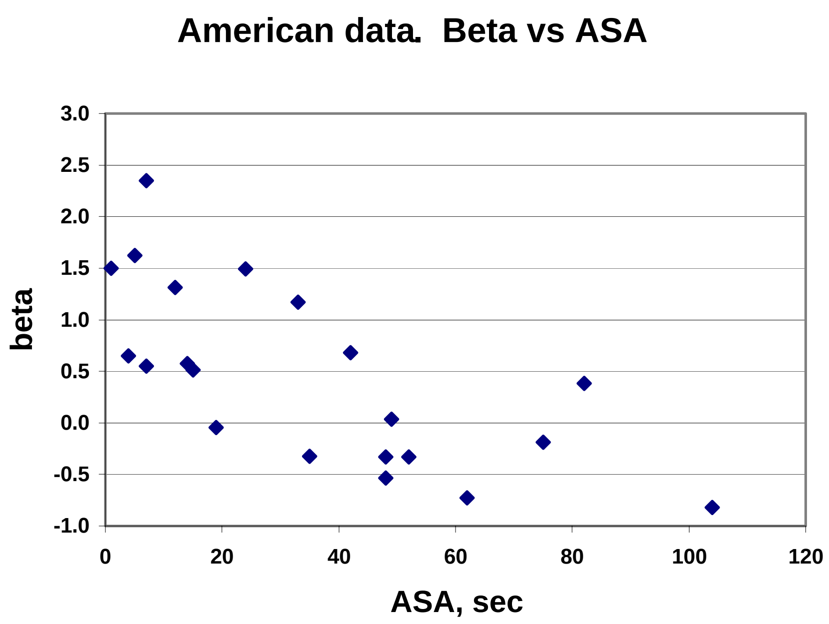
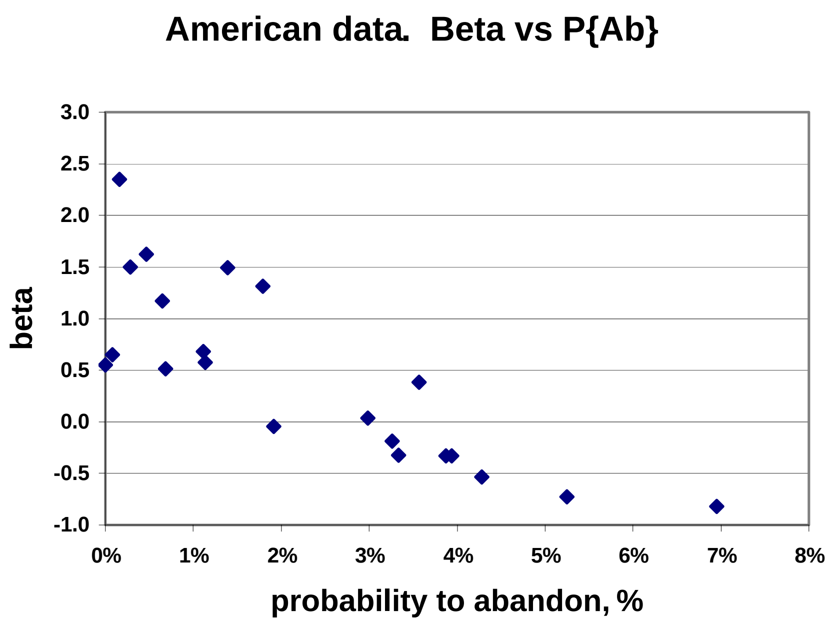
Empirical Service Grade (Beta)





QED Relevance: Wharton CC-Forum

6/13/00 - Tue

Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
Total	**20,577**	**19,860**	**~3.0%**	**30**	**307**	**95.1%**	**85.4%**	**222.7**	**234.6**	**95.0%**
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	866	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	2l	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1.286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	**1,364**	**1,338**	**1.9%**	**33**	**296**	**99.0%**	84.1%	**222.5**	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	**1,174**	**1,160**	**1.2%**	**10**	**302**	**95.5%**	93.6%	**203.8**	209.8	97.1%
13:00	**1,018**	**999**	**1.9%**	**9**	**314**	**95.4%**	91.2%	**182.9**	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	86.0%	**206.1**	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	615	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	89.1%	5.8	1.4	416.2%

Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%

Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience=E(S)		
1600	100	96%	0:09	50%

Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience=E(S)		
1600	100	96%	0:09	50%
AND		could have		**%Abandon**
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

Operational Aspects of Impatience

Recall earlier Q, E and QED Scenarios (E(S) = 3:45):

λ/hr	N	OCC	ASA	% Wait = 0
1599	100	99.9%	59:33	1%
1599	105	95.2%	0:23	51%
1600	100	100%	infinity	0%
BUT	with	Patience=E(S)		
1600	100	96%	0:09	50%
AND		could have		**%Abandon**
1600	100	97.3%	0:23	2.7 %
1600	95	98.4%	0:23	6.5%
1800	105	97.7%	0:23	3.4%

QED with **(Im)patient** Customers:

The "fittest" survive and wait less – much less!

