QED Q's:

Quality- and Efficiency-Driven Queues

with a focus on Call/Contact Centers

Avishai Mandelbaum

Technion, Haifa, Israel

http://ie.technion.ac.il/serveng

TAU, Stat + OR, January 2008

Based on joint work with Students, Colleagues, . . .

Technion SEE Lab: P. Feigin, S. Zeltyn, V. Trofimov, RA's, . . .

Contents

- Introduction to Service Science / Engineering / **QED Q's**
- The Anatomy of "Waiting for Service"
- The Basic (Operational) Call-Center Model:
  Palm/Erlang-A (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?**
  Framework - Asymptotic Regimes: **QED, ED, ED+QED**
- Explain **Practice**: "Right Answers for the Wrong Reasons"
- Technion's **SEE** (Service Enterprise Engineering): **DataMOCCA**

Main Messages

1. Simple Useful Models at the Service of **Complex Realities**.

Note: Useful must be Simple; Simple rooted in **deep** analysis.

2. Data-Based Research & Teaching is a Must & Fun.

Supported by **DataMOCCA** = Data **MO**dels for **C**all **C**enter **A**nalysis.

3. Human Complexity requires the **Basic-Research Paradigm** (Physics, . . .): **Measure, Model, Experiment, Validate, Refine, etc.**

4. Ancestors & **Practitioners** often knew/apply the **“right answer"**: simply did/do not have our tools/desire/need to prove it so.

Supported by **Erlang (1910+), Palm (1940+)**,..., thoughtful managers.

5. Service Science / Management / Engineering are emerging **Academic Disciplines**. For example, universities and

USA NSF (SEE), IBM (SSME), Germany IAO (ServEng), ...

Background Material (Downloadable)

- Technion's "**Service-Engineering" Course** ($\geq$ 1995): http://ie.technion.ac.il/serveng

- Google Scholar - search <Call Centers>:
  - Gans (U.S.A.), Koole (Europe), and M. (Israel): "Telephone Call Centers: Tutorial, Review and Research Prospects." MSOM, 2003.
  - Brown, Gans, M., Sakov, Shen, Zeltyn, Zhao: "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective." JASA, 2005.

- Trofimov, Feigin, M., Ishay, Nadjharov: "DataMOCCA: Models for Call/Contact Center Analysis." Technion Report, 2004-2006.

- M. "Call Centers: Research Bibliography with Abstracts." Version 7, December 2006.

Queueing Science: Data-Based QED's Q's

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency** **must** be traded off against each other.

For example, **M/M/1 in heavy-traffic**: **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Yet, heavily-loaded queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait **"seconds"** for **minutes** service;
- **Transportation**: Search **"minutes"** for **hours** parking;
- **Hospitals**: Wait **"hours"** in ED for **days** hospitalization in IW's;

and, moreover, a significant fraction are not delayed in queue. (For example, in well-run call-centers, **50%** served "immediately", along with over **90%** agents' utilization, is not uncommon) **?**

Prerequisite: Data

Averages Prevalent.

But I need data at the level of the **Individual Transaction**: For each service transaction (during a phone-service in a call center, or a patient's stay in a hospital), its **operational history** = time-stamps of events.

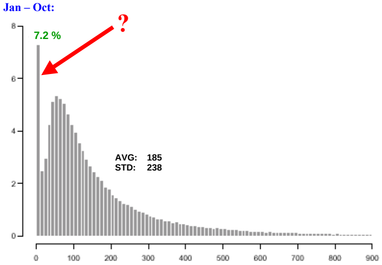
Sources: **"Service-floor"** (vs. Industry-level, Surveys, . . .)

