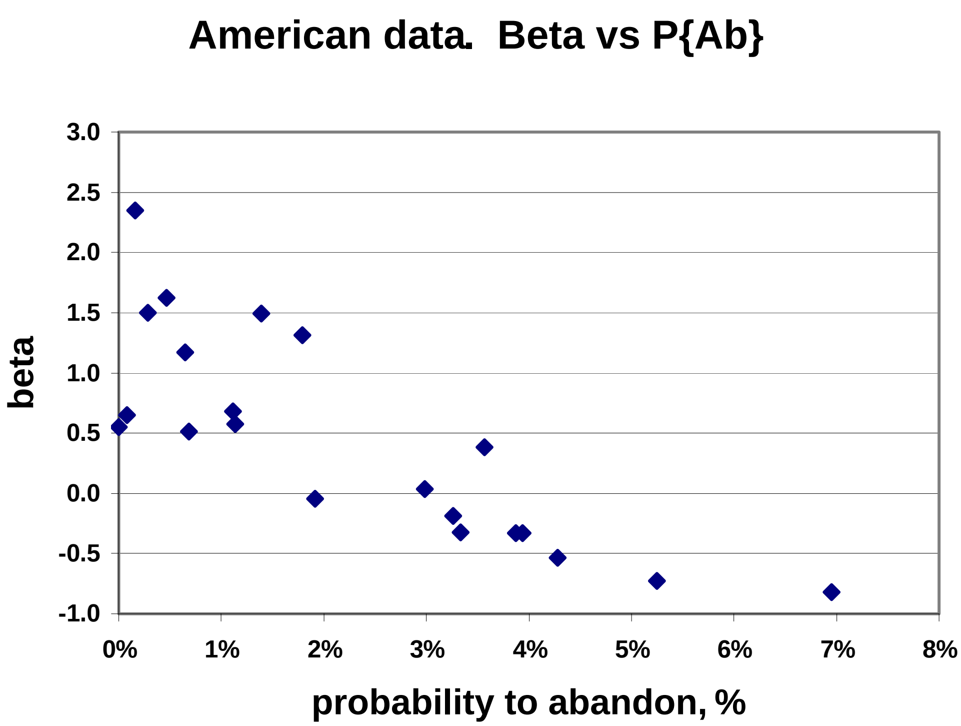
Class 13

QED (QD, ED) Queues
Erlang-A (M/M/n+G) in the QED & ED Regime

- Motivation, via Data & Infinite-Servers;
- QED Erlang-A: Garnett's Theorem;
- The right answer for the wrong reasons - revisited;
- M/M/n+G: Zeltyn's Approximations (QD, ED);
- Rules of Thumb;
- Cost Minimization for Erlang-A (with Zeltyn);
- Constraint-Satisfaction; The 80-20 Rule.

QED Erlang-A: Practical Motivation





QED Erlang-A: Theoretical Motivation

QED staffing: $n \approx R + \beta\sqrt{R}$.

Assume $\boldsymbol{\theta = \mu}$, namely "average service-time" = "average (im)patience".

Recall and Note:

- If $\theta = \mu$, the number-in-system of M/M/n+M has the same distribution of a corresponding M/M/∞ (both are the same Birth&Death process). Formally, in steady-state: $\boldsymbol{L(\mathrm{M/M/n+M}) \stackrel{d}{=} L(\mathrm{M/M/\infty})}$.
- The steady-state distribution of M/M/∞ with parameters λ and μ is **Poisson(R)**, where $R = \lambda/\mu$ (offered-load).
- For R not too small, Poisson(R) is approximately Normal(R,R). Formally: $\boldsymbol{L(\mathrm{M/M/\infty}) \stackrel{d}{\approx} R + Z\sqrt{R}}$, where $\boldsymbol{Z}$ is standard normal.

We now use these facts to estimate the delay-probability for Erlang-A, in which $\theta = \mu$:

$$\mathrm{P}\{W_q(\mathrm{M/M/n+M}) > 0\} \stackrel{\mathrm{PASTA}}{=} \mathrm{P}\{L(\mathrm{M/M/n+M}) \geq n\} \stackrel{\theta=\mu}{=} \mathrm{P}\{L(\mathrm{M/M/\infty}) \geq n\}$$

Standardizing $L \approx R + Z\sqrt{R}$ reveals the QED regime, specifically how square-root staffing yields a non-degenerate delay-probability:

$$\mathrm{P}\{W_q > 0\} \approx \mathrm{P}\left\{Z \geq \frac{n-R}{\sqrt{R}}\right\} \approx 1 - \Phi(\beta)\,.$$

The Erlang-A Queue in the QED-Regime

Theorem (with Garnett & Reiman, 2002)

The following **points of view** are equivalent:

0. QED: $\mathrm{P}\{W_q > 0\} \approx \alpha,$ for some $0 < \alpha < 1$;

1. Manager: $n \approx R + \beta\sqrt{R},$ for some $-\infty < \beta < \infty$;

2. Servers: $\text{Occupancy} \approx 1 - \frac{\beta + \gamma}{\sqrt{n}}$;

3. Customers: $\mathrm{P}\{\mathrm{Ab}\} \approx \frac{\gamma}{\sqrt{n}},$ for some $0 < \gamma < \infty$;

in which case

$$\alpha \;=\; \alpha(\beta, \frac{\mu}{\theta}) \;=\; \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} ,$$

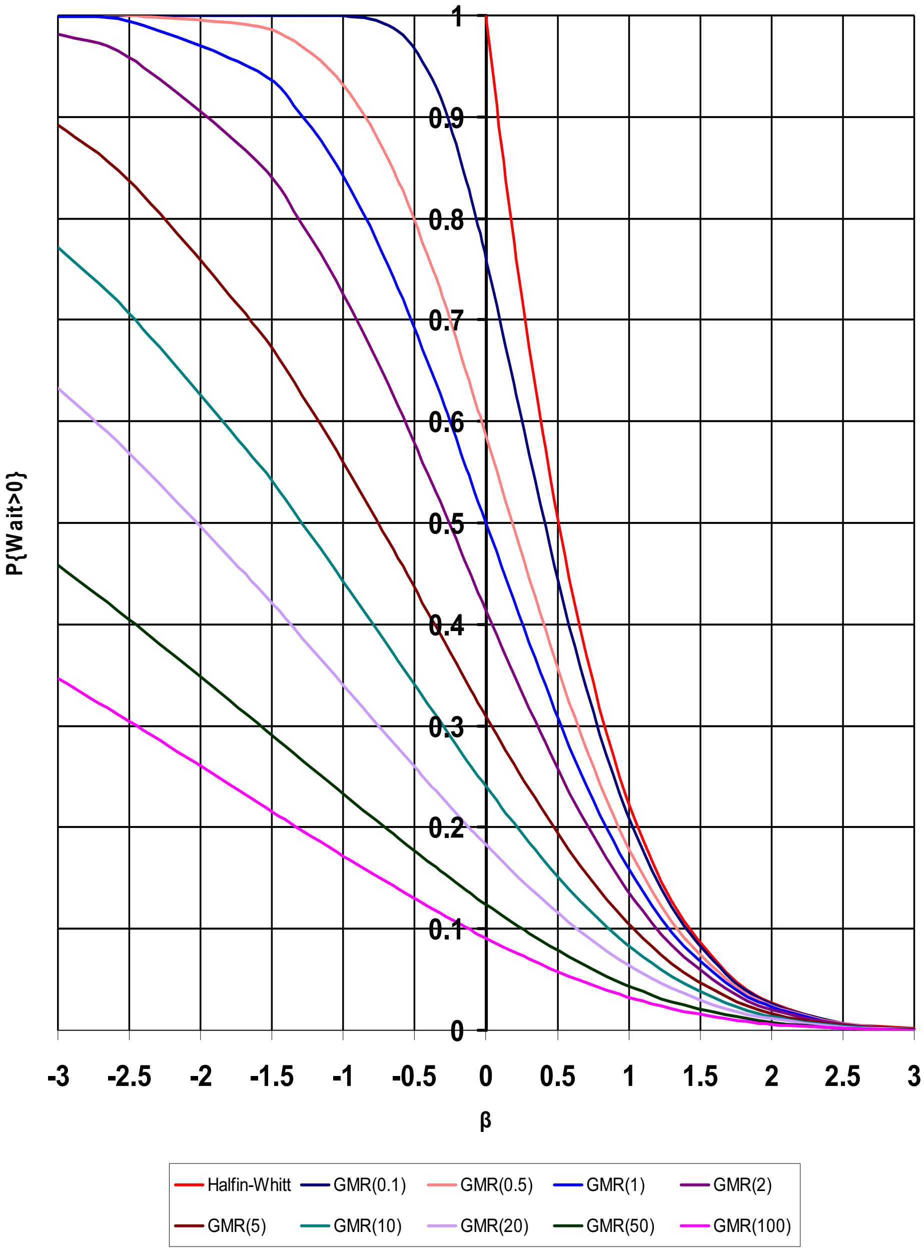
which we call the **Garnett Delay-Function(s)**;