Erlang-A: Erlang-C with Exponential Patience / **A**bandonment

Downloadable implementation: 4CallCenters(.com)

Asymptotic Operational Regimes

Example of Half-Hour 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

Asymptotic Operational Regimes

Efficiency-Driven (ED) regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**

- 100% occupancy;
- high P{Ab};
- considerable ASA;
- $\mathbf{P\{W > 0\} \approx 1}$.

Offered load

$$R_{ED} \triangleq \frac{\lambda}{\mu} = 1061 : \frac{1800}{306} = 180.37\,.$$

Definition:

$$n = R_{ED} \cdot (1 - \gamma) \qquad \gamma > 0.$$

In our case, *service grade*

$$\gamma = 1 - \frac{n}{R_{ED}} = 1 - \frac{163.4}{180.37} = 0.094 \approx \mathrm{P}\{\mathrm{Ab}\}\,.$$

- This case is similar to traditional queues in heavy traffic;
- See recent papers of Whitt (2004).

Quality-Driven (QD) regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**

- Occupancy far below 100%;
- negligible P{Ab};
- very small ASA;
- $\mathbf{P\{W > 0\} \approx 0}$.

Offered load

$$R_{QD} \;=\; \frac{\lambda}{\mu} \;=\; 615 : \frac{1800}{328} \;=\; 112.07\,.$$

Definition:

$$n \;=\; R_{QD} \cdot (1+\gamma) \qquad \gamma > 0.$$

Service grade

$$\gamma \;=\; \frac{n}{R_{QD}} - 1 \;=\; \frac{135}{112.07} - 1 \;=\; 0.205\,.$$

Quality and Efficiency-Driven (QED) regime

Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**

- High occupancy, but not 100%;
- small P{Ab} and ASA;
- $\mathbf{P\{W > 0\} \approx \alpha}$, $\quad 0 < \alpha < 1.$

$$R_{QED} \;=\; \frac{\lambda}{\mu} \;=\; 1212 : \frac{1800}{304} \;=\; 204.69\,.$$

Definition:

$$n \;=\; R_{QED} + \beta\sqrt{R_{QED}}\,, \qquad -\infty < \beta < \infty\,.$$

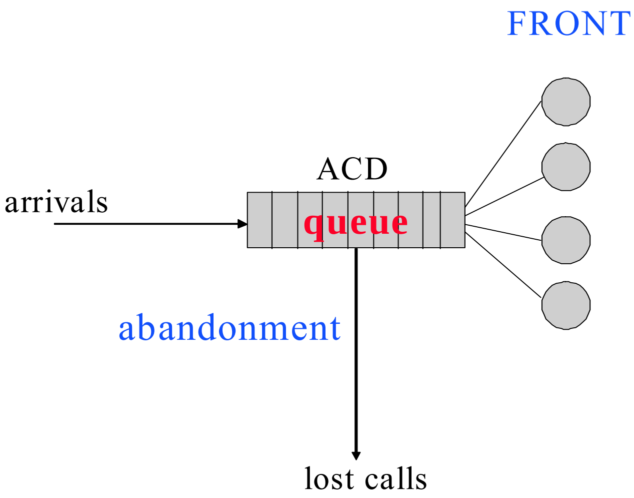
Service grade

$$\beta \;=\; \frac{n - R_{QED}}{\sqrt{R_{QED}}} = \frac{206.1 - 204.69}{\sqrt{204.69}} \;=\; 0.10\,.$$

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

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

Erlang-A (with G-Patience): M/M/N+G



QED Theorem (Garnett, M. and Reiman '02; Zeltyn '03)

Consider a sequence of M/M/N+G models, **N=1,2,3,…**

Then the following **points of view** are equivalent:

- **QED** $\quad \%\{\text{Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1;$
- **Customers** $\quad \%\{\text{Abandon}\} \approx \frac{\gamma}{\sqrt{N}}, \qquad 0 < \gamma;$
- **Agents** $\quad \text{OCC} \approx 1 - \frac{\beta + \gamma}{\sqrt{N}} \qquad -\infty < \beta < \infty;$
- **Managers** $\quad N \approx R + \beta\sqrt{R}, \qquad R = \lambda \times \text{E(S)}$ not small;

QED performance (ASA, ...) is easily computable, all in terms of β (the square-root ~~safety~~ staffing level) – see later.

Covers also the Extremes:

$\alpha = 1$: N = R - γ R **Efficiency-driven**

$\alpha = 0$: N = R + γ R **Quality-driven**

QED Approximations (Zeltyn)

λ – arrival rate,

μ – service rate,

N – number of servers,

G – patience distribution,

$\boldsymbol{g_0}$ – patience density at origin $\quad (\boldsymbol{g_0 = \theta}$, if $\exp(\boldsymbol{\theta}))$.