- Administrative (Court, via "paper analysis")
- Face-to-Face (Bank, via bar-code readers)
- Telephone (Call Centers, via ACD / CTI)
- Expanding:
  - Hospitals (via RFID)
  - IVR (VRU), internet, chat (multi-media)
  - Operational + Financial + Marketing / Clinical history

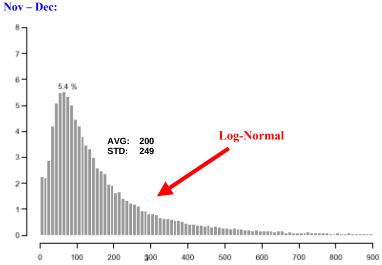
Beyond Averages (+ The Human Factor)

Histogram of Service Times in an Israeli Call Center

January-October



November-December



- **7.2% Short-Services:** Agents' “Abandon" (improve bonus, rest)
- **Distributions**, not only Averages, must be measured.
- **Lognormal** service times prevalent in call centers (Why?)

Present Focus: Call Centers

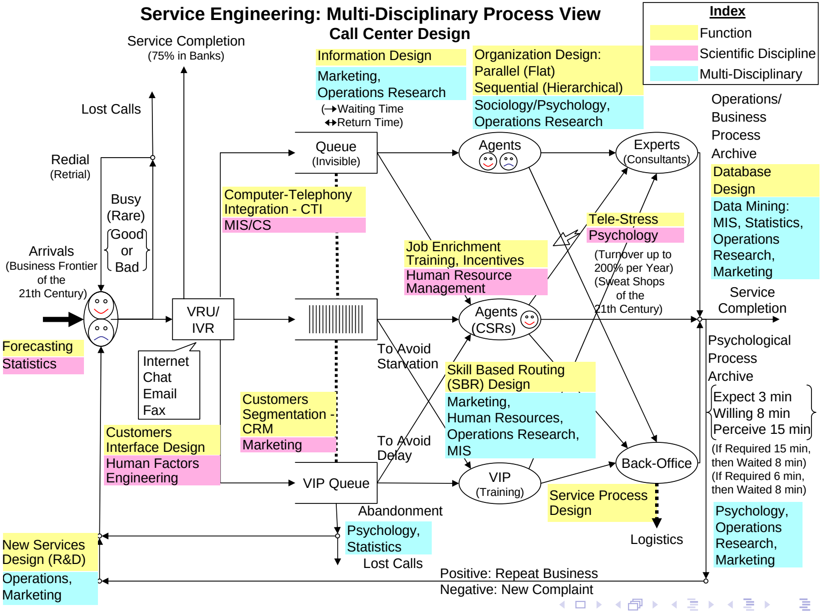
U.S. Statistics (Relevant Elsewhere)

- Over 60% of annual business volume via the telephone
- 100,000 – 200,000 call centers
- 3 – 6 million employees (**2% – 4% workforce**)
- 1000's agents in a "single" call center = 70 % costs.
- 20% annual growth rate
- \$200 – \$300 billion annual expenditures

Call-Center Environment: Service Network

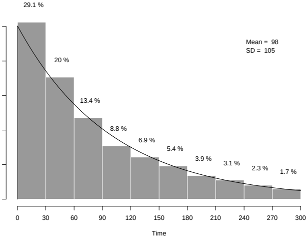
Call-Centers: “Sweat-Shops of the 21st Century"

Call-Center Network: Gallery of Models

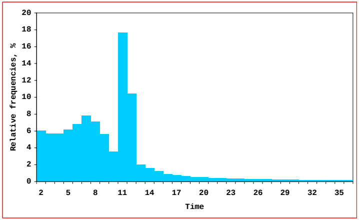


Beyond Averages: Waiting Times in a Call Center

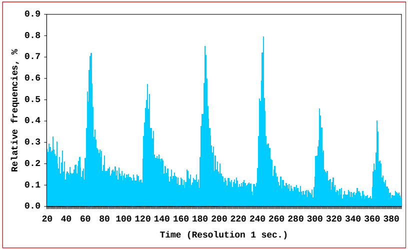
Small Israeli Bank



Large U.S. Bank



Medium Israeli Bank



The "Anatomy of Waiting" for Service

Common Experience:

- Expected to wait 5 minutes, Required to 10,
- Felt like 20, Actually waited 10,
- ... etc.