here $\hat{\beta} \;\triangleq\; \beta\sqrt{\frac{\mu}{\theta}},$ and

$$\gamma \;=\; \alpha \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] .$$

Erlang-A: The Garnett Delay-Functions

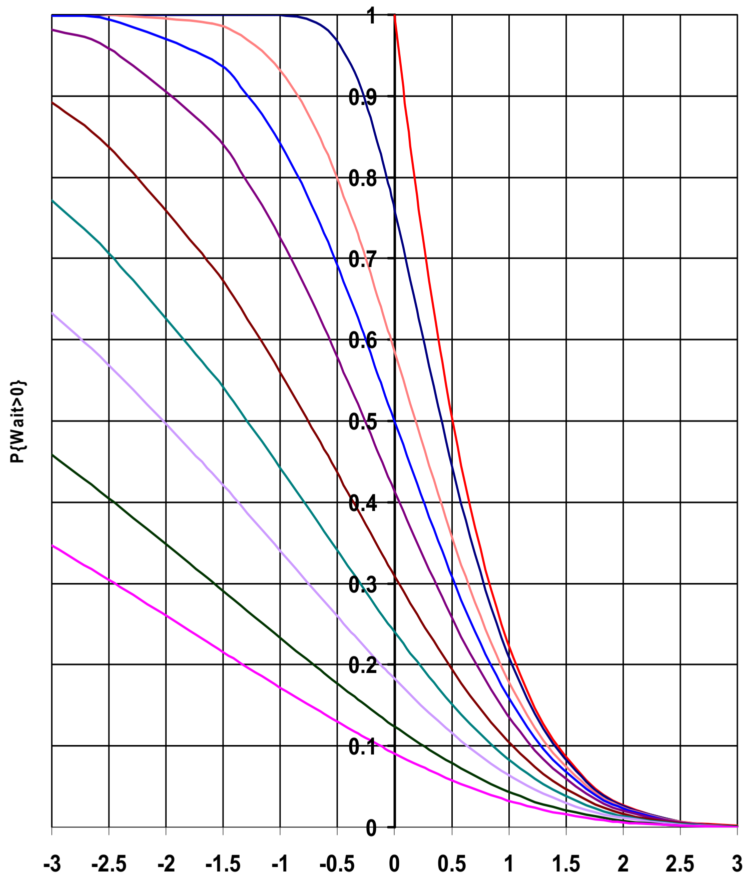
$P\{W_q > 0\}$ vs. the QOS parameter β, for varying patience θ/μ.



GMR(x) describes the asymptotic probability of delay as a function of β when $\theta/\mu = x$. Here, θ and μ are the abandonment and service rate, respectively.

Note: **Erlang-C** = limit of **Erlang-A**, as patience ↑ indefinitely.

Understanding the Garnett Functions



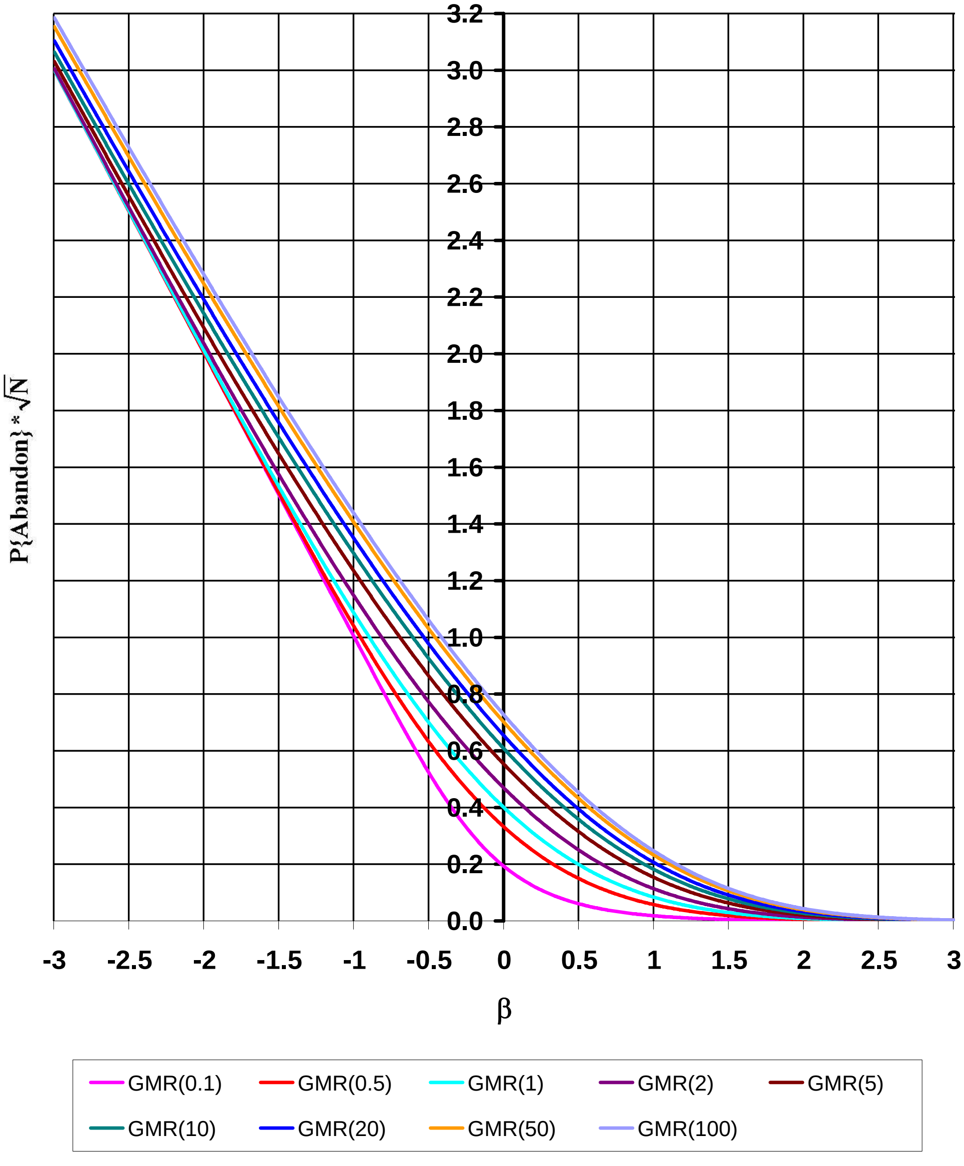
- **Fix a staffing-level** (service-grade) and let patience ↑: then delays ↑; in particular, the Garnett functions ↑ to the Halfin-Whitt function (infinite-patience).