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

$$\mathrm{P}\{\mathrm{Ab}\} \approx \frac{1}{\sqrt{\boldsymbol{N}}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] \cdot \left[\sqrt{\frac{\mu}{g_0}} + \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

$$\mathrm{P}\left\{W > \frac{T}{\sqrt{N}}\right\} \approx \boxed{\left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}} \cdot \frac{\bar{\Phi}\left(\hat{\beta} + \sqrt{g_0\mu} \cdot T\right)}{\bar{\Phi}(\hat{\beta})},$$

$$\mathrm{P}\left\{\mathrm{Ab}\,\middle|\, W > \frac{T}{\sqrt{N}}\right\} \approx \frac{1}{\sqrt{\boldsymbol{N}}} \cdot \sqrt{\frac{g_0}{\mu}} \cdot \left[h\left(\hat{\beta} + \sqrt{g_0\mu} \cdot T\right) - \hat{\beta}\right].$$

Here

$$\hat{\beta} = \beta\sqrt{\frac{\mu}{g_0}}$$

$$\bar{\Phi}(x) = 1 - \Phi(x)\,,$$

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

- Generalizing Garnett, M., Reiman (2002) (Palm 1943–53)
- No Process Limits

Efficiency-Driven Approximations

(Zeltyn; Whitt)

$G =$ (Im)Patience distribution

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

Assume the equation

$$G(x) = \gamma$$

has a unique solution x^*.

Then

$$\begin{aligned}
\mathrm{P}\{\mathrm{Ab}\} &\approx \gamma \quad \text{(insensitive to } G\text{)} \\
\mathrm{P}\{W > T\} &\approx \begin{cases} 1 - G(T), & T < x^* \\ 0, & T > x^* \end{cases}, \\
\mathrm{P}\{\mathrm{Ab} \mid W > T\} &\approx \gamma - G(T)\,, \quad 0 \le T < x^*\,.
\end{aligned}$$

- **Derivation**: Laplace Method, based on Baccelli & Hebuterne (1981)
- Towards Dimensioning (with Borst, Reiman)

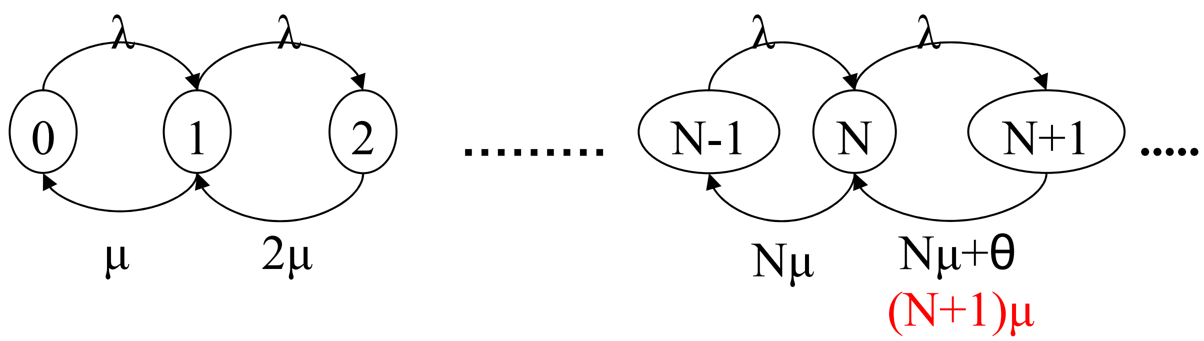
Erlang-A: Moderate (Im)patience

- M/M/N + M queue, with

  service rate μ equals θ abandonment rate

- L_t: number-in-system at time t (Birth & Death)

- For **any** N, transition-rates for $\{L_t, t \geq 0\}$:



Note: The same transition rates as M/M/∞

Square-Root Staffing: Motivation

$$P\{W_q(M/M/N+M) > 0\} \underset{PASTA}{=}$$

$$P\{L(M/M/N+M) \geq N\} \underset{\theta=\mu}{=}$$

$$P\{L(M/M/\infty) \geq N\}$$

Fact: $L(M/M/\infty) \sim \text{Poisson}(R)$; $\quad R = \lambda/\mu \quad$ offered load

Square-Root Staffing: Motivation

$$P\{W_q(M/M/N+M)>0\} \underset{PASTA}{=}$$

$$P\{L(M/M/N+M)\geq N\} \underset{\theta=\mu}{=}$$

$$P\{L(M/M/\infty)\geq N\}$$

Fact: $L(M/M/\infty) \sim \text{Poisson}(R); \quad R=\lambda/\mu$ offered load

For R not too small:

$$\mathrm{L(M/M/\infty)} \overset{\mathrm{d}}{\approx} \mathrm{Normal(R,R)} \overset{\mathrm{d}}{=} \mathrm{R} + \mathrm{Z}\sqrt{\mathrm{R}}$$

$$\Rightarrow \qquad P\{W_q>0\} \approx P\left\{Z \geq \frac{N-R}{\sqrt{R}}\right\} = 1-\phi\left(\frac{N-R}{\sqrt{R}}\right)$$

Square-Root Staffing: Motivation

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

Fact: $L(M/M/\infty) \sim \text{Poisson}(R)$; $R=\lambda/\mu$ offered load

For R not too small:

