Sergey Zeltyn March 2005

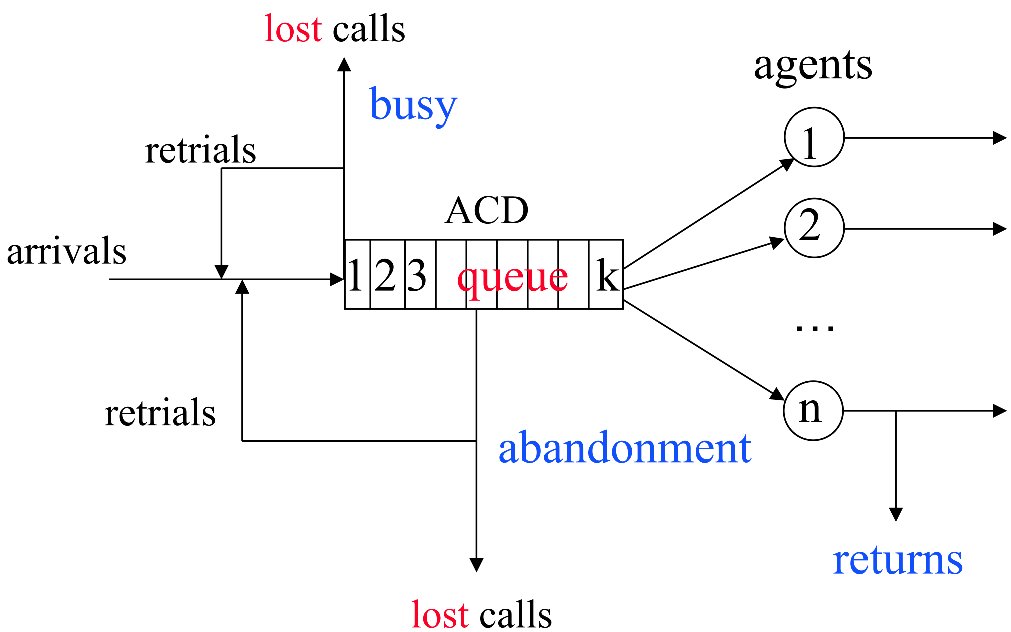
STAT 991. Service Engineering.
The Wharton School. University of Pennsylvania.

Operational Regimes in Call Centers: Empirically-Based Queueing Theory

Based on:

- Mandelbaum A. *Service Engineering* course, Technion.
  `http://iew3.technion.ac.il/serveng2005`
- Mandelbaum A. and Zeltyn S. Call Centers with Impatient Customers: Many-Server Asymptotics of the M/M/n+G queue. Submitted to *QUESTA*.

Schematic representation of a telephone call center



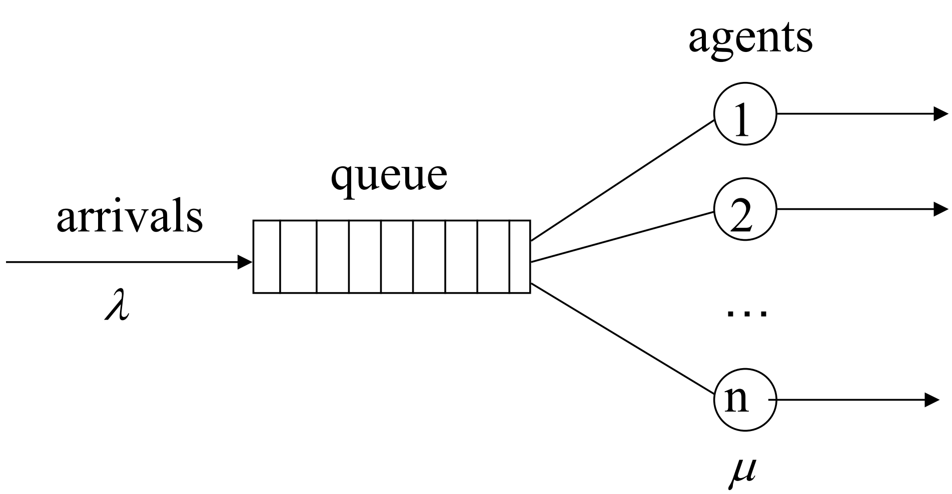
How to model?

Review of queueing models

Basic models:

- Poisson arrivals, rate λ;
- n exponential servers, rate μ.

M/M/n (Erlang-C) queue

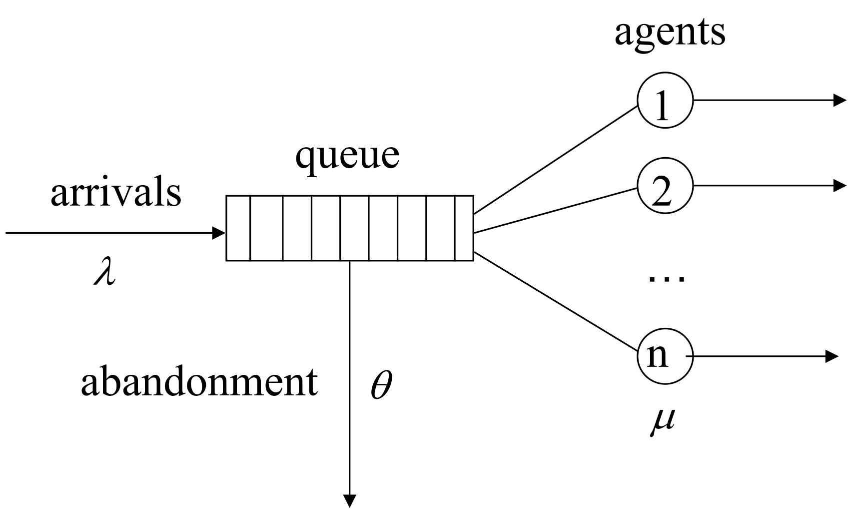


M/M/n/k queue

k trunks, $k - n$ slots in queue.

Important special case: **M/M/n/n (Erlang-B)**.

M/M/n+M (Erlang-A) queue



- **Patience time** $\tau \sim \exp(\theta)$:
  time a customer is willing to wait for service;
- **Offered wait** V:
  waiting time of a customer with infinite patience;
- If $\tau \le V$, customer abandons; otherwise, gets service;
- **Actual wait** $W = \min(\tau, V)$.
- Always stable;
- $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W]$.

Performance measures

Include

- L_q – number of customers in the queue,
- W – waiting time of a customer in the queue,
- $\mathrm{P}\{\mathrm{Ab}\}$ – probability to abandon,
- $P\{W \le T; \mathrm{Sr}\}$ – fraction of well-served,
- Agents' utilization.

Examples of **performance goals**:

- $\mathrm{P}\{\mathrm{Ab}\} \le 3\%$;
- $\mathrm{P}\{W \le 20 \text{ sec}; \mathrm{Sr}\} \ge 80\%$;
- $\mathrm{E}[W] \le 10$ sec;
- $\mathrm{P}\{W > 0\} \le 50\%$.

Staffing problem: find minimal n s.t. performance goal(s) are satisfied.
(Then shifts and specific agents should be assigned.)
A specific problem can be solved via 4CallCenters.

Lacks insight, **Rules of thumb.**
"How many agents needed if arrival rate doubles?"
"How sensitive is performance to 50% error in patience estimate?"

Motivation:
The Right Answer for the Wrong Reason

Recall: $R = \lambda/\mu$ is the **offered load** (measured in Erlangs): minutes of work that arrive per minute.

Deterministic ("naive") approach:
Staffing according to working load: $\boldsymbol{n = R}$.

Erlang-C: tele-queue "explodes".

What if **abandonment** is taken into account?

Erlang-A: E[S]=3 min, E[τ]=3 min

λ/hr	n	Occupancy	P$\{W > 0\}$	E[W]	P{Ab}
20	1	63.2%	63.2%	1:06.2	36.8%
100	5	82.5%	56.0%	0:31.6	17.5%
500	25	92.0%	52.7%	0:14.3	8.0%
2,500	125	96.4%	51.2%	0:06.4	3.6%
9,000	450	98.1%	50.6%	0:03.4	1.9%
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	∞	1 ?	50% ?	0 ?	0 ?

Motivation: The Right Answer for the Wrong Reason

Erlang-A: E[S]=3 min, E[τ]=6 min

λ/hr	n	Occupancy	P{$W > 0$}	E[W]	P{Ab}
2,500	125	97.0%	59.6%	0:10.6	3.0%
9,000	450	98.4%	59.1%	0:05.6	1.6%

Moderate-to-large $n \Rightarrow$ reasonable-to-good performance.

Motivation:
What can be reached? At what cost?
Quality-Driven Operational Regime

U.S. retail company. ACD Report.