An attempt at "Modeling the Experience":

1.	Time that a customer	**expects** to wait	
2.		**willing** to wait	((Im)Patience: τ)
3.		**required** to wait	(Offered Wait: V)
4.		**actually** waits	($W_q = \min(\tau, V)$)
5.		**perceives** waiting.	

Experienced customers $\Rightarrow$ Expected = Required

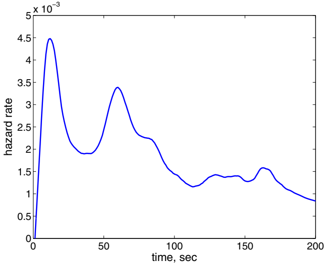
"Rational" customers $\Rightarrow$ Perceived = Actual.

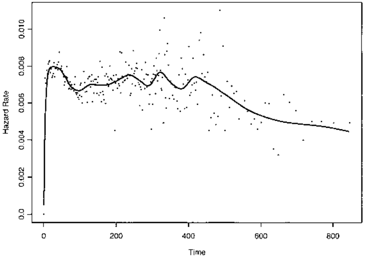
Then left with (τ, V).

Call Center Data: Hazard Rates (Un-Censored)

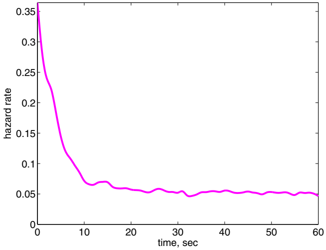
(Im)Patience Time τ | **Required/Offered Wait V**

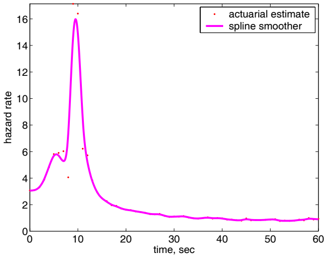
Israel





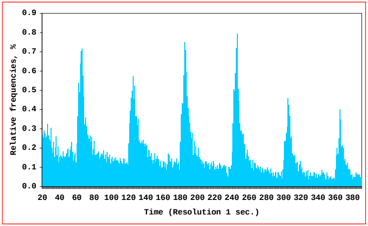
U.S.





Note: 5% abandoning $\Rightarrow$ 95% (im)patience-observations **censored** !

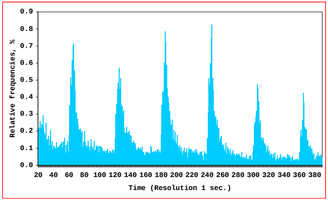
A “Waiting-Times" Puzzle at a Large Israeli Bank



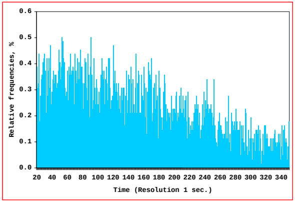
Peaks Every 60 Seconds. Why?

- Human: **Voice-announcement** every 60 seconds.
- System: **Priority-upgrade** (unrevealed) every 60 sec’s (Theory?)

Served Customers



Abandoning Customers



Models for Performance Analysis

- **(Im)Patience**: r.v. τ = Time a customer is **willing to wait**
- **Offered-Wait**: r.v. V = Time a customer is **required to wait** (= Waiting time of a customer with infinite patience).

- **Abandonment** $= \{\tau \leq V\}$
- **Service** $= \{\tau > V\}$
- **Actual Wait** $W_q = \min\{\tau, V\}$.