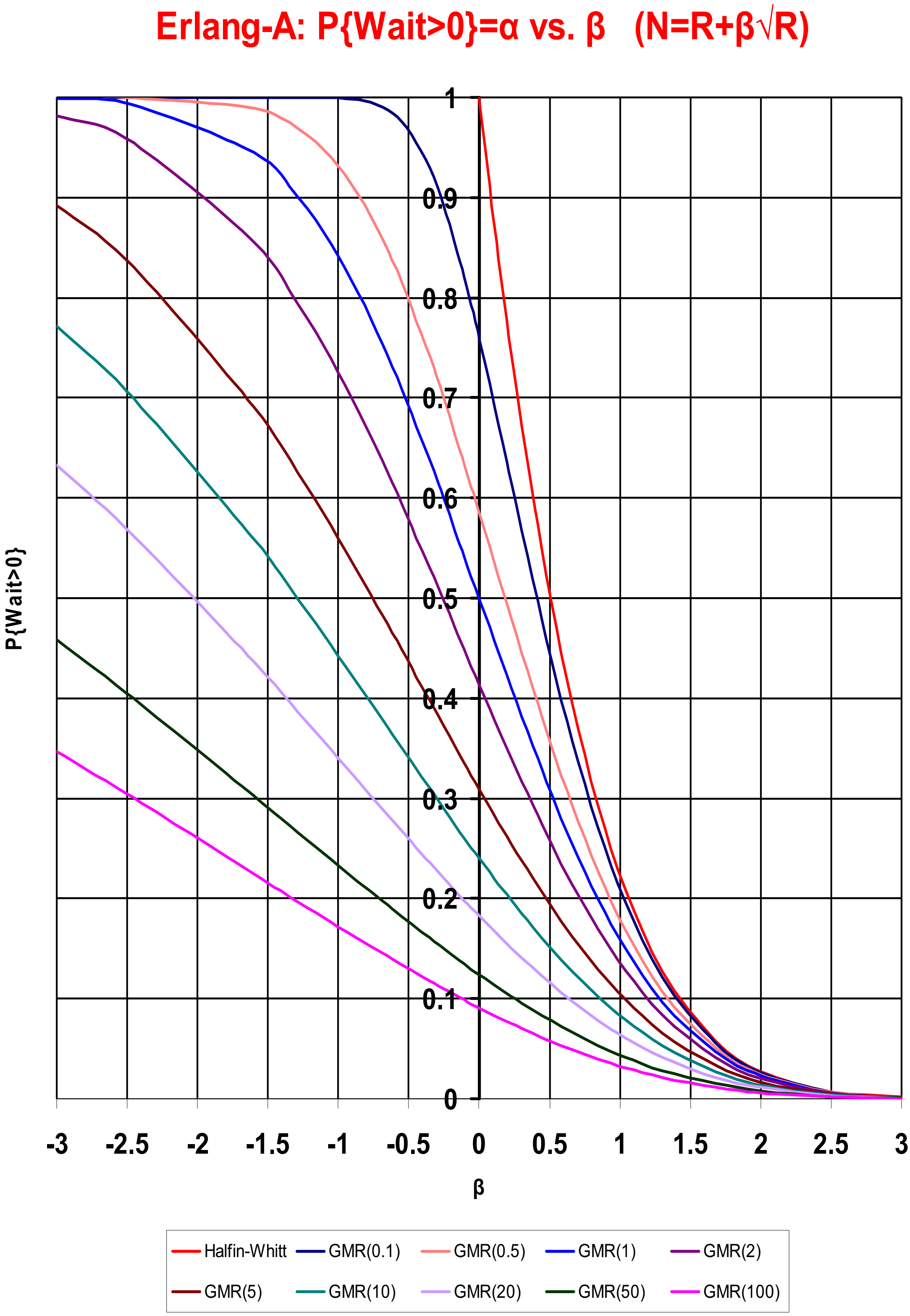
$$\mathrm{L(M/M/\infty)} \overset{\mathrm{d}}{\approx} \mathrm{Normal(R,R)} \overset{\mathrm{d}}{=} \mathrm{R}+\mathrm{Z}\sqrt{\mathrm{R}}$$

$$\Rightarrow \quad P\{W_q>0\} \approx P\left\{Z\geq \frac{N-R}{\sqrt{R}}\right\} = 1-\phi\left(\frac{N-R}{\sqrt{R}}\right)$$