ie	Avg Speed Ans [W]	Avg Aban Time	ACD Calls [λ]	Avg ACD Time [1/μ]	Avg ACW Time	Aban Calls [#Aban]	% ACD Time	% Ans Calls	Avg Pos Staff [N]	Calls Per Pos	%Serv Lev	%Aux Time	%ACW Time	%ACD Time [P]
Totals	:00:02	:00:28	10456	:03:47	:00:25	46	53	95	70	149		8		
12:00 AM*	:00:00	:00:00	26	:04:31	:00:02	1	76	61	7	4	61	2	16	61
12:30 AM*	:00:03	:04:10	14	:07:27	:00:33	1	89	52	5	3	48	1	26	63
1:00 AM*	:00:00		9	:04:54	:11:29	0	91	90	1	7	90	0	26	65
5:30 AM*			0			0	0		0	0		33	0	0
6:00 AM*	:00:00		12	:03:21	:00:19	0	21	100	7	2	100	9	2	19
6:30 AM*	:00:00		27	:02:51	:00:20	0	32	100	14	2	100	5	3	29
7:00 AM*	:00:00		62	:03:34	:00:15	0	38	100	21	3	100	13	4	34
7:30 AM*	:00:00		93	:03:11	:00:34	0	36	100	30	3	100	7	4	32
8:00 AM*	:00:00		120	:03:37	:00:40	0	39	100	47	3	100	8	6	33
8:30 AM*	:00:00		193	:03:04	:00:14	0	44	100	61	3	100	10	7	37
9:00 AM*	:00:01		293	:03:25	:00:25	0	54	99	75	4	97	9	7	47
9:30 AM*	:00:02	:00:06	361	:03:45	:00:22	2	60	97	91	4	93	8	8	52
10:00 AM*	:00:02	:00:01	416	:03:49	:00:26	1	63	97	94	4	98	5	8	55
10:30 AM*	:00:00		349	:03:35	:00:33	0	52	99	96	4	99	6	8	44
11:00 AM*	:00:00		352	:03:50	:00:27	0	51	100	102	3	100	7	8	45
11:30 AM*	:00:00		349	:03:44	:00:18	0	49	100	97	4	100	8	5	45
12:00 PM*	:00:01		354	:03:59	:00:18	0	52	95	95	4	95	8	5	47
12:30 PM*	:00:00		336	:03:38	:00:21	0	52	99	97	3	99	9	8	46
1:00 PM*	:00:00		347	:03:53	:00:32	0	51	99	98	4	99	11	6	44
1:30 PM*	:00:00		368	:03:52	:00:14	0	56	98	99	4	99	11	7	60
2:00 PM*	:00:01		393	:03:55	:00:17	0	51	100	106	4	100	10	5	46
2:30 PM*	:00:00		403	:03:58	:00:13	0	54	100	112	4	100	10	4	50
3:00 PM*	:00:00	:00:04	410	:04:02	:00:16	1	57	98	110	4	98	8	5	51
3:30 PM*	:00:00		347	:03:59	:00:14	0	60	100	100	3	100	7	5	45
4:00 PM*	:00:00		382	:03:48	:01:37	0	54	100	96	4	100	6	7	47
4:30 PM*	:00:00		379	:03:41	:00:19	0	55	99	97	4	99	8	5	50
5:00 PM*	:00:00		411	:03:53	:00:19	0	53	100	109	4	100	9	5	46
5:30 PM*	:00:01		387	:03:58	:00:19	0	58	99	96	4	99	10	6	51
6:00 PM*	:00:01	:00:21	371	:03:28	:00:25	1	53	98	91	4	98	9	6	47
6:30 PM*	:00:00		260	:03:26	:00:13	0	41	100	90	3	100	8	4	37
7:00 PM*	:00:00		289	:03:24	:00:17	0	42	100	78	3	100	9	5	38

Quality-Driven Operational Regime. Performance Analysis

10:00-10:30 am, with 94 agents;
416 calls;
2 seconds ASA.

$$
\begin{array}{llcl}
\text{Service time} & \text{E[S]} & = & \text{ACD Time + ACW Time} \\
 & & = & 3{:}49 \; + \; 0{:}26 \; = \; 4{:}15
\end{array}
$$

$$
\begin{array}{llcl}
\text{Offered load} & R & = & \lambda \times \text{E}[S] \\
 & & = & 416 \times (\text{ 4:15 / 30 min }) \\
 & & = & \text{1768 min / 30 min} \; = \; \text{59 Erlangs}
\end{array}
$$

$$
\begin{array}{llcl}
\text{Occupancy} & \rho & = & R/n \\
 & & = & 59/94 \; = \; 63\%
\end{array}
$$

Compare with "% ACD Time" column of ACD report.

Rule of Thumb: $n \; \approx \; R \cdot (1 + \gamma)$,
$\gamma > 0$ – service grade.

Motivation: Operational Regimes

Health Insurance. Charlotte – Center. 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;
- $\mathbf{P\{W > 0\} \approx \alpha}$, $\quad 0 < \alpha < 1$.

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

(Very close to $n = 206.1$, recall deterministic staffing.)

Definition:

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

Service grade

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

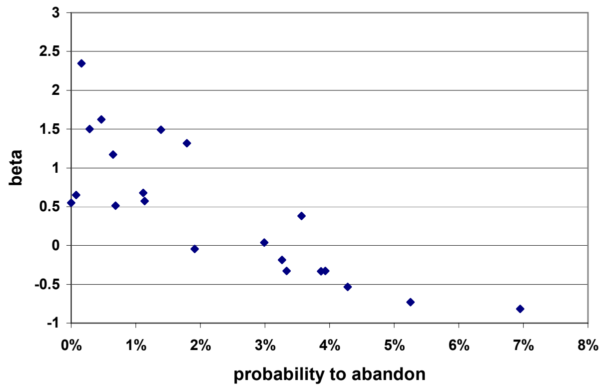
Square-Root Staffing Rule: Described by Erlang in 1924!
"In use at the Copenhagen Telephone Company since 1913".

QED Regime: Examples

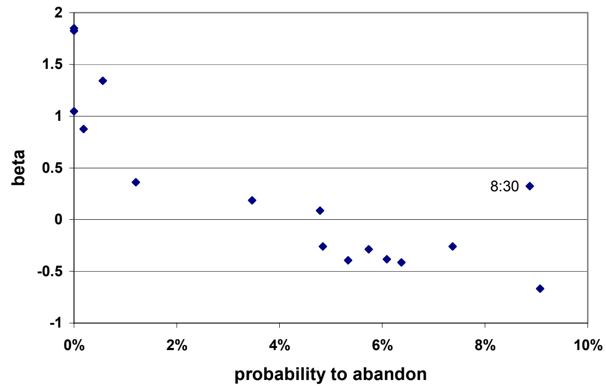
Two call center: U.S. (health insurance) and Italian (banking).

Service grade - correlation with abandonment

U.S. data

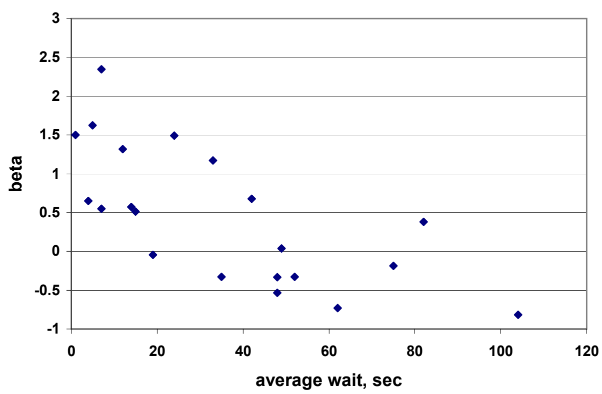


Italian data

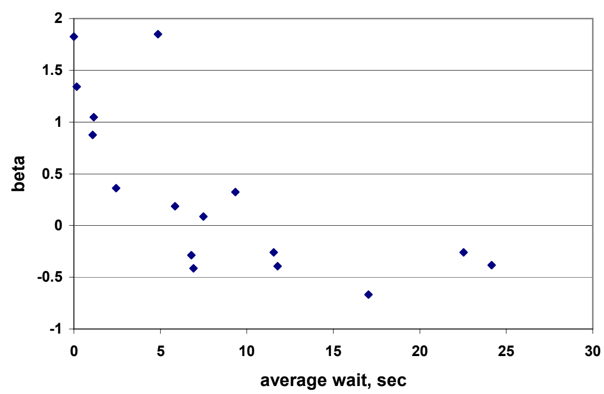


Service grade - correlation with average wait

U.S. data



Italian data



Erlang-B Queue: QED Regime

Recall: $E_{1,n} \;=\; \mathrm{P}\{\text{Blocked}\} \;=\; \dfrac{R^n}{n!} \Big/ \sum_{j=0}^{n} \dfrac{R^j}{j!}\,.$

Theorem (Jagerman, 1974)

As $n \to \infty$, the following 3 statements are equivalent:

1. $n \approx R + \beta\sqrt{R}, \qquad -\infty < \beta < \infty,$

2. $\sqrt{n}(1-\rho) \to \beta,$

3. $\sqrt{n}E_{1,n} \to \alpha, \qquad 0 < \alpha < 1,$

in which case

$$\alpha \;=\; h(-\beta) \;=\; \frac{\phi(-\beta)}{\bar{\Phi}(-\beta)} \;=\; \frac{\phi(\beta)}{\Phi(\beta)}\,,$$

where $\phi, \Phi, \bar{\Phi}$ and h are density, cdf, survival function and hazard rate of $N(0,1)$, respectively.

Proof. 1 $\sim$ 2 – straightforward.

1 $\Rightarrow$ 3. Assume $n \approx R + \beta\sqrt{R}$.

$$E_{1,n} = \frac{\mathrm{P}\{X_R = n\}}{\mathrm{P}\{X_R \le n\}}$$

where $X_R \sim \mathrm{Poiss}(R)$.

$$\begin{aligned} \mathrm{P}\{X_R \le n\} &= \mathrm{P}\left\{\frac{X_R - R}{\sqrt{R}} \le \frac{n-R}{\sqrt{R}}\right\} \\ &\overset{CLT,1}{\approx} \mathrm{P}\{N(0,1) \le \beta\} = \Phi(\beta)\,. \end{aligned}$$

$$\begin{aligned} \mathrm{P}\{X_R = n\} &= \mathrm{P}\{n-1 < X_R \le n\} \\ &= \mathrm{P}\left\{\frac{n-R-1}{\sqrt{R}} < \frac{X_R - R}{\sqrt{R}} \le \frac{n-R}{\sqrt{R}}\right\} \\ &\approx \mathrm{P}\{\beta - \frac{1}{\sqrt{R}} \le N(0,1) \le \beta\} \\ &\approx \frac{1}{\sqrt{R}} \cdot \phi(\beta) \approx \frac{1}{\sqrt{n}} \cdot \phi(\beta)\,. \end{aligned}$$

3 $\Rightarrow$ 1. $n = R + \beta\sqrt{R} + o(\sqrt{R})$ iff
$\forall \epsilon > 0 \quad R + (\beta - \epsilon)\sqrt{R} \le n \le R + (\beta + \epsilon)\sqrt{R}$ for large n.
Assume not true. E.g., for some subsequence

$$n > R + (\beta + \epsilon)\sqrt{R}\,.$$

$E_{1,n}$ decreasing in $n \quad \Rightarrow \quad \limsup \sqrt{n} E_{1,n} < h(-\beta - \epsilon)$.
$h(\cdot)$ increasing function $\quad \Rightarrow \quad h(-\beta - \epsilon) < h(-\beta)$
$\Rightarrow$ contradiction to **3**.