- **Fix a target delay-probability** (service level): then, as impatience ↑, less servers (smaller service-grade) are required to achieve the target (convincing managers to use Erlang-A).

- With $\beta = 0$ $(n = R)$ and $\mu = \theta$, 50% are served immediately. Compare with Erlang-C in which $n = R + 0.5\sqrt{R}$ was required. But there is **no free lunch**: **2%** abandon! (under $n = 400$) see next page.

Erlang-A: % Abandonment

$\% Ab \times \sqrt{n}$ vs. β, for varying (im)patience (θ/μ):



Note the behavior: slope $-\beta$, for (relatively) large negative β and over all (im)patience levels. For an explanation, think **ED**: $n = R + \beta\sqrt{R} = R - \gamma R$; hence $\gamma \approx -\beta/\sqrt{R} \approx -\beta/\sqrt{n}$, and γ is $P\{Ab\}$ in the ED-Regime.

"The Right Answer for the Wrong Reason" - Revisited

If $\beta = 0$, the QED staffing level $n \approx R + \beta\sqrt{R}$ becomes

$$n = R = \frac{\lambda}{\mu} = \lambda \cdot E[S],$$

which is equivalent to the following **deterministic** rule:
Assign a number of agents that equals the offered load.
(Common in stochastic-ignorant operations.)

Erlang-C: queue "explodes".

Erlang-A: Assume $\mu = \theta$. Then $\mathrm{P}\{W_q = 0\} \approx 50\%$.

If $n = 100$, $\mathrm{P}\{\mathrm{Ab}\} \approx 4\%$ (twice the value 2% in the graph - why?), and $\mathrm{E}[W_q] \approx 0.04 \cdot \mathrm{E}[S]$ (why?).

Overall, reasonable (good?) service level, which will in fact improve with scale. For example, with $n = 400$, both $P\{Ab\}$ and $\mathrm{E}[W_q]$ reduce to half their value under $n = 100$ (why?).
(Note: Changes in n go hand in hand with same changes in λ, assuming μ remains fixed.)

The Effect of Patience:
Suppose now $\mu = 0.1 \cdot \theta$ (highly impatient customers).
Via the Garnett Functions, suffices $n = R - \sqrt{R}$ to achieve $\mathrm{P}\{W_q = 0\} \approx 50\%$, but this comes at the cost of somewhat over 10% abandoning, with $n = 100$ (and 5% with $n = 400$); though $\mathrm{E}[W_q]$ decreases to one fourth of the above, assuming μ remains unchanged.

Erlang-A in the QED Regime: Operational Performance Measures

$$\mathrm{P}\{W_q > 0\} \approx \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad \hat{\beta} = \beta\sqrt{\frac{\mu}{\theta}}$$

$$\mathrm{E}\left[W_q | W_q > 0\right] \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1}{\theta\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right]$$

$$\mathrm{P}\{\mathrm{Ab}\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}$$

$$\mathrm{P}\{Ab | W_q > 0\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right]$$

$$\mathrm{P}\left\{\frac{W_q}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}} \,\middle|\, W_q > 0\right\} \approx \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{\frac{\theta}{\mu}} \cdot t\right)}{\bar{\Phi}(\hat{\beta})}$$

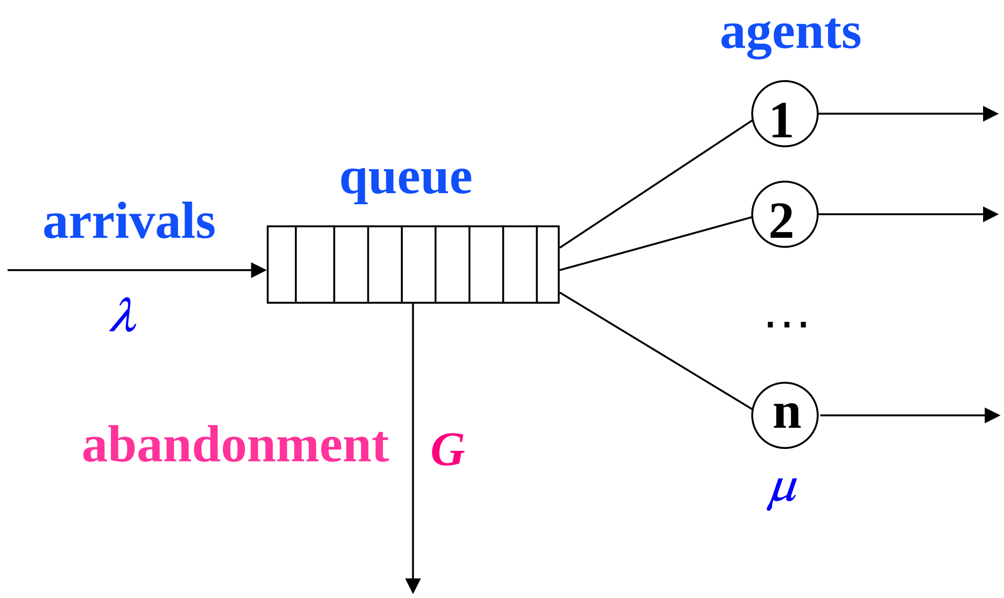
$$\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, \frac{W_q}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{\theta}{\mu}}\right) - \hat{\beta}\right]$$

$$E\left[\frac{W_q}{\mathrm{E}[S]} \,\middle|\, Ab\right] \approx \frac{1}{\sqrt{n}} \cdot \frac{1}{2}\sqrt{\frac{\mu}{\theta}} \cdot \left[\frac{1}{h(\hat{\beta}) - \hat{\beta}} - \hat{\beta}\right]$$

Here

$$\bar{\Phi}(x) = 1 - \Phi(x)\,,$$
$$h(x) = \phi(x)/\bar{\Phi}(x)\,, \quad \text{hazard rate of } N(0,1).$$

M/M/n+G in the QED Regime



Density of (im)patience G: $g = \{g(x), x \geq 0\}$.

Assume $g_0 \triangleq g(0) > 0$.

QED regime: $n \approx R + \beta\sqrt{R}$.

QED approximations: Use the Erlang-A formulae (from the previous page), **substituting g_0 instead of θ**.

How to estimate g_0? As $\hat{\theta}$ in Erlang-A!

Why? Recall **Erlang-A**: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$ used for estimating θ (either via $\hat{\theta}$ = [#Abandoning] / [Total Waiting Time]; or by regression of half-hours' [%Abandoning] over [Expected-Waits]).