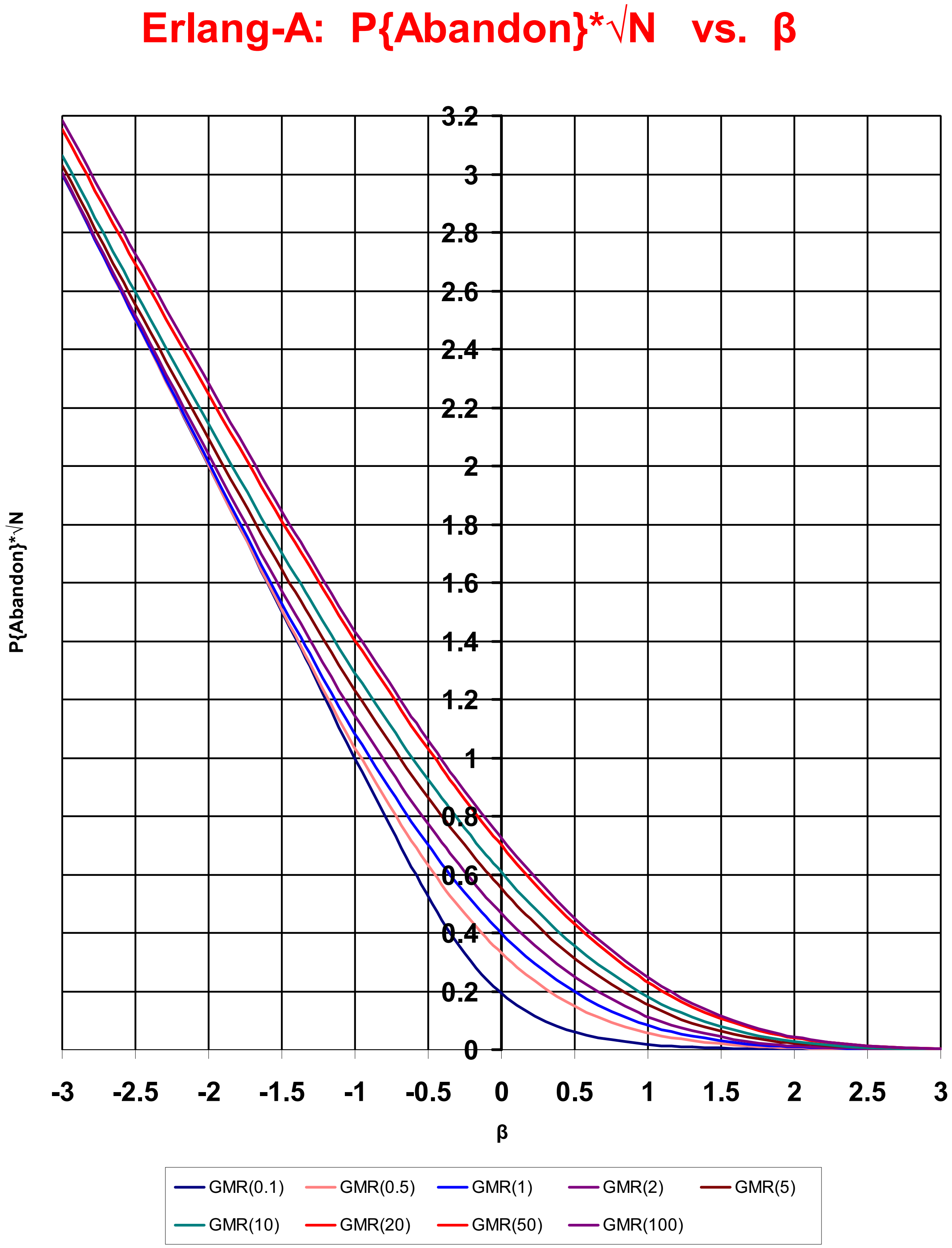
Given target delay-probability $\alpha = 1-\phi\left(\frac{\mathrm{N-R}}{\sqrt{\mathrm{R}}}\right)$

$$\Rightarrow \quad N = R+\beta\cdot\sqrt{R}, \quad \text{with} \quad \beta=\phi^{-1}(1-\alpha)$$

N is the "least integer for which" $P\{W_q>0\}\leq\alpha$



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.


Erlang-A: P{Abandon}*√N vs. β
P{Abandon}*√N
3.2
3
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
β
GMR(0.1)
GMR(0.5)
GMR(1)
GMR(2)
GMR(5)
GMR(10)
GMR(20)
GMR(50)
GMR(100)