Erlang-C Queue

Recall:

$$\mathrm{P}\{W > 0\} \;\triangleq\; E_{2,n} = \sum_{i \geq n} \pi_i = \frac{R^n}{n!}\;\frac{1}{1-\rho}\;\cdot \pi_0 \,,$$

where

$$\pi_0 \;=\; \left[\sum_{j=0}^{n-1} \frac{R^j}{j!} \;+\; \frac{R^n}{n!(1-\rho)}\right]^{-1} \;.$$

Palm's relation between Erlang-C and Erlang-B:

$$E_{2,n} = \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}}$$

Waiting time distribution:

$$\frac{W}{1/\mu} \;=\; \begin{cases} 0 & \text{wp} \;\; 1 - E_{2,n} \\ \exp\left(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho}\right) & \text{wp} \;\; E_{2,n} \end{cases}$$

Erlang-C Queue: QED Regime

Theorem (Halfin & Whitt, 1981)

The following 3 statements are equivalent:

1. *Manager's* view: $n \approx R + \beta\sqrt{R}, \qquad 0 < \beta < \infty,$

($\beta\sqrt{R}$ – **safety staffing**)

2. *Server's* view: $\sqrt{n}(1-\rho) \to \beta,$

3. *Customer's* view: $E_{2,n} \to \alpha, \qquad 0 < \alpha < 1,$

in which case

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

the Halfin-Whitt function.

4. In addition $\mathrm{E}[W|W>0] \approx \frac{1}{\sqrt{n}} \cdot \frac{1}{\mu\beta}.$

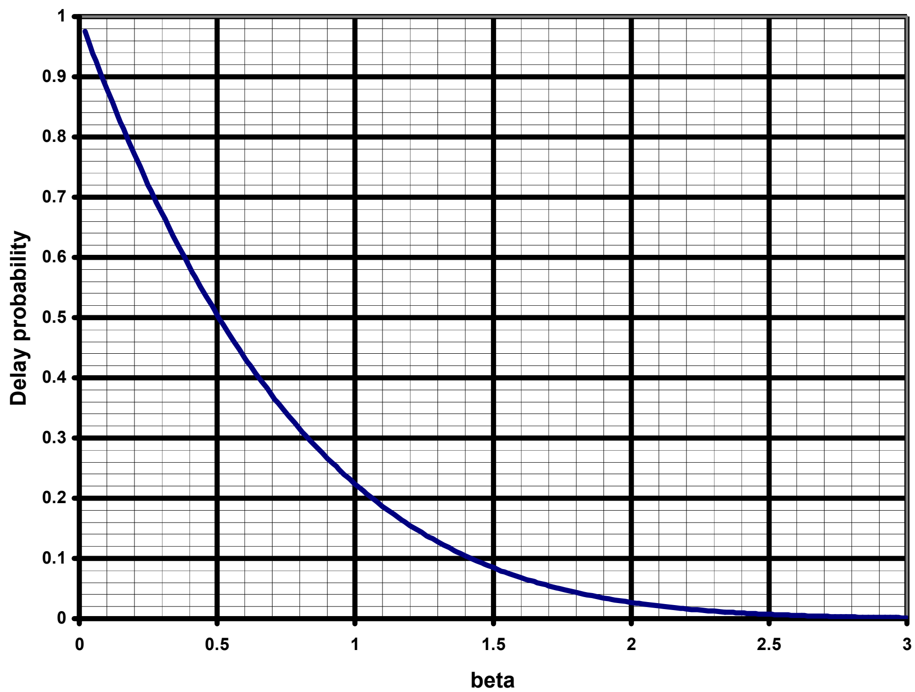
Proof. 1 ⇒ 3. Follows from the Palm's relation:

$$E_{2,n} \;=\; \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}}$$

$$\approx\; \frac{h(-\beta)/\sqrt{n}}{\beta/\sqrt{n} + h(-\beta)/\sqrt{n}} \;=\; \left[1 + \frac{\beta}{h(-\beta)}\right]^{-1} .$$

4 follows from **2** and wait distribution in Erlang-C.

The Halfin-Whitt Delay Function



Assume offered load $R = 1000$.

- $\beta = 0.5 \;\rightarrow\; \beta\sqrt{R} = 16,\; \mathrm{P}\{W > 0\} \approx 50\%$;
- $\beta = 2 \;\rightarrow\; \beta\sqrt{R} = 63,\; \mathrm{P}\{W > 0\} \approx 2\%$.

Erlang-C Queue: ED Regime

What if service goal is $\mathrm{E}[W] \le C$?

Assume $n = R + \gamma, \quad \gamma > 0$. Then

1. $n \cdot (1 - \rho) = \gamma$,

2. $\mathrm{P}\{W > 0\} \approx 1$,

3. $W \overset{d}{\approx} \exp(\gamma\mu)$.

Example. (4CallCenters)

$\mathrm{E}[S] = 6$ min ($\mu = 10$), γ=1.

λ/hr	n	ρ	$\mathrm{P}\{W > 0\}$	$\mathrm{E}[W]$
10	2	50%	33.3%	2:00
50	6	83.3%	58.8%	3:32
250	26	96.2%	78.2%	4:42
1000	101	99%	88.3%	5:18
9000	901	99.9%	95.9%	5:45
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
∞	∞	1	1	6:00

$\mathrm{E}[W|W > 0]$ remains constant (6:00).

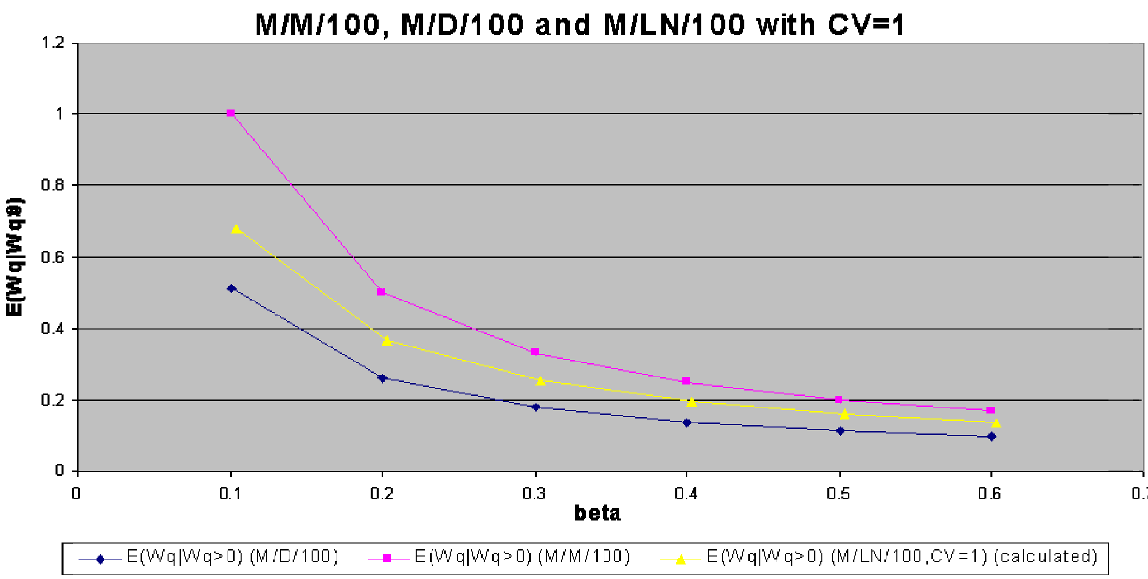
Decrease n by 1 $\rightarrow$ queue "explodes".

General Service Times in the QED Regime.

Mandelbaum & Schwartz, 2002.

Compare three M/G/n systems with E[S] = 1 (simulation):

- M/D/100, deterministic service times;
- M/M/100, exponential service times;
- M/LN/100, lognormal service times, $C_s = \sigma(S)/\mathrm{E}[S] = 1$.



Khintchine-Pollaczek approximation (n fixed, $\rho \uparrow 1$):

$$\mathrm{E}[W|W>0] \approx \frac{1}{n} \cdot \frac{\mathrm{E}[S]}{1-\rho} \cdot \frac{1+C_s^2}{2}.$$

Not accurate for lognormal distribution!

Queues with abandonment – impact of service distribution seems smaller (Whitt, 2004).

Theoretical Motivation: Square-Root Staffing in Erlang-A

Assume $\theta = \mu$.

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

Fact. If $\theta = \mu$, number-in-system distributions of M/M/n+M and M/M/∞ are identical. (The same B&D process.)

$$\begin{aligned} \mathrm{P}\{W(\mathrm{M/M/n{+}M}) > 0\} &\overset{\text{PASTA}}{=} \mathrm{P}\{L(\mathrm{M/M/n{+}M}) \geq n\} \\ &\overset{\theta=\mu}{=} \mathrm{P}\{L(\mathrm{M/M/}\infty) \geq n\} \end{aligned}$$

From lecture on classical queues:

$$L(\mathrm{M/M/}\infty) \sim \mathrm{Poisson}(R)\,.$$

For large R

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

Hence,

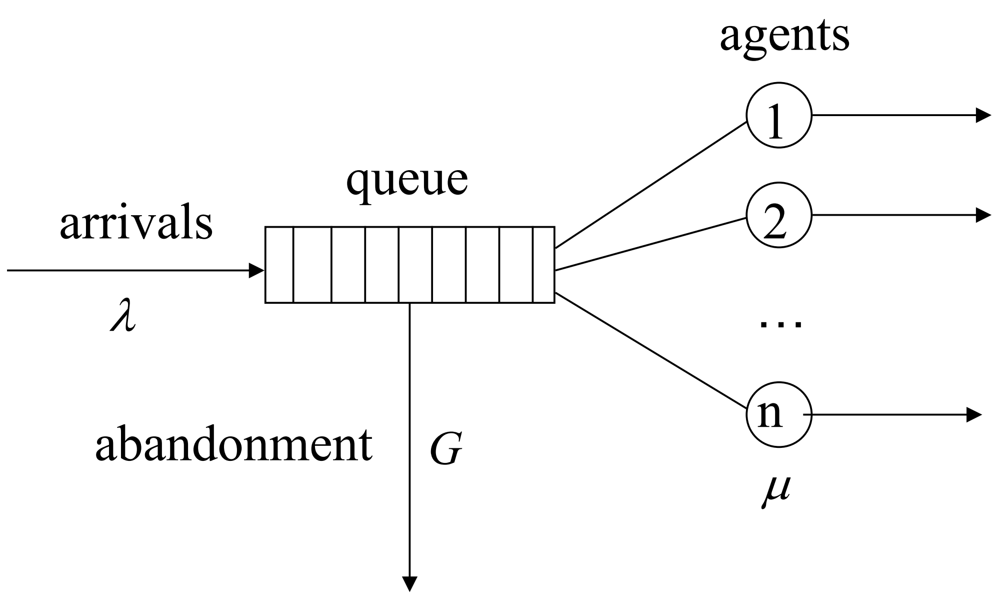
$$\mathrm{P}\{W > 0\} \approx \mathrm{P}\left\{Z \geq \frac{n-R}{\sqrt{R}}\right\} \approx \bar{\Phi}(\beta)\,.$$

Solution for $\theta \neq \mu$: Garnett, Mandelbaum & Reiman, 2002.