M/M/n+G: It turns out that, in the QED regime:

$$\mathrm{P}\{\mathrm{Ab}\} \approx g_0 \cdot \mathrm{E}[W_q] \ .$$

Hence, one estimates g_0 exactly as $\hat{\theta}$ in Erlang-A.

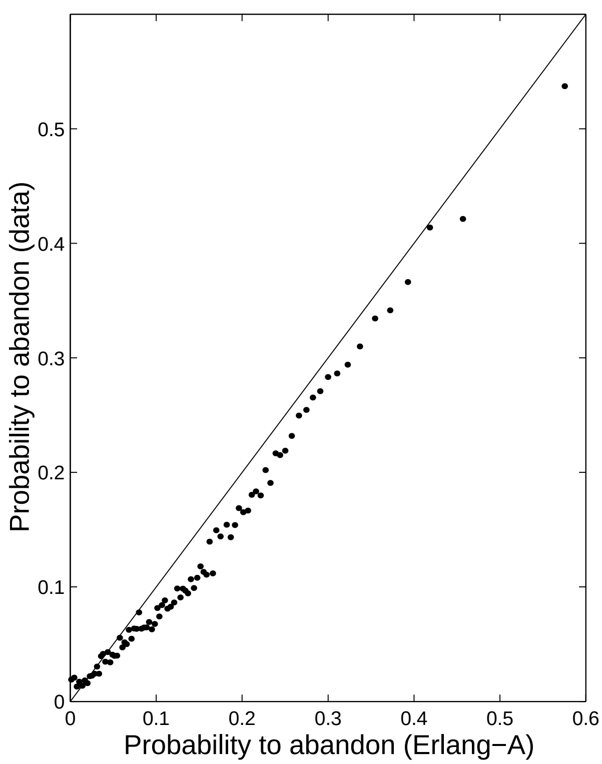
Erlang-A: Fitting a Simple Model to a Complex Reality

Question: Can one **usefully** apply the Erlang-A model to systems with **non-exponential** patience?

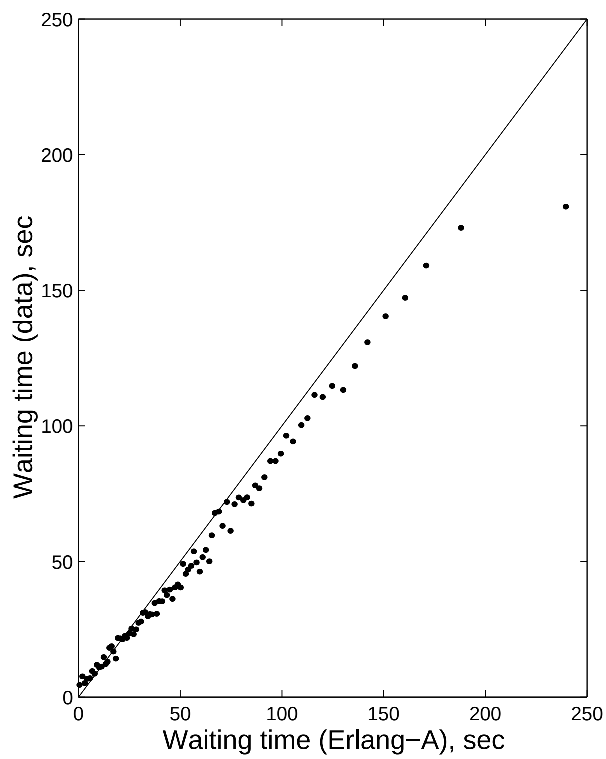
YES!

Erlang-A Formulae vs. Data Averages (Israeli Bank)

$P\{Ab\}$

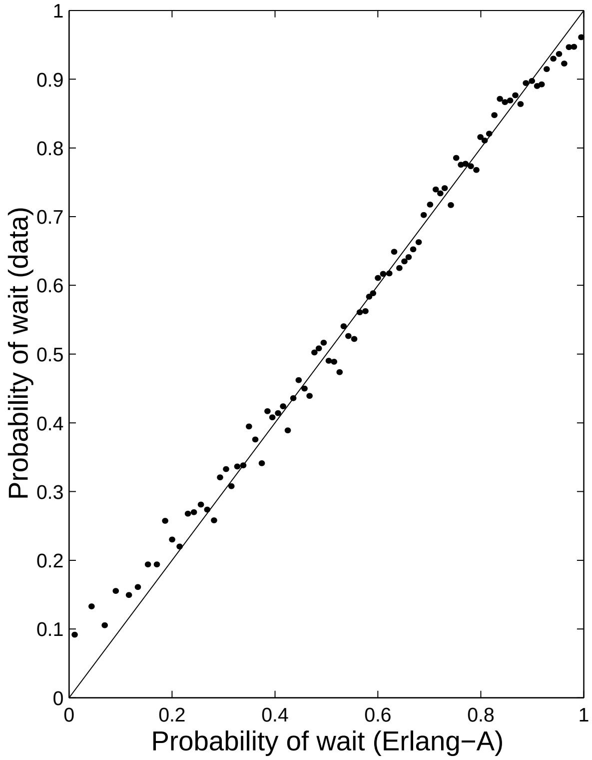


$E[W_q]$



Erlang-A: Fitting a Simple Model to a Complex Reality II

$\mathrm{P}\{W_q > 0\}$

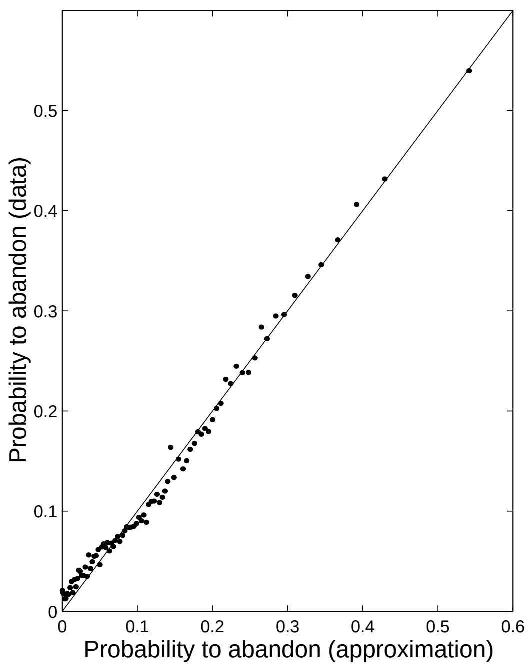


Summary:

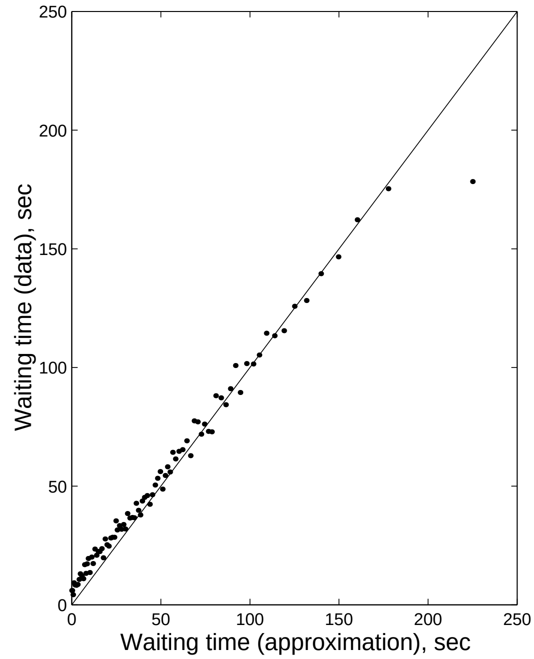
- Points: Hourly data (averages) vs. Erlang-A predictions;
- Formulae with continuous n (special-functions) used to account for non-integer n;
- **Patience estimated via $\mathrm{P}\{\mathrm{Ab}\}/\mathrm{E}[W_q]$;**
- **Erlang-A estimates provide close upper bounds**.