Designing a Call Center

Approximate Performance Measures

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

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

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

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

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

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

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

Here

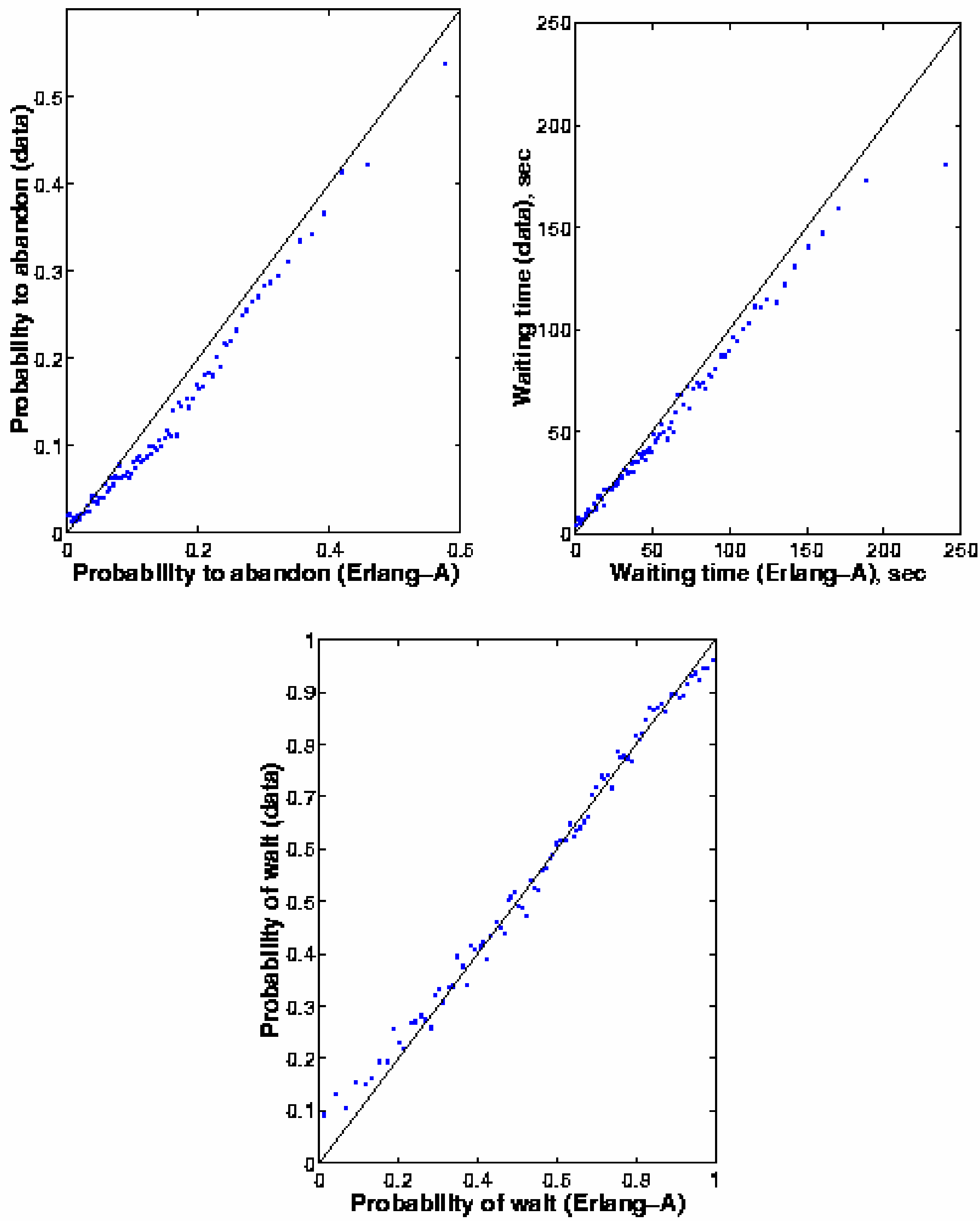
$$\hat{\beta} = \beta\sqrt{\frac{\mu}{\theta}},$$

$$\bar{\Phi}(x) = 1 - \Phi(x),$$

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

Fitting a Simple Model to a Complex Reality

Erlang-A Formulae vs. Data Averages



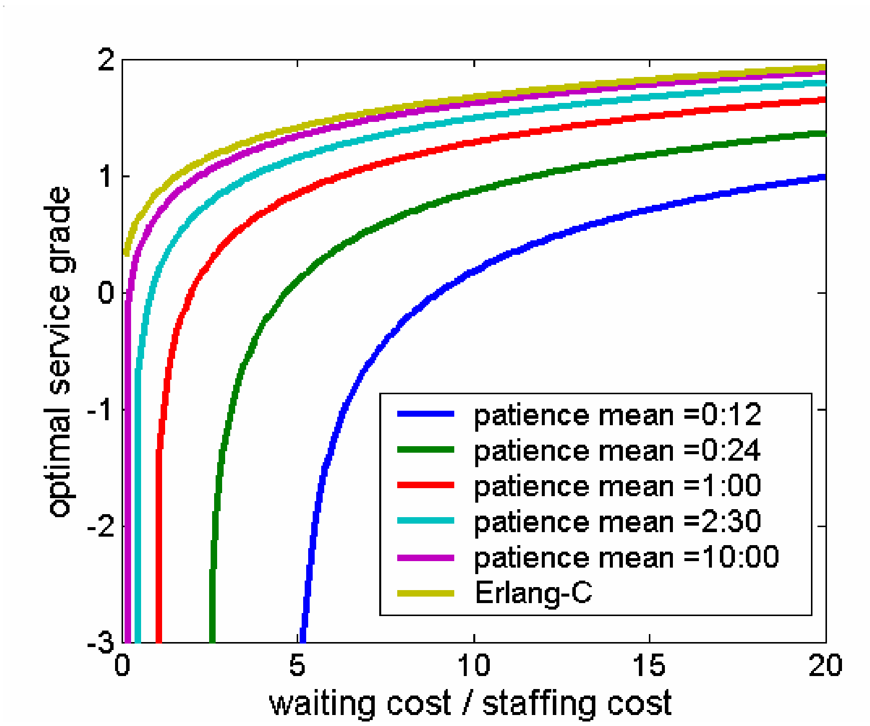
QED Dimensioning: $\sqrt{\cdot}$ Safety-Staffing