But we consider more general model.

M/M/n+G Queue

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



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.

M/M/n+G Queue: 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.

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

Finally,

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

Erlang-A: Substitute $\bar{G}(u) = e^{-\theta u}$,

$$H(x) = \frac{1}{\theta} \cdot (1 - e^{-\theta x})\,.$$

Performance measures calculated via building blocks:

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

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

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

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

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.

Erlang-A: Substitute $\theta = g_0$.

Main case: performance measures

- Probability of wait converges to constant:

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

Check: $g_0 = \mu \quad \Rightarrow \quad \mathrm{P}\{W > 0\} = \bar{\Phi}(\beta)$.

- 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 $\dfrac{1}{\sqrt{n}}$:

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

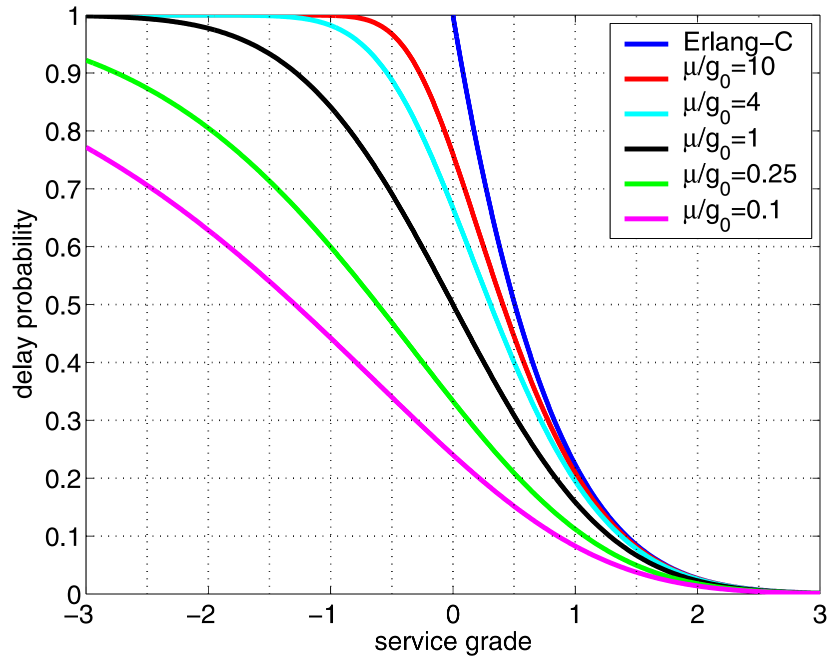
- Ratio between $\mathrm{P}\{\mathrm{Ab}\}$ and $\mathrm{E}[W]$ converges to patience density at the origin:

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

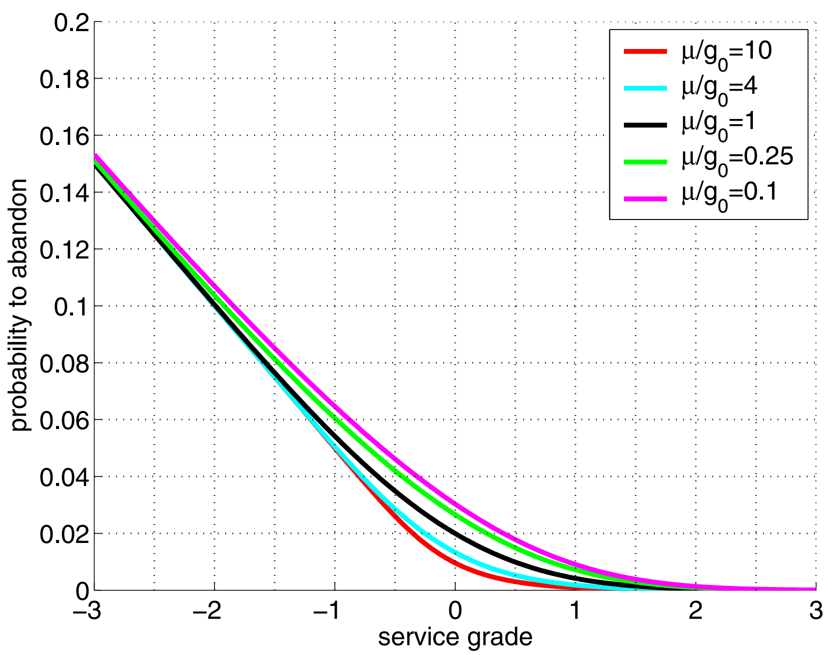
- Asymptotic distribution of wait:

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

QED Regime: Delay Probability



QED Regime: Probability to Abandon (n=400)



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

QED Operational Regime: Discussion

Points of view.

- **Customers:** $\mathrm{P}\{W > 0\} \approx \alpha, \quad \mathrm{P}\{\mathrm{Ab}\} \approx \frac{\gamma}{\sqrt{n}}$;
- **Agents:** Offered load per Server $= \frac{R}{n} \approx 1 - \frac{\beta}{\sqrt{n}}$;
- **Managers:** $n \approx R + \beta\sqrt{R}$.

$\beta = 0$: right answer for wrong reasons.

(Common in stochastic-ignorant operations.)

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

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

Equivalent to deterministic rule: assign number of agents equal to *offered load*.

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

According to patience distribution.

- **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, $\mathrm{E}[W] \to 0$ slowly. $\mathrm{P}\{\mathrm{Ab}\} \approx -\beta/\sqrt{n}$.
  - 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).
  - $\mathrm{P}\{W > 0\}$ decreases at rate $1/\sqrt{n}$;
  - $\mathrm{P}\{\mathrm{Ab}|V > 0\} \approx \mathrm{P}\{\mathrm{Blk}\}$;
  - $\mathrm{P}\{\mathrm{Ab}\} \approx h(-\beta)/\sqrt{n}$, asymptotic loss probability for Erlang-B.
- **Scaled balking.**
  Customer, not served immediately, balks with probability $p_b/\sqrt{n}$.
  Results are similar to the main case.

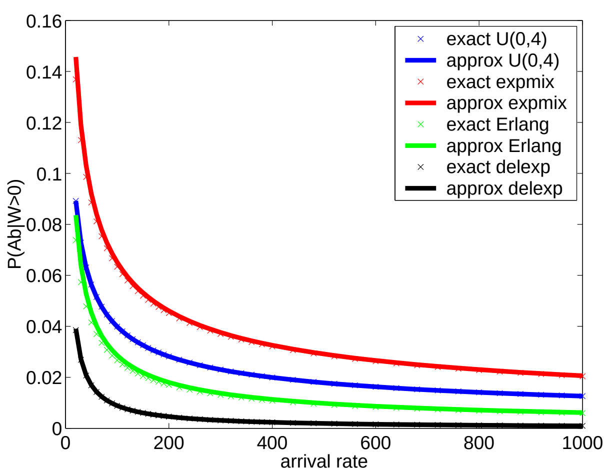
QED Regime: Numerical Experiments–1

Patience distributions:

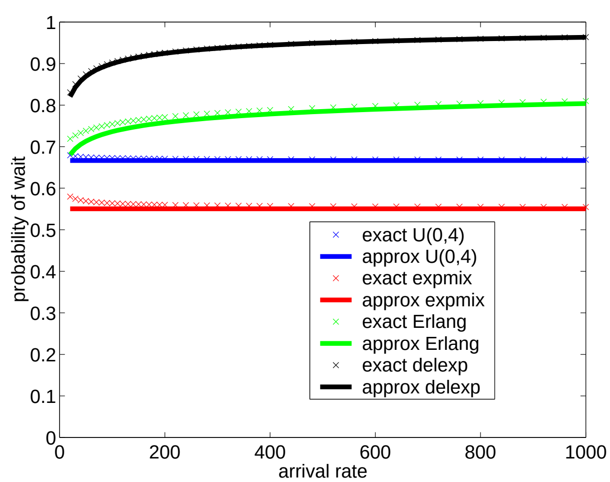
- *Uniform* on [0,4], $g_0 = 0.25$;
- *Hyperexponential*, 50-50% mixture of exp(mean=1) and exp(mean=1/3), $g_0 = 2/3$;
- *Erlang*, two exp(mean=1) phases, $g_0 = 0$;
- *Delayed exponential*, 1 + exp(mean=1), $g_0 = 0$.

Service grade $\beta = 0$.

Probability to abandon given delay vs. arrival rate



Probability of wait vs. arrival rate

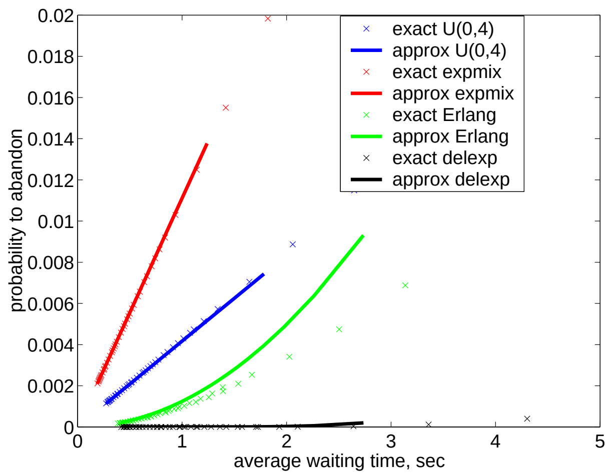


P{Ab} convergence rates: $1/\sqrt{n}$, $1/\sqrt{n}$, $n^{-2/3}$, exp, respectively.