Fitting Erlang-A Approximations

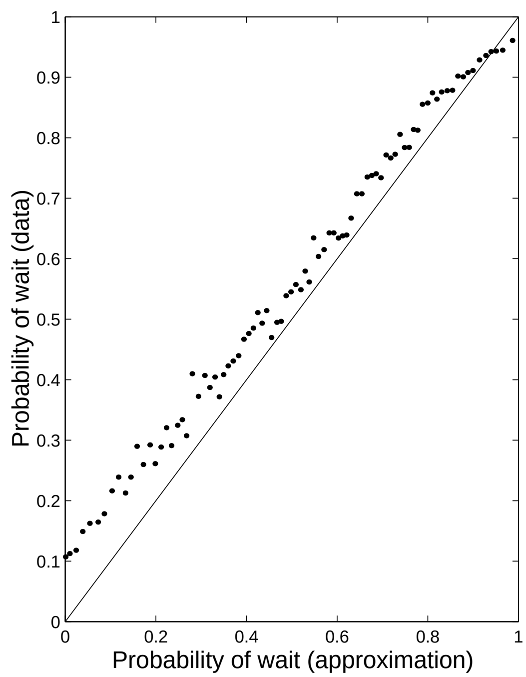
P{Ab}



$\mathrm{E}[W_q]$



$\mathrm{P}\{W_q > 0\}$



Quality-Driven M/M/n+G (QD)

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

Staffing level:

$$n \approx R \cdot (1 + \delta)\,, \qquad \delta > 0\,.$$

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

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

- Average wait of delayed customers:

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

- Linear relation between $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W_q]$:

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

- Asymptotic distribution of wait:

$$\mathrm{P}\left\{ \frac{W_q}{\mathrm{E}(S)} > \frac{t}{n} \;\middle|\; W_q > 0 \right\} \;\sim\; e^{-(1-\rho)t}\,, \qquad \rho = \frac{\lambda}{n\mu}\,.$$

Comparison with QED: Simpler here, hence worth having. Often, order $1/n$ replaces $1/\sqrt{n}$ (though, note conditioning).

Efficiency-Driven M/M/n+G (ED)

Let γ be a QOS parameter, $0 < \gamma < 1$.
Assume $G(x) = \gamma$ has a unique solution $x^* = G^{-1}(\gamma)$, at which $g(x^*) > 0$.

Staffing level:

$$n \approx R \cdot (1 - \gamma)\,, \qquad \gamma > 0\,.$$

- $\mathrm{P}\{\boldsymbol{W_q > 0}\} \approx \mathbf{1}$.
- Abandonment-Probability converges to:

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

- Offered-Wait converges to x^*:

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

- Waiting distribution (asymptotically):

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

  where G^* is the distribution of $\min(x^*, \tau)$, namely

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

Operational Regimes: Rules-of-Thumb

Assume that the **Offered-Load** R is not too small (more than several 10's for QED, more than 100 for ED and QD).

ED regime: $n \approx R - \delta R, \qquad 0.1 \le \delta \le 0.25\,.$

- Essentially **all** customers are delayed;
- %Abandoned $\approx$ δ (10-25%);
- Average-wait $\approx$ 30 seconds - 2 minutes.

QD regime: $n \approx R + \gamma R, \qquad 0.1 \le \gamma \le 0.25\,.$
Essentially **no** delays.

QED regime: $n \approx R + \beta\sqrt{R}, \qquad -1 \le \beta \le 1\,.$

- %Delayed between 25% and 75%;
- %Abandoned is 1-5%;
- Average wait is one-order less than average service-time (eg. seconds vs. minutes).

Operational Regimes: Performance

Assume that **offered load** R is not small (more than several 10's for QED, more than 100 for ED and QD).

ED regime: $n \approx R - \delta R, \quad 0.1 \le \delta \le 0.25\,.$

- Essentially **all** customers are delayed;
- %Abandoned $\approx \delta$ (10-25%);
- Average wait $\approx$ 30 seconds - 2 minutes.

QD regime: $n \approx R + \gamma R, \quad 0.1 \le \gamma \le 0.25\,.$
Essentially **no** delays.

QED regime: $n \approx R + \beta\sqrt{R}, \quad -1 \le \beta \le 1\,.$

- %Delayed between 25% and 75%;
- %Abandoned is 1-5%;
- Average wait is one-order less than average service time (seconds vs. minutes).

Economies of Scale (EOS)

For our purpose:

Economies of Scale (EOS) prevail if load-increase by a factor m "requires" staffing-increase by **less** than m.

In what sense **"Requires"** ?

- **Achieve** management goal(s) (**constraint satisfaction**), or
- **Optimize** management goal(s) (**optimize cost / profit**).

Constraint Satisfaction **easier to formulate** (simpler data) and **solve** (hence more prevalent); but, as we saw (recall the 80:20 rule), Performance Optimization is easier to **grasp**.

Pooling QD Erlang-A's

Pool m identical service operations (call centers) with parameters $(\lambda, \mu, n, \theta)$.
Sustain the same QD operational regime, namely staffing levels:
$n \approx R + \delta R\,, \qquad \delta = 0.25\,,$ for concreteness.

Use 4CallCenters to calculate the following:

E[*S*]=6 min, E[τ]=9 min

λ/hr	n	Occupancy	P{Ab}	E[W_q]	P{$W_q > 0$}
8	1	57.6%	28.0%	2:31	57.6%
32	4	71.5%	10.6%	0:58	42.5%
128	16	78.0%	2.5%	0:14	23.4%
512	64	79.8%	0.2%	0:01	4.9%
2,048	256	80.0%	0.0%	0:00	0.0%
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	∞	**80%**	**0%**	**0:00**	**0%**

Occupancy converges to $1/(1+\delta)$; here $1/1.25 = 80\%$.

EOS: Performance Measures improve at an exponential rate.

Pooling ED Erlang-A's

$n \;\approx\; R - \gamma R\,, \qquad \gamma = 1/6\,.$

E[S]=6 min, E[τ]=9 min

λ/hr	n	Occupancy	P{Ab}	E[W_q]	P{$W_q > 0$}
12	1	73.4%	38.8%	3:29	73.4%
48	4	89.8%	25.2%	2:16	75.6%
192	16	97.5%	18.7%	1:41	85.4%
768	64	99.8%	16.8%	1:31	97.2%
3,072	256	100.0%	16.7%	1:30	100.0%
↓	↓	↓	↓	↓	↓
∞	∞	**100%**	**16.7%**	**1:30**	**100%**

P{Ab} and **E[W_q]** converge as is:

$\text{P}\{\text{Ab}\} \to \gamma; \quad \text{E}[W_q] \to \gamma \cdot \text{E}[\tau]$.

Thus, in the ED-Regime, there is **no EOS** for large n.