Dimensioning, with Borst, Reiman and Zeltyn:
Asymptotically optimal $\sqrt{\ }$ - safety staffing **(conjectured).**

$$N \approx R + y^*\left(r\,;\frac{\theta}{\mu}\right)\cdot\sqrt{R}$$

r = cost of delay / cost of staffing;

$y^*\left(r\,;\frac{\theta}{\mu}\right)$ = optimal service grade: independent of λ!

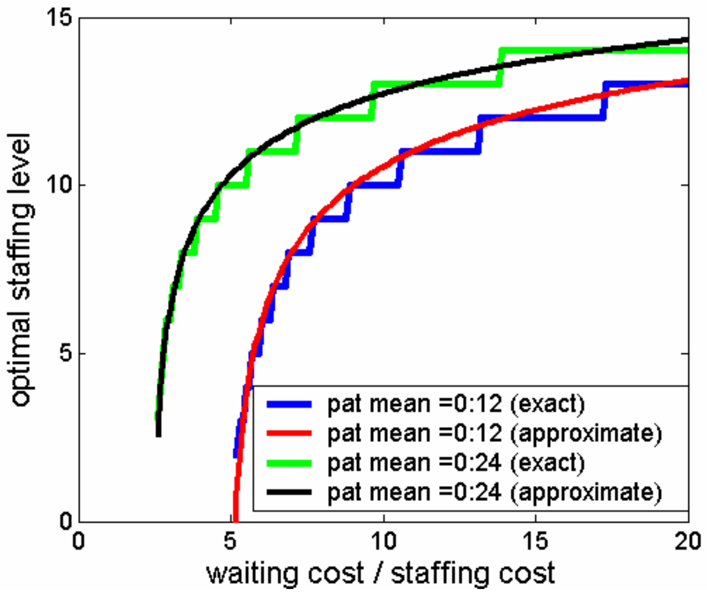


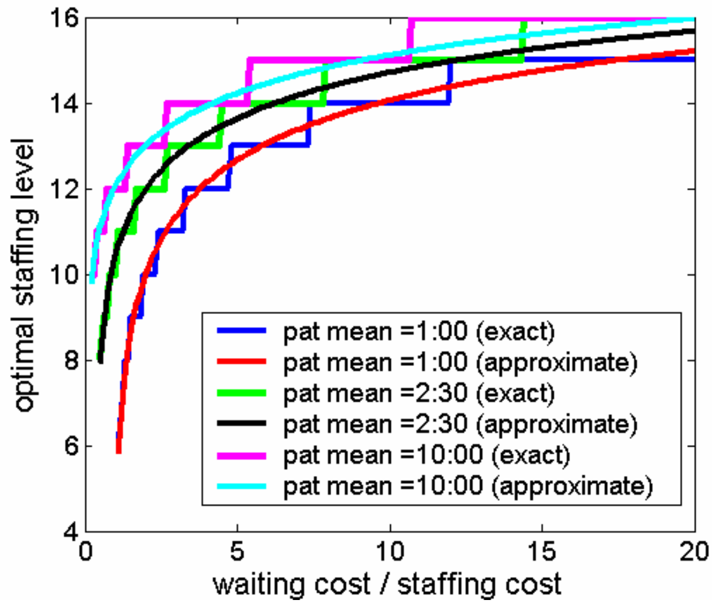
As $\theta \downarrow 0$, $y^*\left(r\,;\frac{\theta}{\mu}\right)$ increases to $y^*(r)$ (M/M/N).

Note: $r < \frac{\theta}{\mu}$ implies that "no service" is optimal.

