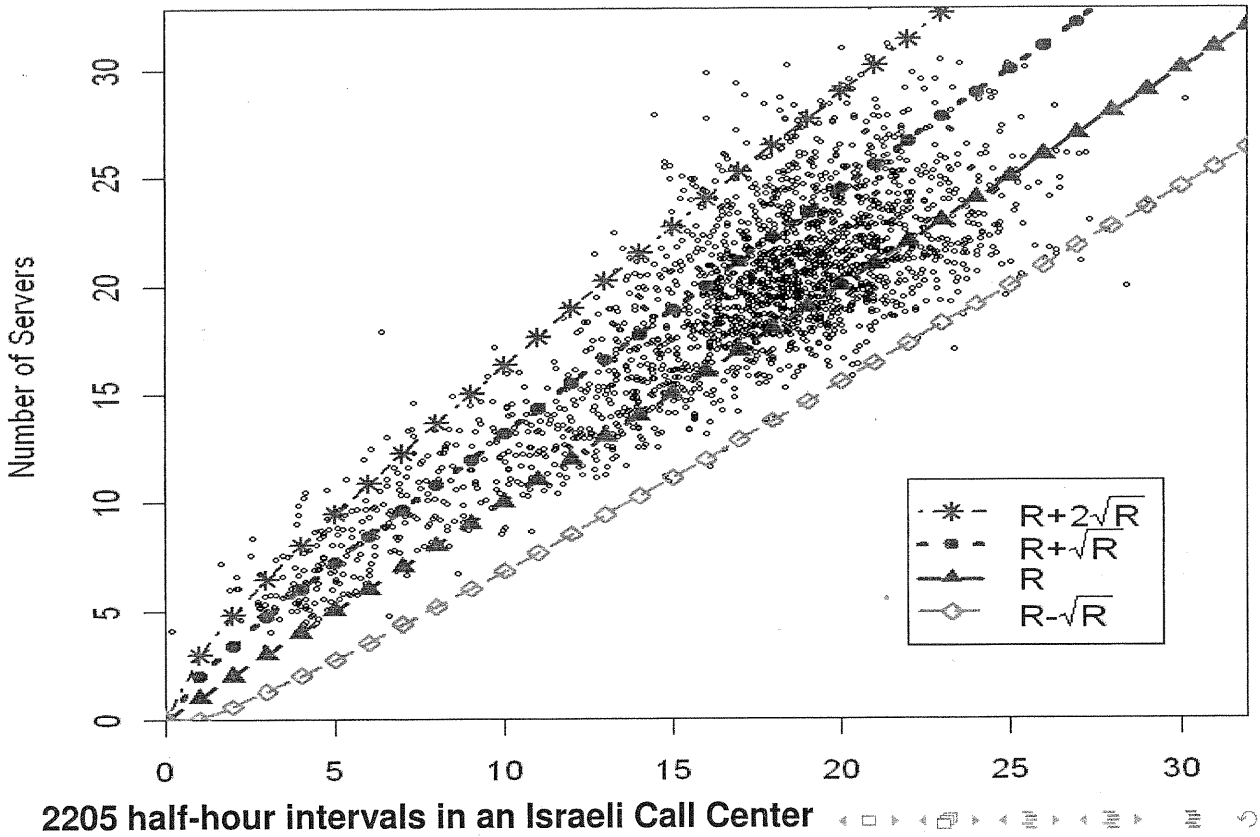
QED Call Center: Staffing (N) vs. Offered-Load (R)