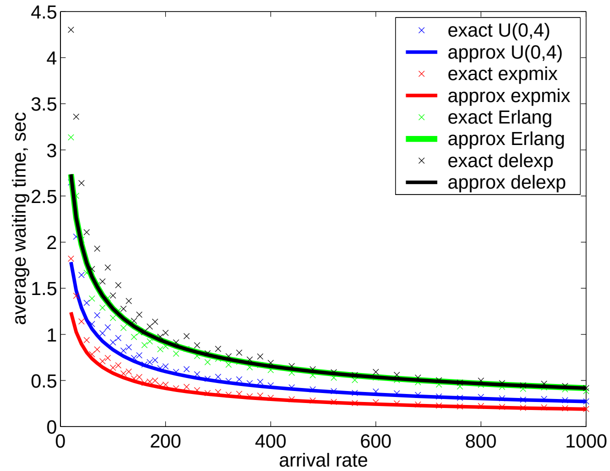
QED Regime: Numerical Experiments–2

Service grade $\beta = 1$.

Probability to abandon
vs. average waiting time



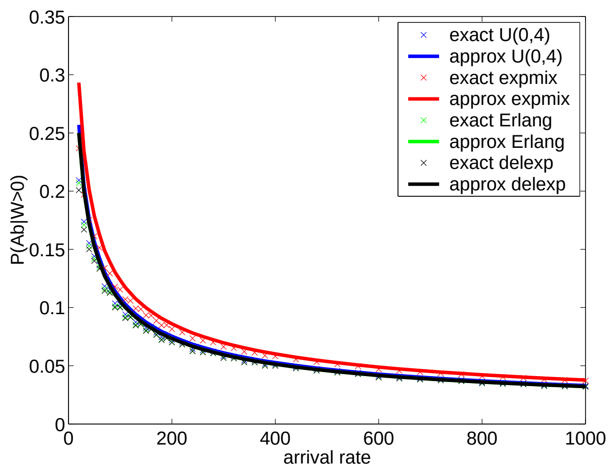
Average waiting time
vs. arrival rate



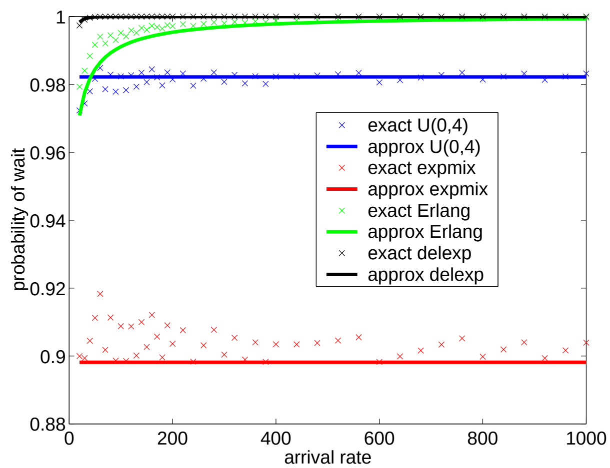
Note linear patterns in the first plot.

Service grade $\beta = -1$.

Probability to abandon given delay
vs. arrival rate



Probability of wait
vs. arrival rate



Convergence to $-\beta/\sqrt{n}$ for probability to abandon.

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

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

- Asymptotic distribution of wait:

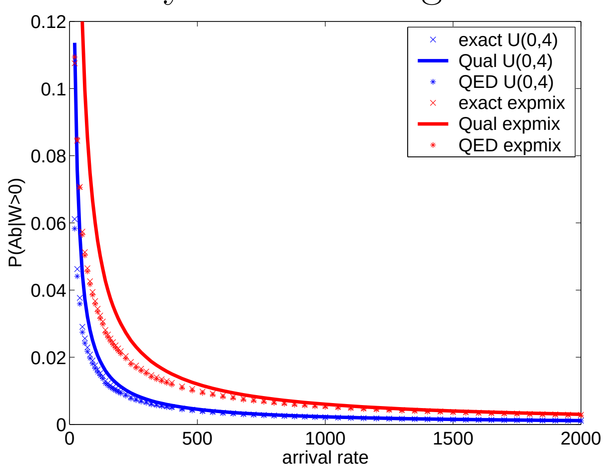
$$\mathrm{P}\left\{\frac{W}{\mathrm{E}(S)} > \frac{t}{n} \,\middle|\, W > 0\right\} \sim e^{-(1-\rho)t}, \qquad \rho = \frac{\lambda}{n\mu}.$$

QD Regime: Numerical Experiments

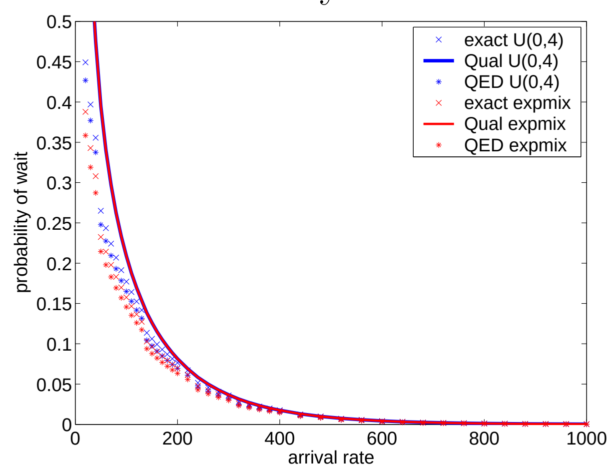
Patience distributions: Uniform, hyperexponential.

Service grade $\gamma = 1/9$, $\rho = 0.9$.

Probability to abandon given delay



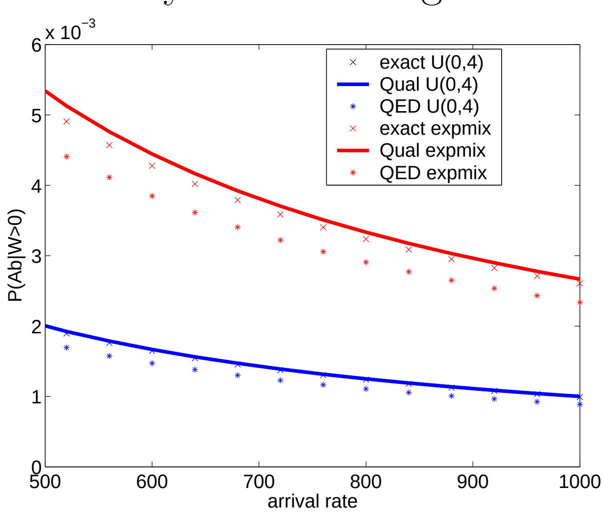
Probability of wait



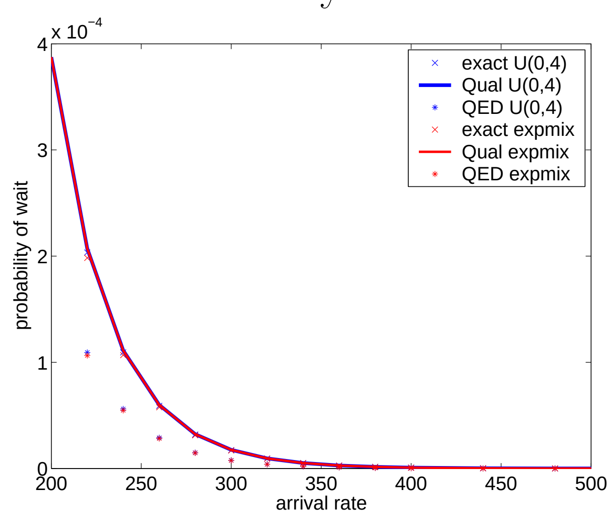
Overall, QED approximations are better than QD.

Service grade $\gamma = 0.25$, $\rho = 0.8$. Large arrival rate.

Probability to abandon given delay



Probability of wait



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 on 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:

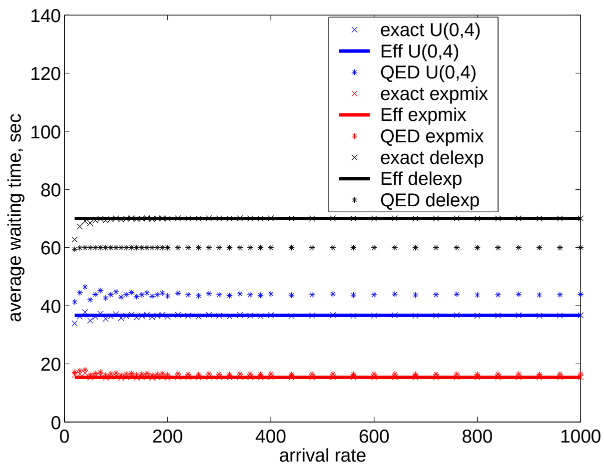
$$W \;\xrightarrow{w}\; G^* , \qquad \mathrm{E}[W] \;\rightarrow\; \mathrm{E}[\min(x^*, \tau)] .$$

ED Regime: Numerical Experiments

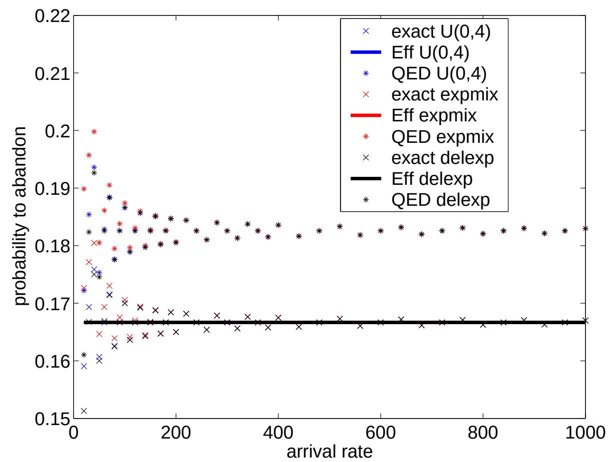
Patience distributions: Uniform, hyperexponential, delayed exponential.