Pooling QED Erlang-A's

$n \approx R + \beta\sqrt{R}, \qquad \beta = 0.$

E[S]=6 min, E[τ]=9 min

λ/hr	n	Occupancy	P{Ab}	E[W_q]	P{$W_q > 0$}
10	1	66.4%	33.6%	3:02	66.4%
40	4	82.4%	17.6%	1:35	60.9%
160	16	91.1%	8.9%	0:48	58.0%
640	64	95.5%	4.5%	0:24	56.5%
2,560	256	97.8%	2.2%	0:12	55.8%
↓	↓	↓	↓	↓	↓
∞	∞	**100%**	**0%**	**0:00**	**55.1%**

Delay probability converges to the appropriate Garnett function:

$$\mathrm{P}\{W_q > 0\} \to \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} = \left[1 + \sqrt{\frac{2}{3}}\right]^{-1} \approx 0.551\,.$$

EOS: P{Ab} and E[W_q] improve at the rate of $1/\sqrt{n}$.

EOS and Constraint Satisfaction

Assume service and abandonment rates are as in the previous example: $\mathrm{E}[S] = 6$ min; $\mathrm{E}[\tau] = 9$ min. Playing with 4CC yields:

ED regime:
"Loose" constraint: $\mathbf{P\{Ab\} \le 10\%}$.
$R = 100 \Rightarrow n = 91;$ $\quad$ $R = 400 \Rightarrow n = 361.$
Almost no EOS! Use $n \approx 90\% \cdot R$ $\quad (= (1-\gamma) \cdot R,\ \ \gamma \approx P\{Ab\})$.

QED regime:
"Moderate" constraint: $\mathbf{P\{Ab\} \le 2\%}$.
$R = 100 \Rightarrow n = 105;$ $\quad$ $R = 400 \Rightarrow n = 399.$
Saved more than 20 agents: 399 instead of $420 = 4 \times 105$.
$\beta = 0.5$ for $R = 100$, $\ \beta = -0.05$ for $R = 400$.
Why EOS? With β fixed, $\mathrm{P}\{\mathrm{Ab}\} \approx c(\beta)/\sqrt{n}$. Thus, $n \uparrow$ implies $P\{Ab\} \downarrow$. Consequently, with $n \uparrow$, $\beta \downarrow$ in order to achieve a given $P\{Ab\}$

QD regime:
"Strict" constraint: $\mathbf{P\{Ab\} \le 0.1\%}$.
$R = 100 \Rightarrow n = 119;$ $\quad$ $R = 400 \Rightarrow n = 432.$
More than 45 agents saved: 432 vs. $4 \times 119 = 476$.
$\delta = 0.19$ for $R = 100$, $\ \delta = 0.08$ for $R = 400$.
Why EOS? With δ fixed, $\mathrm{P}\{\mathrm{Ab}\}$ decreases exponentially in n, etc.

Recall: Cost Minimization in Erlang-C

(With Borst and Reiman, 2004.)

(Equivalently, Profit Maximization, if Revenues proportional to λ.)

$\boldsymbol{\text{Cost} = c \cdot n + d \cdot \lambda \mathbf{E}[W_q]}$,

c – cost of staffing;

d – cost of delay.

Erlang-C: Optimal staffing level:

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

$\beta^*(r)$ = optimal service grade (QOS), independent of λ:

$$\beta^*(r) = \arg \min_{0<y<\infty} \left\{ y + \frac{r \cdot P_w(y)}{y} \right\} ,$$

where (recall the Halfin-Whitt function)

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

Very good approximation:

$$\beta^*(r) \approx \left(\frac{r}{1 + r(\sqrt{\pi/2} - 1)} \right)^{1/2}, \quad 0 < r < 10,$$

$$\approx \left(2 \ln \frac{r}{\sqrt{2\pi}} \right)^{1/2}, \qquad r \geq 10.$$

Erlang-A: Staffing via Optimization

(with Zeltyn, 2006)

We study "Minimize **Costs (Staffing + Waiting)**". Why?

- Comparison easy against Erlang-C;
- W.L.O.G.: $P\{Ab\} = \theta \cdot E[W_q]$ reduces profit- to cost-optimization. Specifically, find n^* that max. average profit per time-unit:

  $$R_s \cdot \lambda \cdot [1 - \mathrm{P}_n\{\mathrm{Ab}\}] - [C_s \cdot n + C_w \cdot \mathrm{E}_n[W_q] \cdot \lambda + C_a \cdot \mathrm{P}_n\{\mathrm{Ab}\} \cdot \lambda],$$

  where R_s is the **revenue** from a single service. This reduces to $c = C_s$ and $d = (R_s \cdot \theta + C_w + C_a \cdot \theta)$ in the following:

Minimize **Cost** $= \boldsymbol{c \cdot n + d \cdot \lambda \mathrm{E}[W_q]}$; here, as before,

c – Staffing Cost;

d – Delay Cost;

$r = d/c$.

Erlang-A. Optimal staffing level:

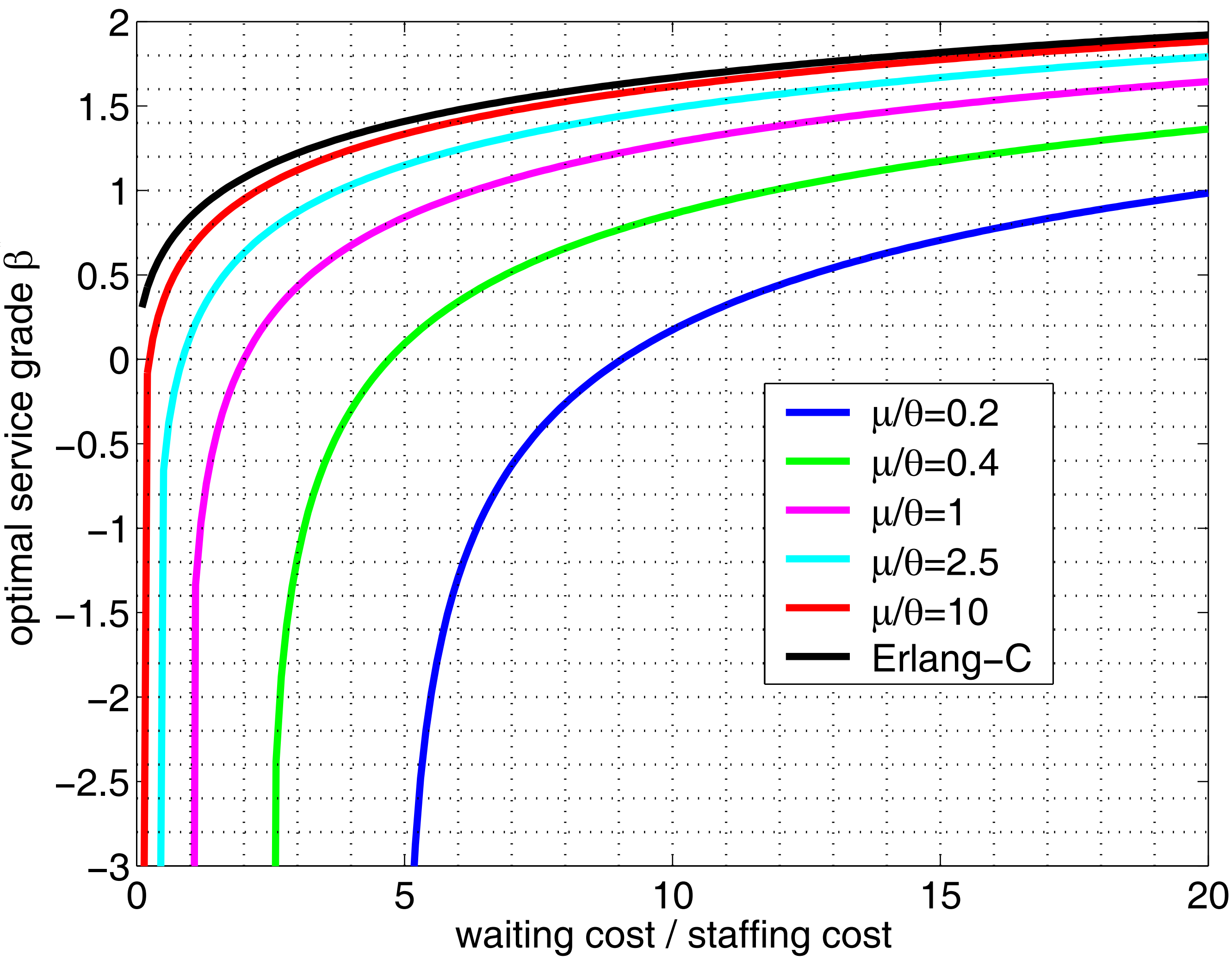
$$n^* \approx R + \beta^*(r; s)\sqrt{R}, \qquad s = \sqrt{\mu/\theta} \quad ,$$