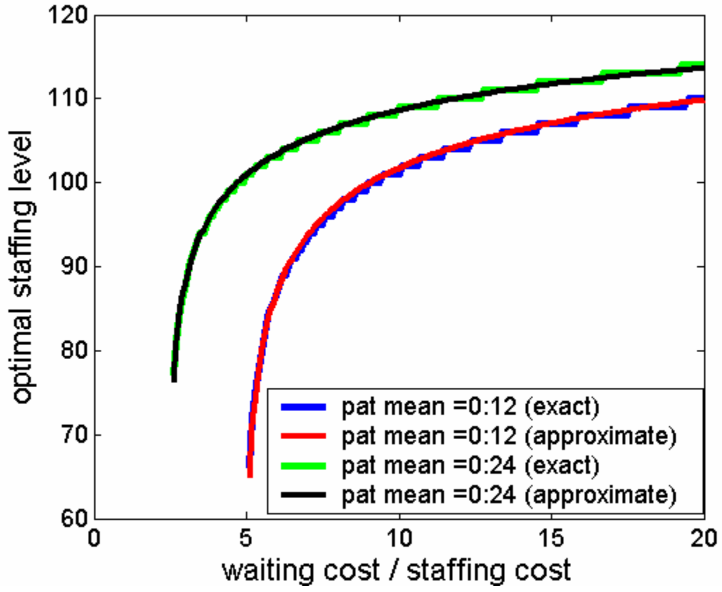
Asymptotically Optimal Staffing

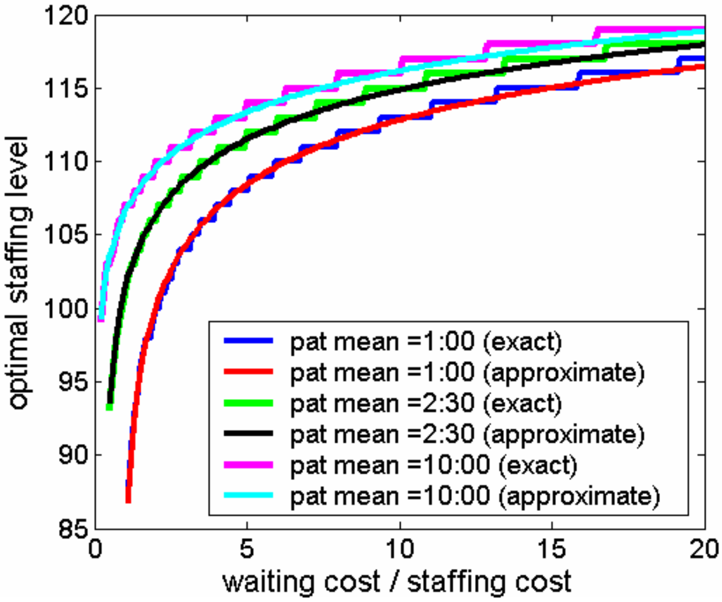
$\lambda = 10,\ \mu = 1$





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





Economics: $\sqrt{\cdot}$ Safety-Staffing

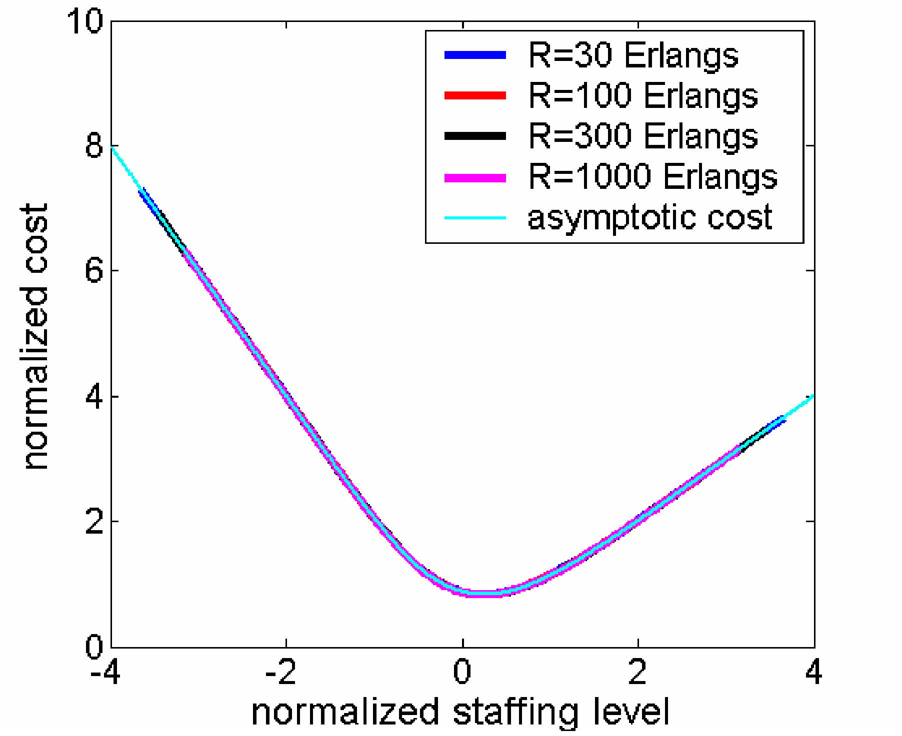
Cost $= c \cdot N + d \cdot \lambda E W_q$ (costs: c-staffing, d-delay).

Optimal staffing level:

$$N^* \approx R + y^*(r;s) \cdot \sqrt{R}, \qquad r = \frac{d}{c}, \quad s = \sqrt{\frac{\mu}{\theta}}.$$

$$y^* = \arg\min_{-\infty<y<\infty} \left\{ c \cdot y + d \cdot \frac{P(y;s)}{y} \cdot ys \cdot [h(ys) - ys] \right\}.$$

Numerical tests exhibit **remarkable** accuracy:
Actual cost function "coincides" with asymptotic cost.



Normalized staffing level = $(N - R)/\sqrt{R}$,

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

Asymptotic cost: $c \cdot y + d \cdot \dfrac{P(y;s)}{y} \cdot ys \cdot [h(ys) - ys]$.