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

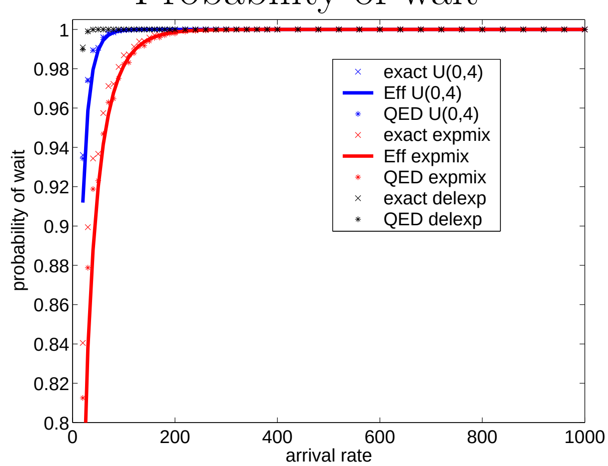
Average waiting time



Probability to abandon



Probability of wait

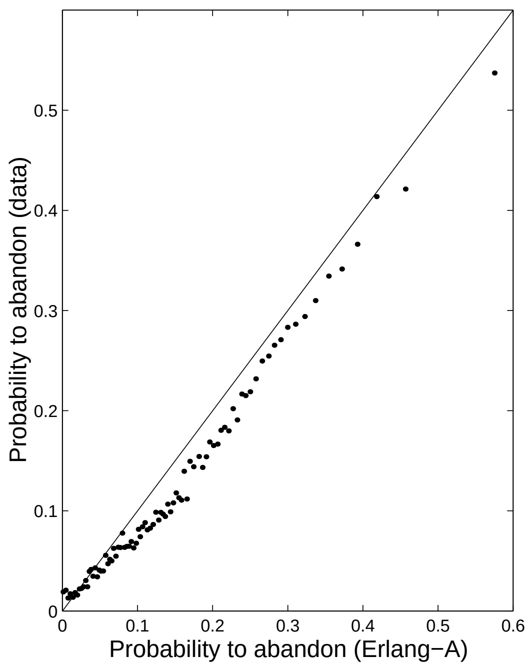


Fluid-limit ED approximations for P{Ab} and E[W] are better than QED.

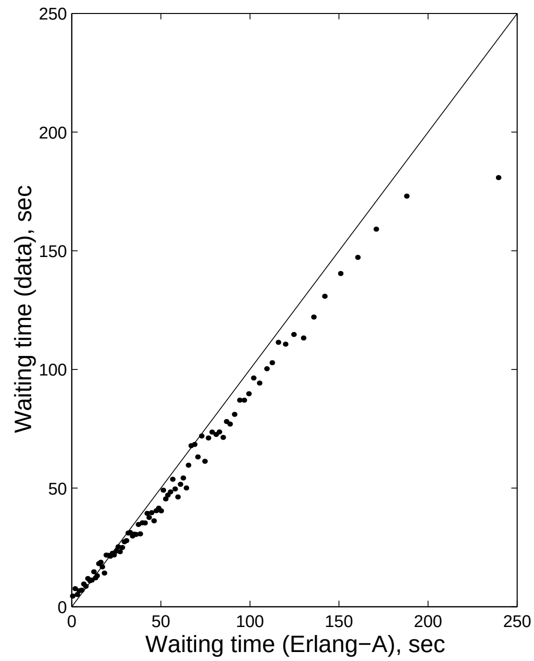
Fitting Erlang-A: Small Call Center

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

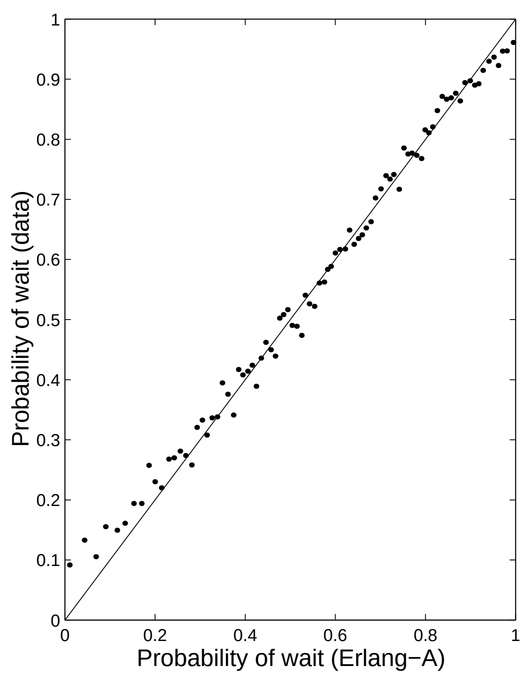
$\mathrm{P}\{\mathrm{Ab}\}$



$\mathrm{E}[W]$

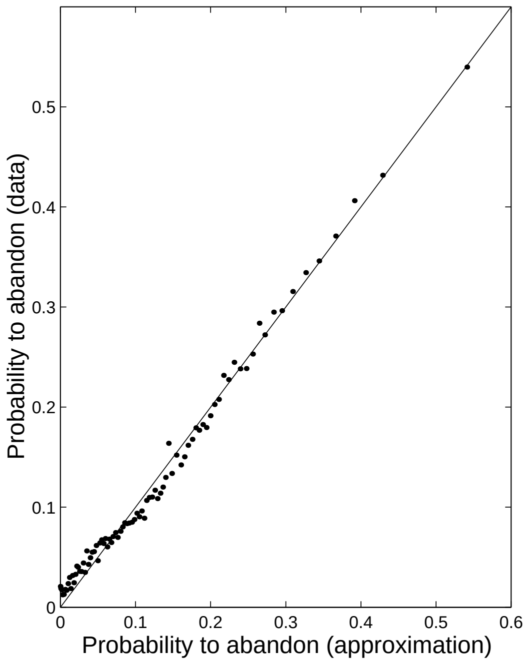


$\mathrm{P}\{W > 0\}$

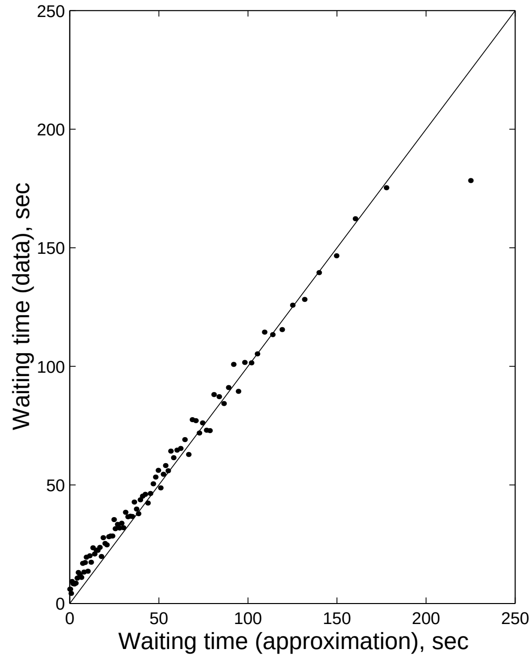


Erlang-A Approximations vs. Data Averages

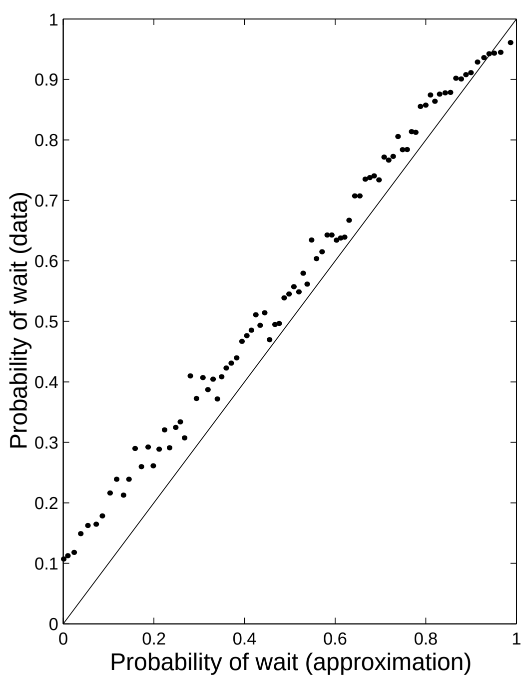
$\mathrm{P}\{\mathrm{Ab}\}$



$\mathrm{E}[W]$



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



Fitting Erlang-A: Small Call Center

Comments and conclusions

- Points: hourly data vs. Erlang-A output;
- Formulae with continuous n used;
- Patience estimated via $\mathrm{P}\{\mathrm{Ab}\}/\mathrm{E}[W]$ relation;
- Erlang-A estimates – close upper bounds;
- Erlang-A QED approximations – even better fit than exact formulae.

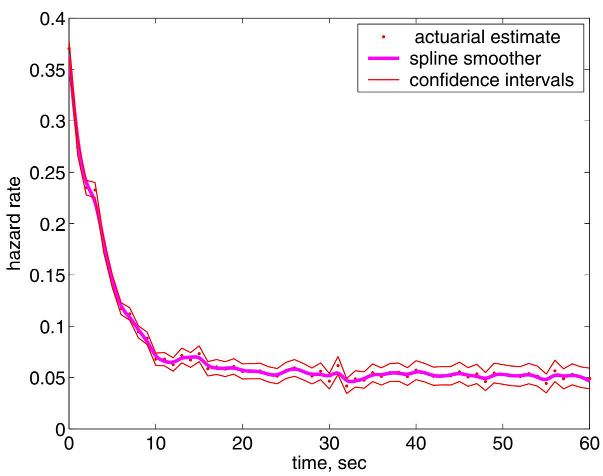
Fitting M/M/n+G: Large Call Center

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

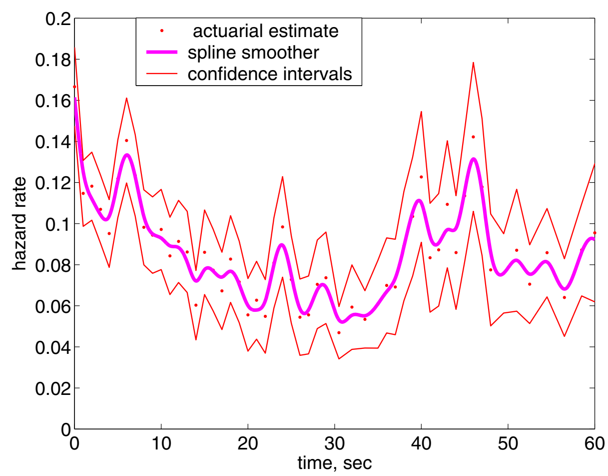
	Calls	$E[S]$	$P\{W > 0\}$	$P\{Ab\}$	$E[W]$
Retail	3,451,743	224.6 sec	30.6%	1.16%	6.33 sec
Telesales	349,371	453.9 sec	24.3%	1.76%	9.66 sec