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

where (recall the Garnnett functions)

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

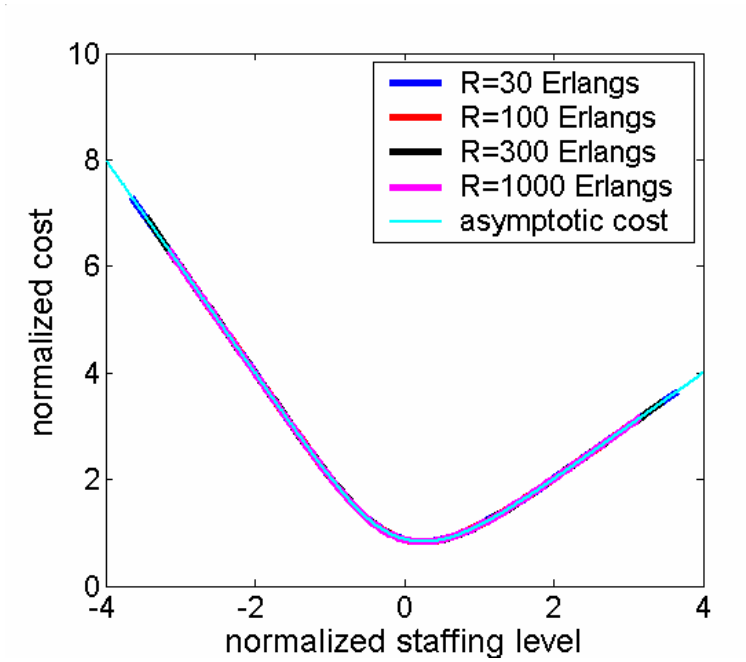
Erlang-A: Optimal Service Grade β^* (QOS)



- As $\theta \downarrow 0$, $\beta^*(r;\sqrt{\mu/\theta})$ increases to $\beta^*(r)$ (Erlang-C = M/M/n).
- $r < \theta/\mu$ implies that "no-service" ($n = 0$) is optimal. Why?
  $d \cdot E[\tau] < c \cdot E[S]$: cheaper to let abandon than to serve!
- $r \le 20 \Rightarrow \beta^* < 2$; $r \le 500 \Rightarrow \beta^* < 3$, as in Erlang-C.
- Numerical tests exhibit **remarkable** accuracy & robustness.

Erlang-A: Actual Cost vs. Asymptotic Cost

$\mu = 1, \theta = 1/3$



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

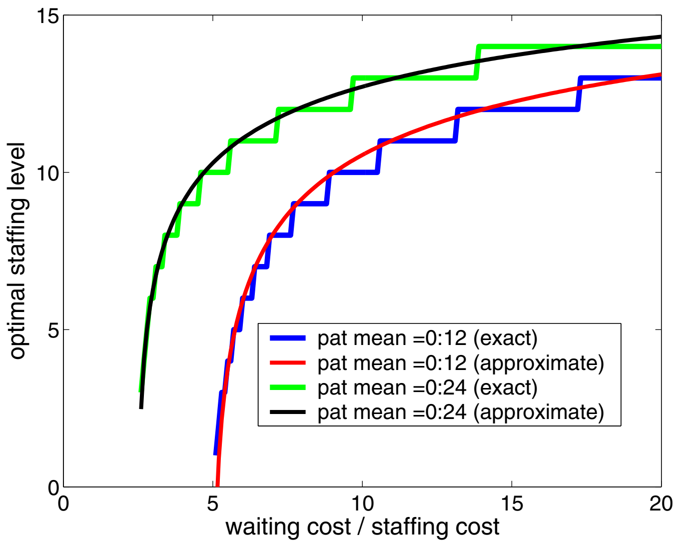
Normalized cost $= (\text{cost} - cR)/\sqrt{R}$;

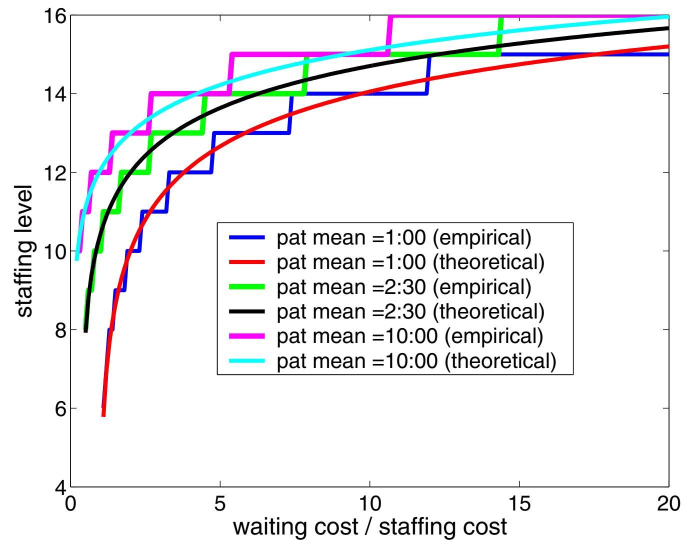
Asymptotic cost $= c \cdot y + d \cdot P_w(y; s) \cdot s \cdot [h(ys) - ys]$,

where y = QED service grade.