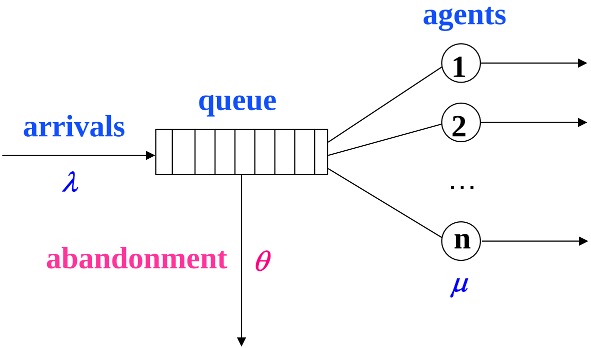
Modeling: τ = **input** to the model, V = **output**.

Operational Performance-Measure calculable in terms of (τ, V):

- eg. **Avg. Wait** = $E[\min\{\tau, V\}]$ ($E[W_q|\text{Served}] = E[V|\tau > V]$)
- eg. **% Abandon** = $P\{\tau \leq V\}$ ($P\{5 \text{ sec} < \tau \leq V\}$)

Application: **Staffing – How Many Agents?** (then: When? Who?)

The Basic Staffing Model: Erlang-A (M/M/N + M)



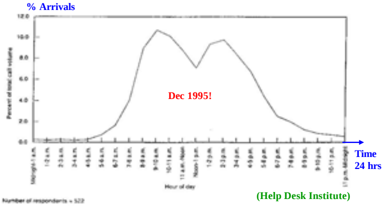
Erlang-A **(Palm 1940's)** = **Birth & Death Q, with parameters:**

- λ – **Arrival** rate (Poisson)
- μ – **Service** rate (Exponential)
- θ – **Impatience** rate (Exponential)
- n – Number of **Service-Agents**.

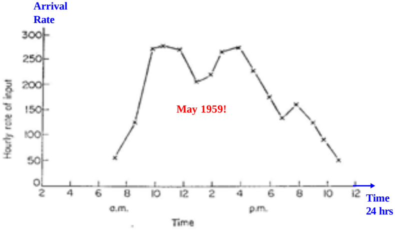
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

Dec. **1995** (U.S. 700 Helpdesks)



May **1959** (England)



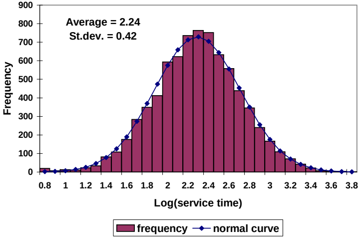
November **1999** (Israel)

Observation:

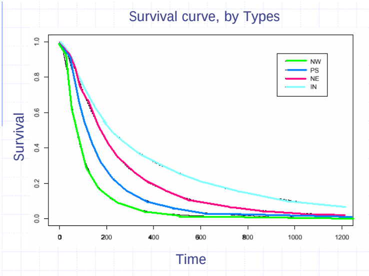
Peak Loads at **10:00 & 15:00**

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



Survival-Functions by Service-Class

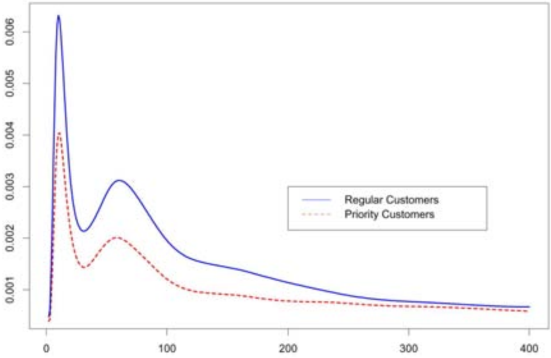


- **New** Customers: 2 min (NW);
- **Regulars**: 3 min (PS);
- **Stock**: 4.5 min (NE);
- Tech-Support: 6.5 min (IN).

Observation: **VIP** require **longer service** times.