Estimates of hazard rate

retail



telesales



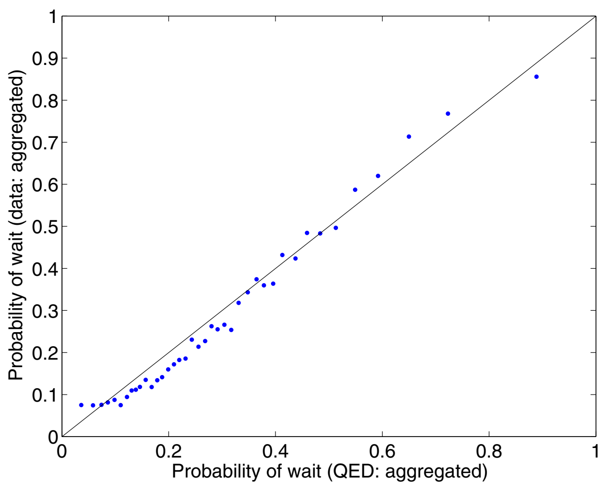
Problems/Challenges:

- Reliable data for number of agents n unavailable;
- Significant variability of hazard rate/density near the origin.

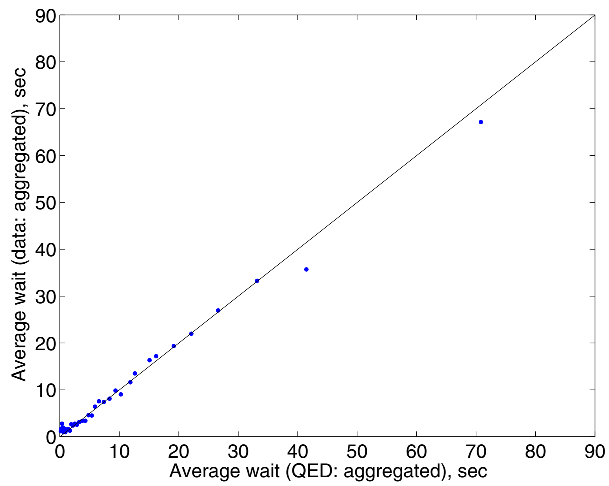
Approach: Estimate n via some performance measure ($P\{Ab\}$).
Fit other performance measure(s).

Substitute g_0 := estimate of $h(0)$ $\Rightarrow$ unsatisfactory fit.
Solution: Substitute g_0 := overall P{Ab}/E[W]
to QED formulae.

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

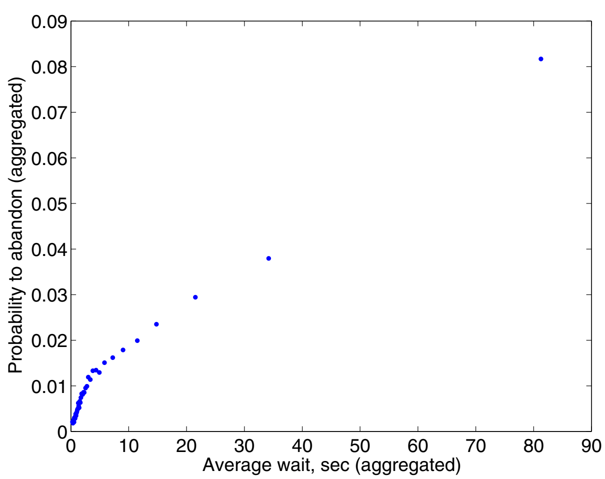


Telesales. $\mathrm{E}[W]$

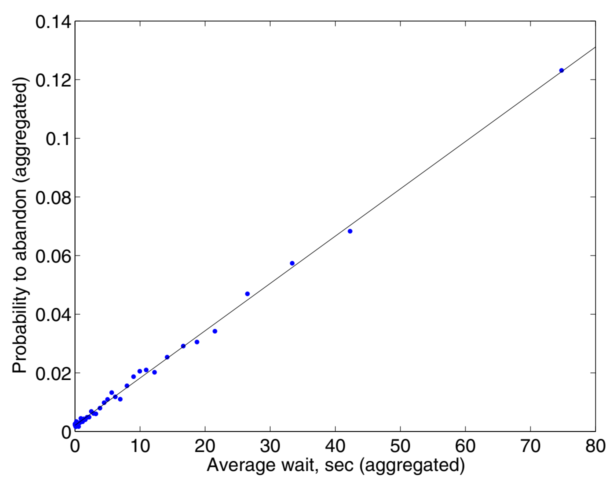


P{Ab}/E[W] relation

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:

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

$y^*(r)$ = optimal service grade, independent of λ:

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

where

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

Erlang-A. Optimal staffing level (conjecture):

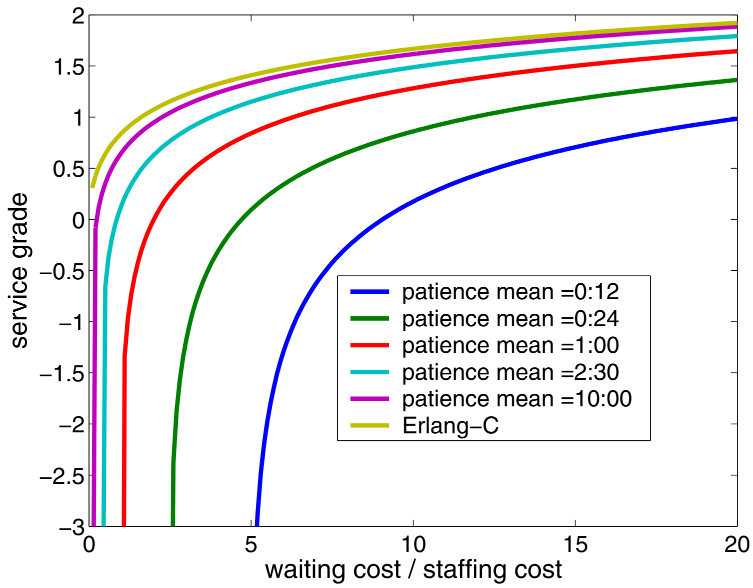
$$n^* \;\approx\; R + y^*(r;s)\sqrt{R}, \qquad s = \sqrt{\mu/\theta}\,,$$

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

where

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

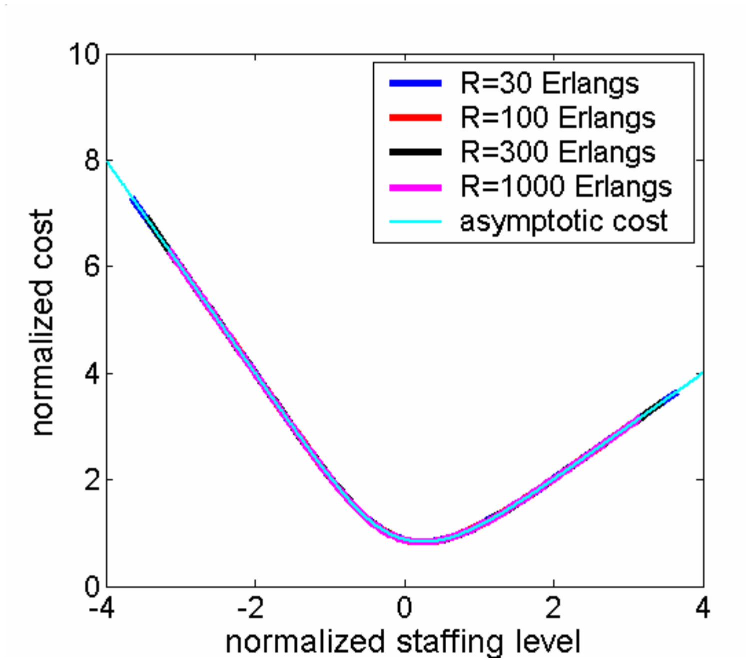
Optimal service grade. $\mathbf{E}[S] = 1$ min.



- $r < \theta/\mu$ 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 $= (n - R)/\sqrt{R}$;

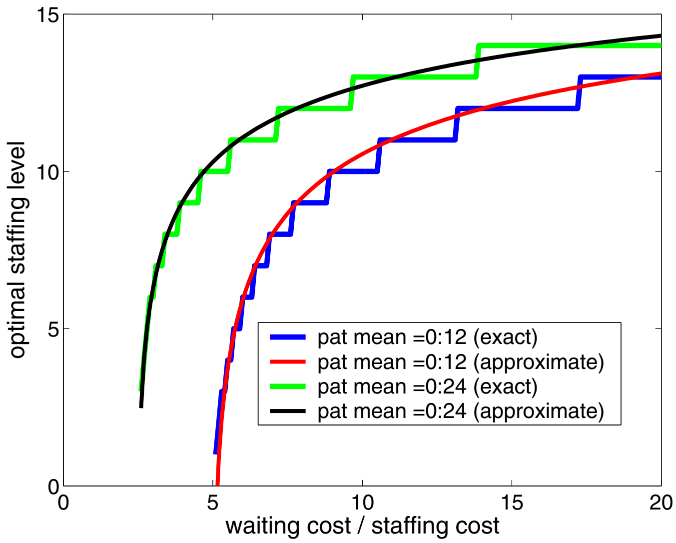
Normalized cost $= (\text{cost} - cR)/\sqrt{R}$;

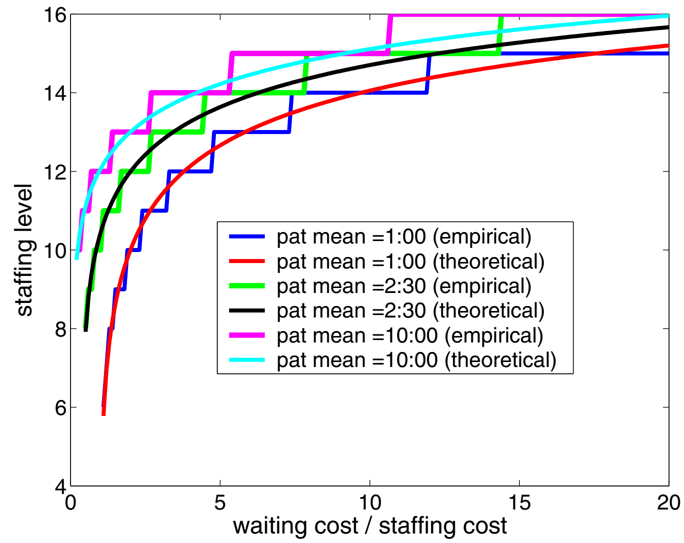
Asymptotic cost $= c \cdot y + d \cdot P_w(y; s) \cdot s \cdot [h(ys) - ys]$,

where $y =$ QED service grade.