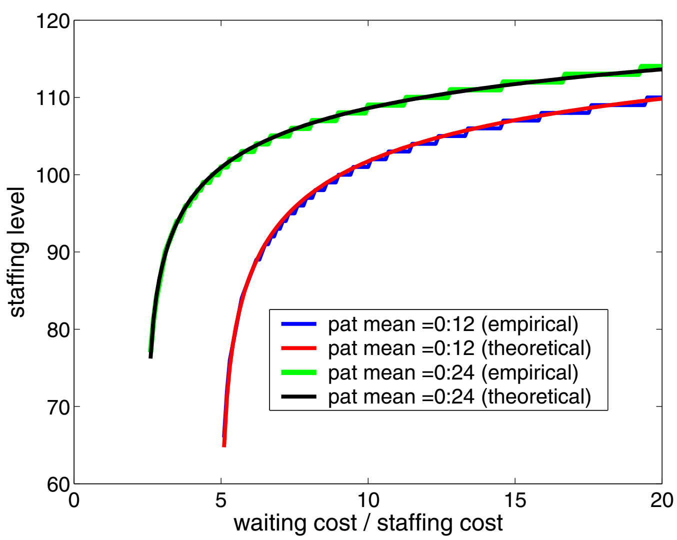
Erlang-A: Optimal Staffing

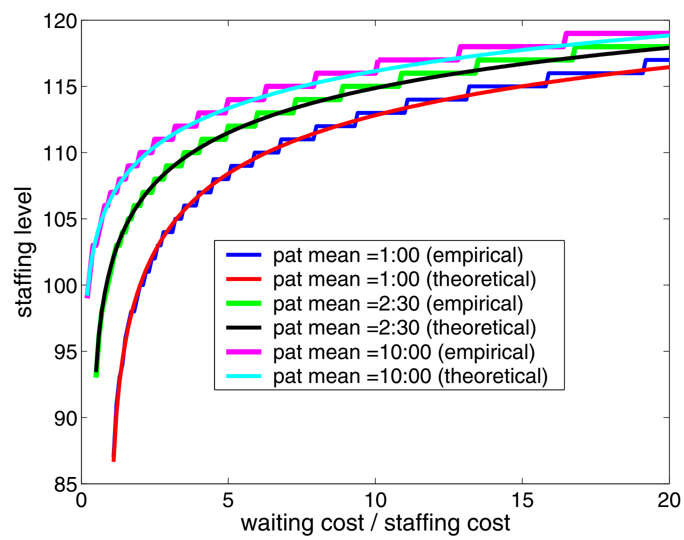
$\lambda = 10, \ \mu = 1$





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

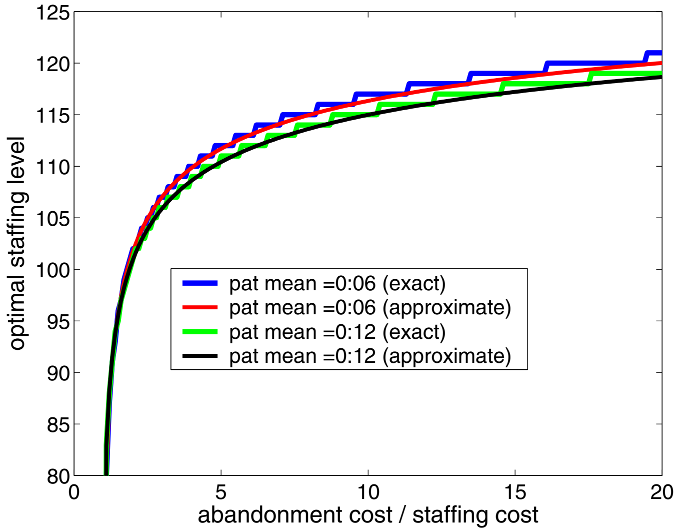


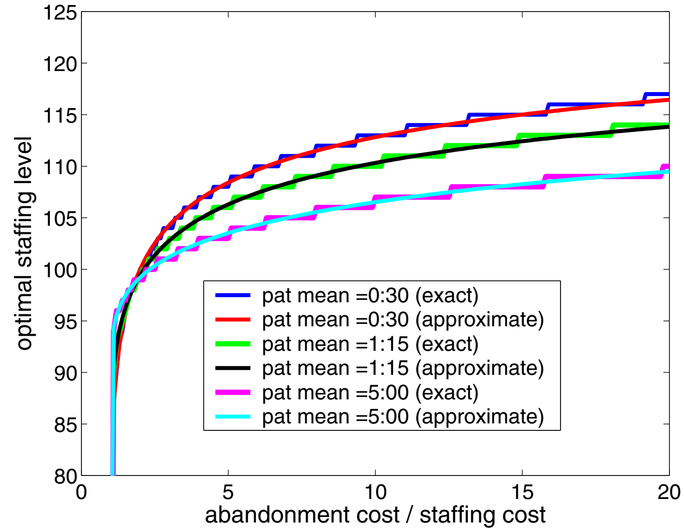


M/M/n+G: Optimal Staffing

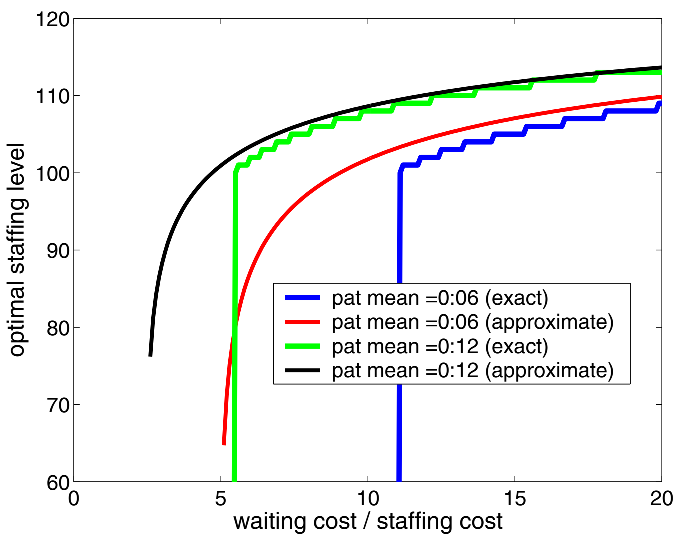
Uniformly Distributed Patience.

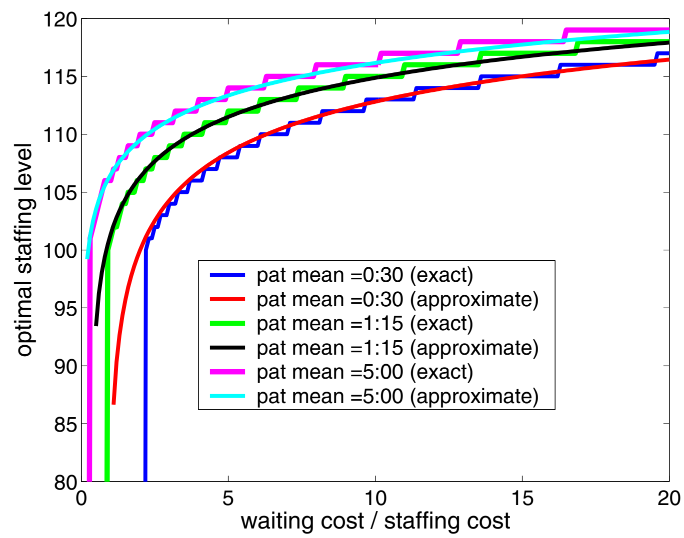
$$\text{Cost} = c \cdot n + d \cdot \lambda \text{P}\{\text{Ab}\}$$





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





The 80-20 Rule: Cost Optimization and Constraint Satisfaction

Prevalent standard:
at least 80% of customers are served within 20 seconds.
Call center: $\lambda = 6000/\text{hr}$, E[S]=4 min (R=400); E[τ]=6 min.
4CallCenters: $n = 394$ agents required $\Rightarrow$ $\beta^* = -0.3$.

According to the graph, $\boldsymbol{d/c \approx 1}$: costs of customers' time and servers' time are nearly equal.

What if $\boldsymbol{d/c = 5}$? $\beta^* = 1 \Rightarrow n^* = 420$;
82.3% served immediately; 98.9% within 20 seconds.
(Comparable Erlang-C: $n^* = 428$, corresponding to $d/c = 10$.)

$\boldsymbol{\theta/\mu = 2/3}$