(Im)Patience while Waiting (Palm 1943-53)

Irritation $\propto$ Hazard Rate of (Im)Patience Distribution
Regular over VIP Customers – Israeli Bank



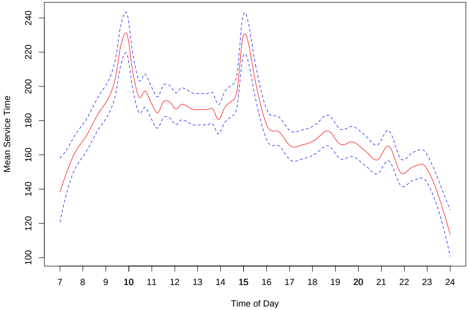
- **Peaks** of abandonment at times of **Announcements**
- **Call-by-Call Data (DataMOCCA)** required (& Un-Censoring).

Observation: **VIP** are **more patient** (Needy)

A “Service-Time" Puzzle at an Israeli Bank

Inter-related Primitives

Average Service Time over the Day – Israeli Bank



Prevalent: **Longest services** at **peak-loads** (10:00, 15:00). **Why?**

Explanations:

- Common: Service protocol different (longer) during peak times.
- Operational: The needy abandon less during peak times; hence the **VIP remain** on line, with their **long service** times.

Erlang-A: Practical Relevance?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (various patterns observed).

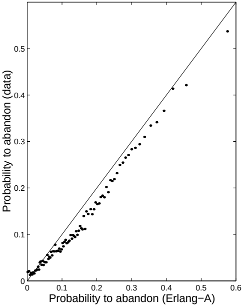
- Building Blocks need **not be independent** (eg. long wait possibly implies long service)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment)
- $\cdots$, and more.

Question: **Is Erlang-A Practically Relevant?**

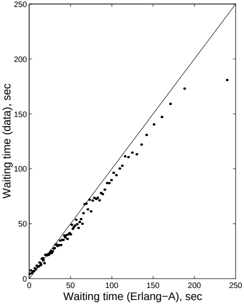
Erlang-A: Fitting a Simple Model to a Complex Reality

- **Small Israeli Banking Call-Center** (10 agents)
- (Im)Patience (θ) estimated via P{Ab} / E[W_q]
- Graphs: **Hourly Performance vs. Erlang-A Predictions**, during 1 year (aggregating groups with 40 similar hours).

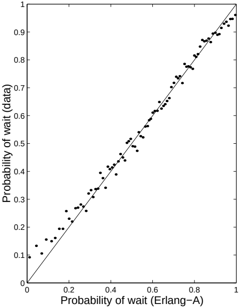
P{Ab}



E[W_q]



$P\{W_q > 0\}$



Erlang-A: Simple, but Not Too Simple

Further Natural Questions:

1. Why does Erlang-A practically work? justify **robustness**.
2. When does it fail? chart **boundaries**.
3. Generalize: time-variation, SBR, networks, uncertainty , . . .

Answers via **Asymptotic Analysis**, as load- and staffing-levels increase, which reveals model-essentials:

- **E**fficiency-**D**riven (**ED**) regime: Fluid models (deterministic)
- **Q**uality- and **E**fficiency-**D**riven (**QED**): Diffusion refinements.

Motivation: Moderate-to-large service systems (**100's - 1000's** servers), notably **call-centers**.

Results turn out **accurate** enough to also cover **10-20** servers. Important – relevant to **hospitals** (nurse-staffing: de Véricourt & Jennings, 2006), ...

Operational Regimes: Conceptual Framework

Assume: **Offered Load** $\boldsymbol{R} = \frac{\lambda}{\mu}$ **(**$= \lambda \times$ E**[*S*])** not too small.

QD Regime: $\boxed{N \;\approx\; R + \delta R}$ $\quad [(N - R)/R \to \delta$, as $N, \lambda \uparrow \infty]$

- Essentially **no** delays: $\quad [P\{W_q > 0\} \to 0]$.

ED Regime: $\boxed{N \;\approx\; R - \gamma R}$

- Garnett, M. & Reiman 2003
- Essentially **all** customers are delayed
- Wait same order as service-time; γ% Abandon (10-25%).