IL Telecom; June-September, 2004; w/ Nardi, Plonski, Zeltyn

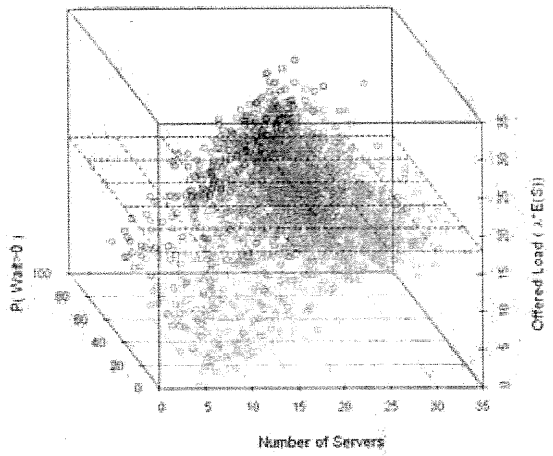


2205 half-hour intervals in an Israeli Call Center

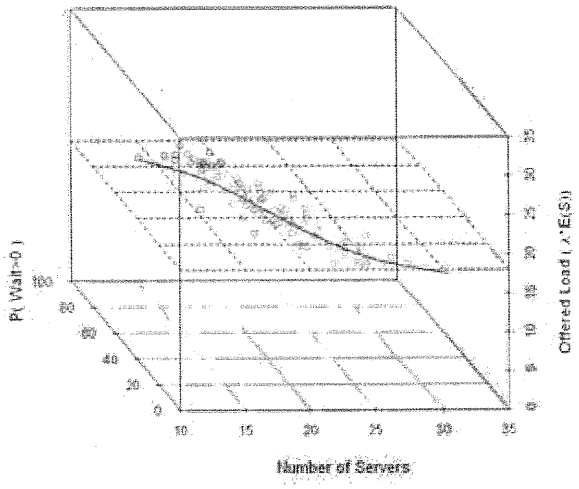
QED Call Center: Performance

Large Israeli Bank

P{$W_q > 0$} vs. (R, N)