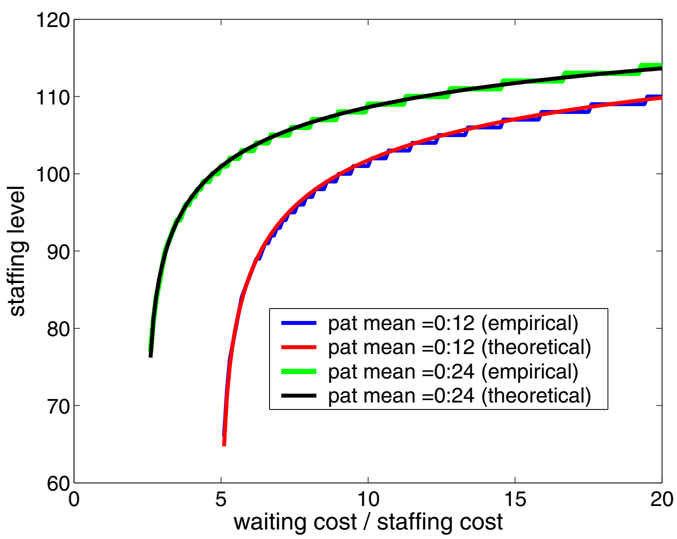
Erlang-A: Optimal Staffing

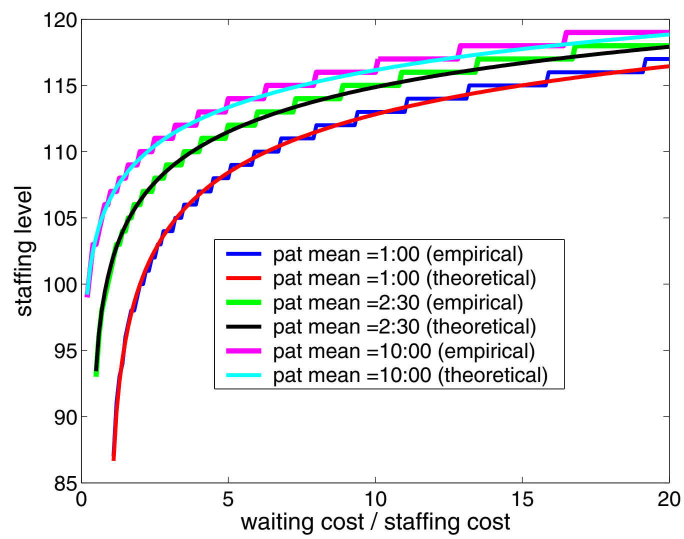
$\lambda = 10, \ \mu = 1$





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

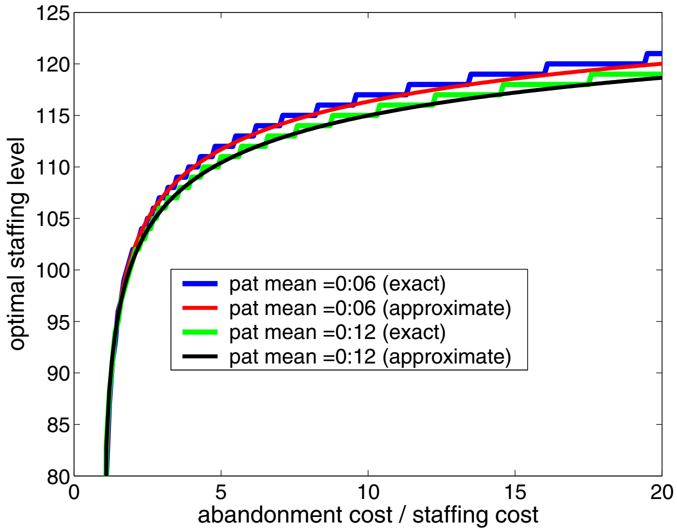


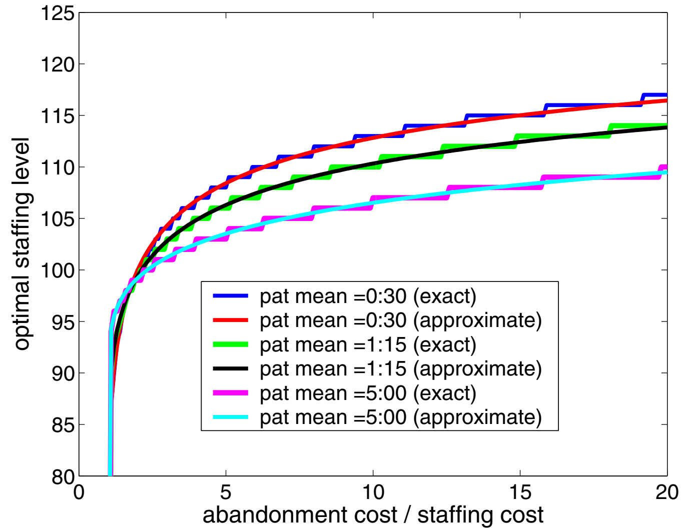


M/M/n+G: Optimal Staffing

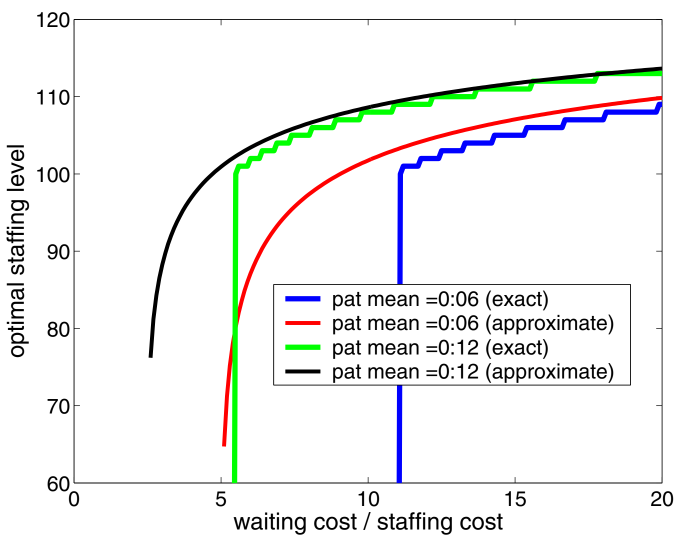
Uniformly Distributed Patience

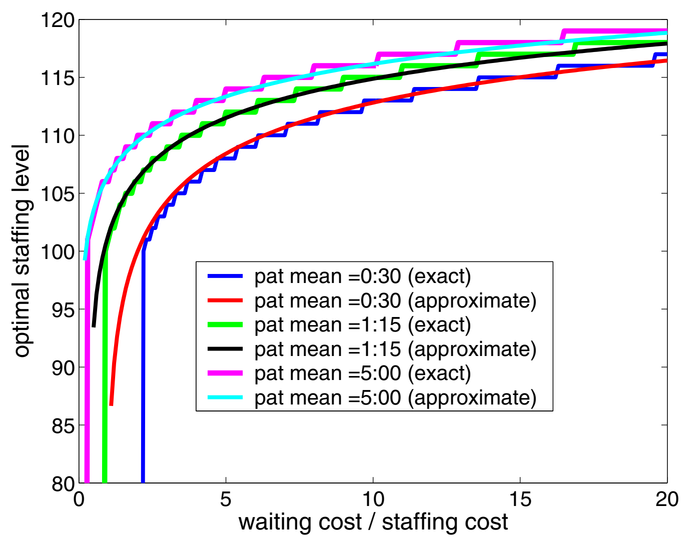
$$\textbf{Cost} = \boldsymbol{c} \cdot \boldsymbol{n} + \boldsymbol{d} \cdot \boldsymbol{\lambda}\mathbf{P}\{\mathbf{Ab}\}$$





$$\textbf{Cost} = \boldsymbol{c} \cdot \boldsymbol{n} + \boldsymbol{d} \cdot \boldsymbol{\lambda}\mathbf{E}[\boldsymbol{W}]$$





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

Deterministic Service Times

- Jelenkovic, Mandelbaum and Momcilovic (2004) Heavy traffic limits for queues with many deterministic servers, *QUESTA*.

M/M/n/k Queue

- Massey and Wallace (2004) An Optimal Design of the M/M/C/K Queue for Call Centers, to appear in *QUESTA*.

Time-Dependent Arrival Rate

- Jennings, Mandelbaum, Massey and Whitt (1996) Server staffing to meet time-varying demand, *Management Science*.
- Feldman, Mandelbaum, Massey and Whitt (2004) Staffing of time-varying queues to achieve time-stable performance, submitted to *Management Science*.

Skills-Based Routing

- Gurvich, Armony and Mandelbaum (2004) Staffing and control of large-scale service systems with multiple customer classes and fully flexible servers, working paper.
- Armony and Mandelbaum (2004) Design, staffing and control of large service systems: The case of a single customer class and multiple server types, working paper.

ED Operational Regime

- Bassamboo, Harrison and Zeevi (2004) Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method.
- Harrison and Zeevi (2004) A method for staffing large call centers using stochastic fluid models, to appear in *MSOM*.
- Whitt (2004) Fluid Models for Many-Server Queues with Abandonments, submitted to *Operations Research*.

Uncertainty about Arrival Rate

- Whitt (2004) Staffing a Call Center with Uncertain Arrival Rate and Absenteeism, submitted to *Management Science*.