QED Regime: $\boxed{N \;\approx\; R + \beta\sqrt{R}}$

- Erlang 1913/24, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

QED+ED: $\boxed{N \;\approx\; (1 - \gamma)R + \beta\sqrt{R}}$

- Zeltyn & M. 2006
- QED refining ED to accommodate "timely-delays": $P\{W_q > T\}$.

QED Theory (Erlang '13; Halfin-Whitt '81; Garnett MSc; Zeltyn PhD)

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** $\quad N \approx R + \beta\sqrt{R}, \quad R = \lambda \times \text{E(S)}$ not small;

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

QED Approximations (Zeltyn, M. '06)

G – patience distribution,

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

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

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

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

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

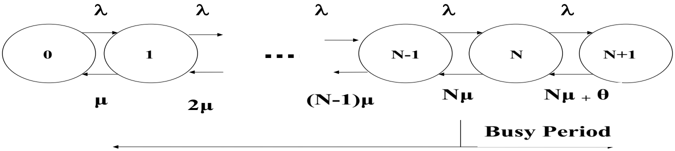
Here

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

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

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

QED Intuition via Excursions: Busy/Idle Periods



$Q(0) = N$: all servers busy, no queue.

Let $T_{N,N-1} =$ Busy Period (down-crossing $N \downarrow N-1$)

$T_{N-1,N} =$ Idle Period (up-crossing $N-1 \uparrow N$)

Then
$$P(Wait > 0) = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1}$$

QED Intuition via Excursions: Asymptotics

Calculate
$$T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1}{\sqrt{N}} \cdot \frac{1/\mu}{h(-\beta)}$$
$$T_{N,N-1} = \frac{1}{N\mu\pi_+(0)} \sim \frac{1}{\sqrt{N}} \cdot \frac{\beta/\mu}{h(\delta)/\delta}, \qquad \delta = \beta\sqrt{\mu/\theta}$$

Both apply as $\sqrt{N}(1-\rho_N) \to \beta, \ -\infty < \beta < \infty.$

Hence,
$$P(Wait > 0) \sim \left[1 + \frac{h(\delta)/\delta}{h(-\beta)/\beta}\right]^{-1}.$$

Special case: $\mu = \theta$ (Impatient):

Then $\boldsymbol{Q} \overset{d}{=} \boldsymbol{M/M/\infty}$, since sojourn-time is $\exp(\mu = \theta)$.

If also $\beta = \mathbf{0}$ (Prevalent): $\boldsymbol{P\{Wait > 0\} \approx 1/2}$.

Process Limits (Queueing, Waiting)

- $\hat{Q}_N = \{\hat{Q}_N(t), t \geq 0\}$: stochastic process obtained by centering and rescaling:

$$\hat{Q}_N = \frac{Q_N - N}{\sqrt{N}}$$

- $\hat{Q}_N(\infty)$: stationary distribution of $\hat{Q}_N$

- $\hat{Q} = \{\hat{Q}(t), t \geq 0\}$: process defined by: $\hat{Q}_N(t) \xrightarrow{d} \hat{Q}(t)$.

$$\begin{tikzcd}
\hat{Q}_N(t) \arrow[r, "t \to \infty"] \arrow[d, "N \to \infty"'] & \hat{Q}_N(\infty) \arrow[d, "N \to \infty"] \\
\hat{Q}(t) \arrow[r, "t \to \infty"'] & Q(\infty)
\end{tikzcd}$$

Approximating (Virtual) Waiting Time

$$\hat{V}_N = \sqrt{N}\, V_N \Rightarrow \hat{V} = \left[\frac{1}{\mu}\,\hat{Q}\right]^+ \qquad \text{(Puhalskii, 1994)}$$

Operational Regimes: Rules-of-Thumb

Constraint	$P\{Ab\}$		$E[W]$		$P\{W > T\}$	
Offered Load	Tight 1-10%	Loose $\geq 10\%$	Tight $\leq 10\% E[\tau]$	Loose $\geq 10\% E[\tau]$	Tight $0 \leq T \leq 10\% E[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\% E[\tau]$ $5\% \leq \alpha \leq 50\%$
Small (10's)	QED	QED	QED	QED	QED	QED
Moderate-to-Large (100's-1000's)	QED	ED, QED	QED	ED, QED if $\tau \stackrel{d}{=} \exp$	QED	ED+QED

ED: $N \approx R - \gamma R \quad (0.1 \leq \gamma \leq 0.25)$.