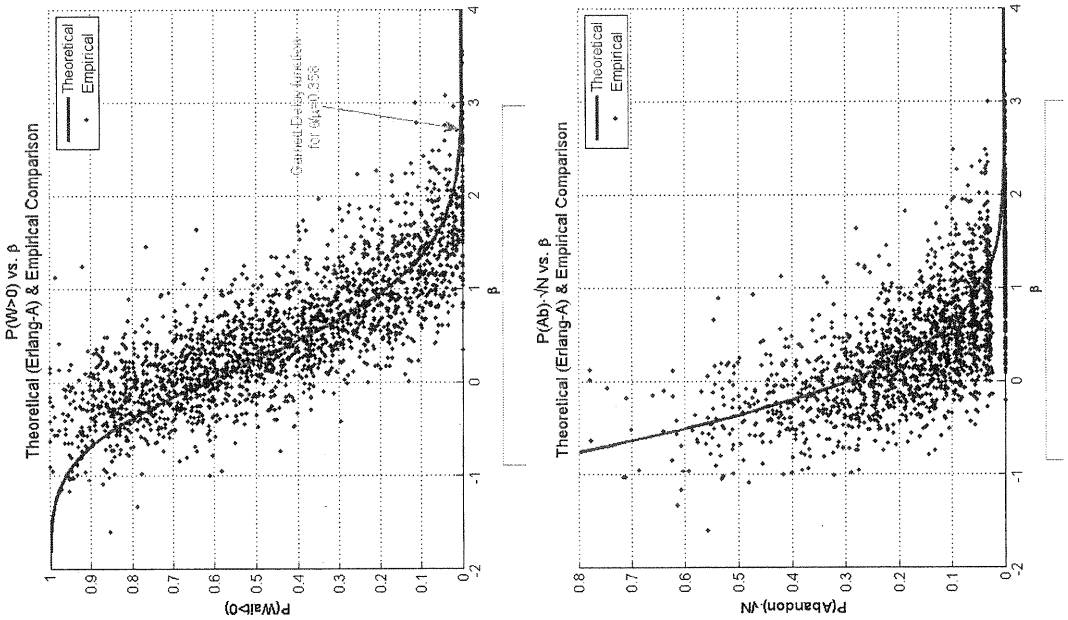


R-Slice: P{$W_q > 0$} vs. N



3 Operational Regimes:

- **QD**: $\leq 25\%$
- **QED**: $25\% - 75\%$
- **ED**: $\geq 75\%$



Operational Regimes: Scaling, Performance, w/ I. Gurvich & J. Huang

Erlang-A	Conventional scaling			MS scaling				NDS scaling		
μ fixed	Sub	Critical	Super	QD	QED	ED	ED+QED	Sub	Critical	Super
Offered load per server	$\frac{1}{1+\delta}<1$	$1-\frac{\beta}{\sqrt{n}}\approx 1$	$\frac{1}{1-\gamma}>1$	$\frac{1}{1+\delta}$	$1-\frac{\beta}{\sqrt{n}}$	$\frac{1}{1-\gamma}$	$\frac{1}{1-\gamma}-\beta\sqrt{\frac{1}{n(1-\gamma)^3}}$	$\frac{1}{1+\delta}$	$1-\frac{\beta}{n}$	$\frac{1}{1-\gamma}$
Arrival rate λ	$\frac{\mu}{1+\delta}$	$\mu-\frac{\beta}{\sqrt{n}}\mu$	$\frac{\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu\sqrt{n}$	$\frac{n\mu}{1-\gamma}$	$\frac{n\mu}{1-\gamma}-\beta\mu\sqrt{\frac{n}{(1-\gamma)^3}}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu$	$\frac{n\mu}{1-\gamma}$
Number of servers	1			n				n		
Time-scale	n			1				n		
Abandonment rate	θ/n			θ				θ/n		
Staffing level	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1-\gamma)+\beta\sqrt{\frac{\lambda}{\mu}}$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta$	$\frac{\lambda}{\mu}(1-\gamma)$
Utilization	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$	1	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{(1-\alpha_2)\hat{\beta}+\alpha_2 h(\hat{\beta})}{\sqrt{n}}$	1	1	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{n}$	1
$\mathbb{E}(Q)$	$\frac{\alpha_1}{\delta}$	$\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta^2}\varrho^n\frac{1}{\sqrt{n}}$	$\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]\alpha_2$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{n\mu}{\theta(1-\gamma)}(\gamma-\frac{\beta}{\sqrt{n(1-\gamma)}})$	$o(1)$	$n\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]$	$\frac{n^2\mu\gamma}{\theta(1-\gamma)}$
$\mathbb{P}(Ab)$	$\frac{1}{n}\frac{1+\delta}{\delta}\frac{\theta}{\mu}\alpha_1$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta})-\hat{\beta}]$	γ	$\frac{1}{\sqrt{2\pi}}\frac{\theta}{\mu}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta})-\hat{\beta}]\alpha_2$	γ	$\gamma-\frac{\beta\sqrt{1-\gamma}}{\sqrt{n}}$	$o(\frac{1}{n^2})$	$\frac{1}{n}\sqrt{\frac{\theta}{\mu}}[h(\hat{\beta})-\hat{\beta}]$	γ
$\mathbb{P}(W_q>0)$	$\alpha_1\in(0,1)$	≈ 1		$\frac{1}{\sqrt{2\pi}}\frac{1+\delta}{\delta}\varrho^n\frac{1}{\sqrt{n}}\approx 0$	$\alpha_2\in(0,1)$	≈ 1	≈ 1	≈ 0	≈ 1	
$\mathbb{P}(W_q>T)$	$\alpha_1 e^{-\frac{\delta}{1+\delta}\mu t}$	$1+O(\frac{1}{\sqrt{n}})$	$1+O(\frac{1}{n})$	≈ 0		$\bar{G}(T)1_{\{G(T)<\gamma\}}$	α_3, if $G(T)=\gamma$	≈ 0	$\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$	$1+O(\frac{1}{n})$
Congestion $\frac{\mathbb{E}W_q}{\mathbb{E}S}$	$\alpha_1\frac{1+\delta}{\delta}$	$\sqrt{n}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]$	$n\mu\gamma/\theta$	$\frac{1}{\sqrt{2\pi}}\frac{(1+\delta)^2}{\delta^2}\varrho^n\frac{1}{n^{3/2}}$	$\frac{1}{\sqrt{n}}\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]\alpha_2$	$\mu\int_0^{x^*}\bar{G}(s)ds$	$\mu\int_0^{x^*}\bar{G}(s)ds-\frac{\mu\beta\sqrt{1-\gamma}}{h_G(x^*)\sqrt{n}}$	$o(\frac{1}{n})$	$\sqrt{\frac{\mu}{\theta}}[h(\hat{\beta})-\hat{\beta}]$	$n\mu\gamma/\theta$

- $\delta>0, \gamma\in(0,1)$ and $\beta\in(-\infty,\infty)$;
- QD:$\varrho=\frac{1}{1+\delta}e^{\frac{\delta}{1+\delta}}<1$;
- ED (ED+QED): $G(x^*)=\gamma$;
- QED:$\alpha_2=[1+\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{h(-\hat{\beta})}]^{-1}$, here $\hat{\beta}=\beta\sqrt{\frac{\mu}{\theta}}$ and $h(x)=\frac{\phi(x)}{\Phi(x)}$;
- ED+QED:$\alpha_3=\bar{G}(T)\bar{\Phi}(\beta\sqrt{\frac{\mu}{g(T)}})$;
- Conventional: critical: $\mathbb{P}(W>T)=\mathbb{P}(\frac{W}{\sqrt{n}}>\frac{T}{\sqrt{n}})$, super: $\mathbb{P}(W>T)=\mathbb{P}(\frac{W}{n}>\frac{T}{n})$; NDS: Super: $\mathbb{P}(W>T)=\mathbb{P}(\frac{W}{n}>\frac{T}{n})$.

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: Occupancy $\approx 1 - \dfrac{\beta + \gamma}{\sqrt{n}}$;

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

in which case

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

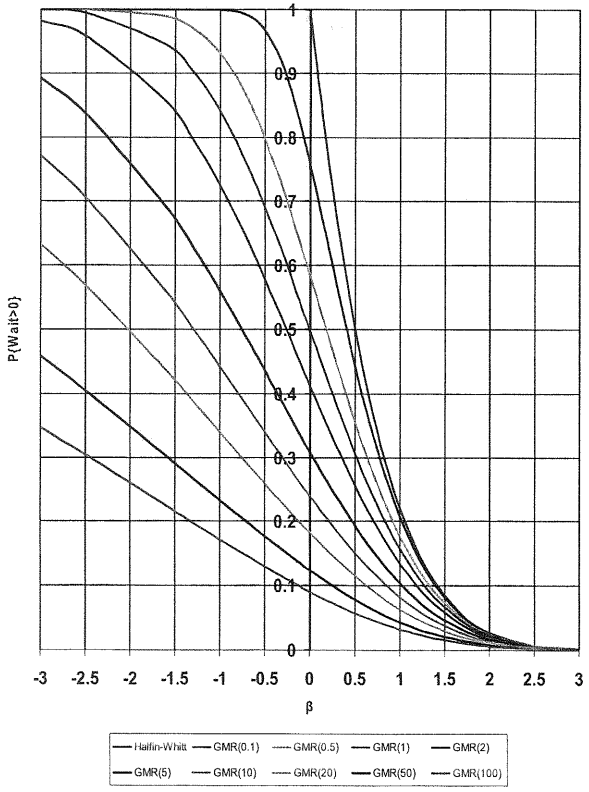
which we call the **Garnett Delay-Function(s)**;

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

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

Erlang-A: The Garnett Delay-Functions

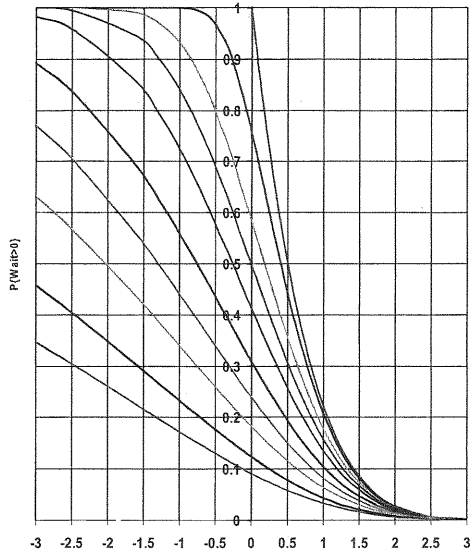
$P\{W_q > 0\}$ vs. the QOS parameter β, for varying patience θ/μ.



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.

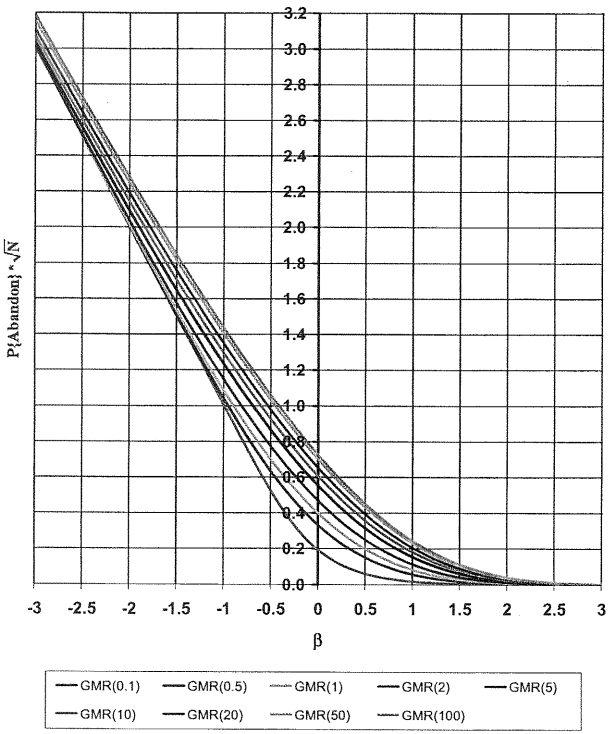
Understanding the Garnett Functions



- **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

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

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

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

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

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

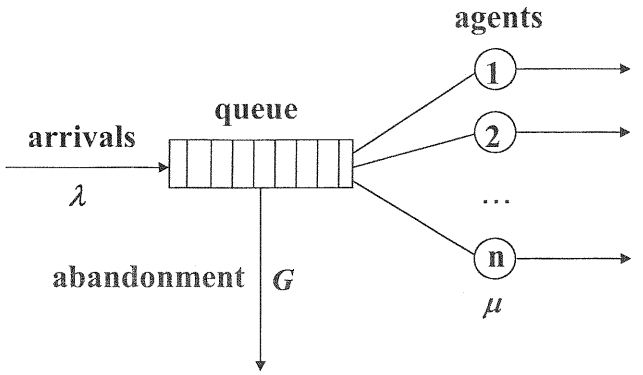
$$\mathrm{P}\left\{\mathrm{Ab} \,\middle|\, \frac{W_q}{\mathrm{E}[S]} > \frac{t}{\sqrt{n}}\right\} \approx \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h\left(\hat{\beta} + t\sqrt{\frac{\theta}{\mu}}\right) - \hat{\beta}\right]$$

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

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.

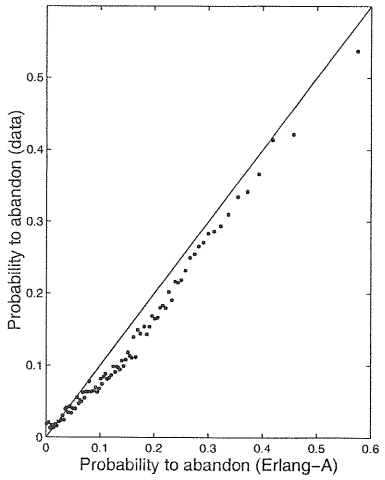
Erlang-A: Fitting a Simple Model to a Complex Reality

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

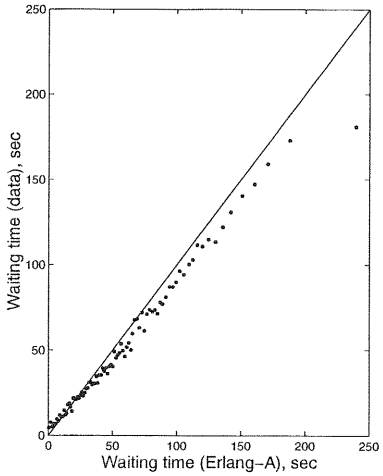
YES!

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

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

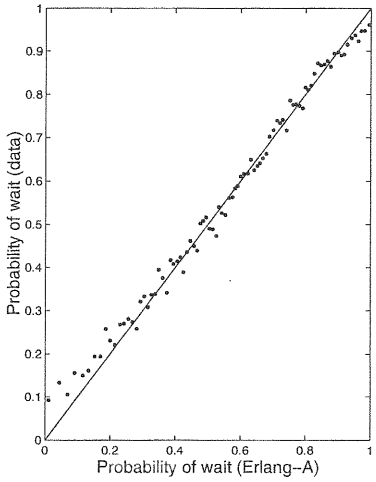


$\mathrm{E}[W_q]$



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.**

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:

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

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

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

- 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 $\approx \delta$ (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).

Class 13

QED Qs - Part II: Erlang-A.

QED Q's: Economies of Scale; Staffing Moderate-to-Large Service Operations

Erlang-A (Abandonment); QEDing Palm's model;

Optional Reading Assignment: Read the article "Healthcare Call Centers: A Technology Migration", by Howard Bernett. Pay special attention to the following:

1. The Call Center Maturity Model (Figure 1);
2. The calls flow within the call centers (Figure 2), starting with the IVR, through the triage nurse, then to one of the advice groups (internal, pediatrics, obstetrics and gynecology) or appointment agents;
3. In regard to the last paragraph of the article, recall the article

   http://iew3.technion.ac.il/serveng/Lectures/Retail.pdf

   in which an attempt was made to define the Industrial Engineer of the Future.

Recitation 14: Introduction to operation regime. Shift scheduling.

HW 11: Empirical Analysis of a call Center via SEEStat (and the Offered-Load). This HW is based on real data, which you will be analyzing with SEEStat. You will first identify problems with the operation of a call center, and then you will find staffing remedies for the difficulties found. The latter will require the use of 4CC.
A central role in the homework is the notion of Offered-Load, especially its time-varying version. There is also a part of the homework where you will check the validity of some congestion laws that were studied in class.

The due date for HW 11 is August 5, 2012.

Final Exam: $\leq$**50%** of the Final Grade. Its structure, is as follows:

Question 1: From previous exams (see our WebSite), or from Recitations, or from Lectures, of very similar to one of these.

Question 2: From Homeworks.

[Those who know well the material from Lectures, Recitations, Homeworks, are very likely to get a final grade of at least 75-80.]

Question 3: A "Practical Question" with some theoretical insight. [With Question 3 answered well, one can get to a final grade of at least 85-90.]

Question 4: A "Theoretical Question" that requires deeper understanding. [Answering well Question 4 is required in order to get to the levels of 95-100 final grade.]

Topics that were left out, or just touched on:

- Skills-Based Routing
- Queueing Networks: Jackson and Non-Parametric
- New-Service Development (or Service-Engineering in Germany)
- Internet-based services (or Contact Centers)
- Appointments - managing demand (Hall, Chapter 8)
- Service Quality
- Forecasting/smoothing (F&F, Chapter 14)
- Location and (functional) design of service facilities (F&F, Chapters 6, 7)
- Marketing (Lovelock, who also has a book dedicated to the subject)
- Human resource management (Lovelock)
- Technology; Automation
- Convergence of Service and Manufacturing:
  Field service, preventive maintenance, supply chain, life-time value
- Significant Service Sectors:
  - Health, Hospitality (Tourism), Financial, Transportation,
  - Telecommunication, Education, Professional Services,...

What's next?

- The "New-Age Industrial Engineer":
  - Industrial Engineers in Services: Banks, Hospitals, Government, etc.
  - Industrial Engineers in the the interface "Manufacturing – Services":
  - Consulting
  - Startups
- Research: Graduate Programs (Technion, Abroad); M.SC., Ph.D., TA'ing.
- Teaching:
  - Projects (Practical but Theoretically-Based)
  - Further Courses: Deeper (Q-Theory, Stochastic Processes); Broader (CRM, HRM, IE)

PLEASE STAY IN TOUCH, ESPECIALLY IF YOU WORK IN "SERVICES" OR ITS RELATIVES (MY GUESS - WITH PROBABILITY 0.75.)