QD: $N \approx R + \delta R \quad (0.1 \leq \delta \leq 0.25)$.

QED: $N \approx R + \beta\sqrt{R} \quad (-1 \leq \beta \leq 1)$.

ED+QED: $N \approx (1 - \gamma)R + \beta\sqrt{R} \quad (\gamma, \beta$ as above).

Back to "Why does Erlang-A Work?"

Theoretical Answer: $M_t^J/G/N_t + G \overset{d}{\approx} (M/M/N + M)_t\ , \quad t \geq 0.$

- **General Patience**: Behavior at the origin is all that matters.
- **General Services**: Empirical insensitivity beyond the mean.
- **Time-Varying Arrivals**: Modified Offered-Load approximations.
- **Heterogeneous Customers**: 1-D state collapse.

Practically: Why do (stochastic-ignorant) Call Centers work?

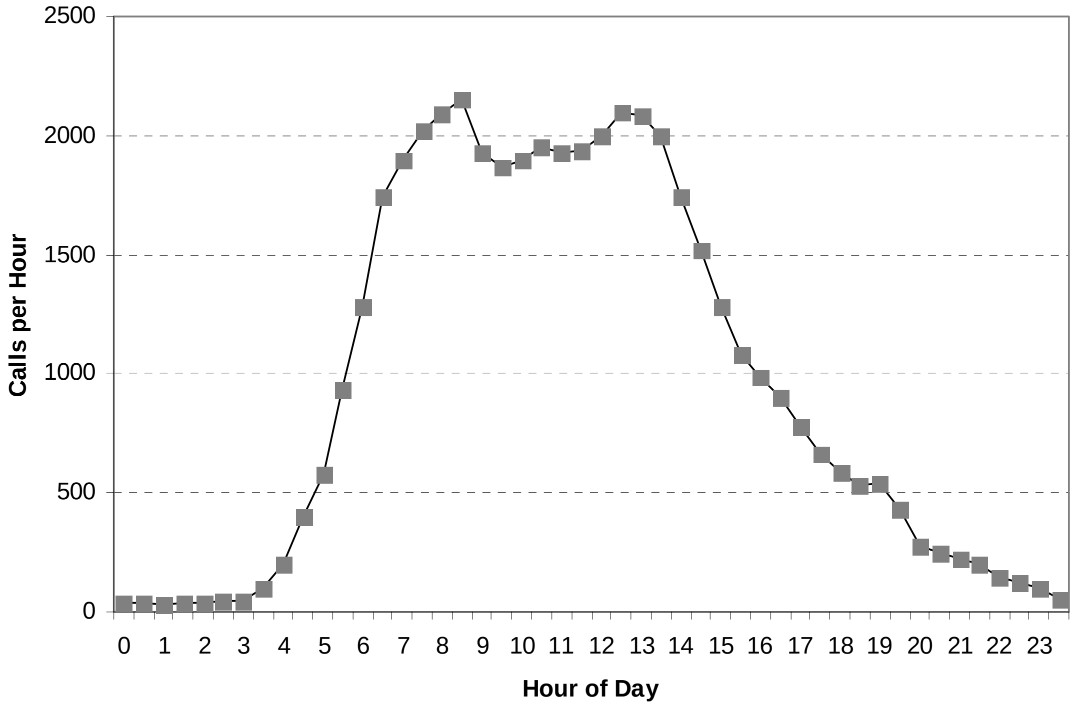
"The right answer for the wrong reason"

Example: "Real" Call Center

(The "Right Answer" for the "Wrong Reasons")

Time-Varying (two-hump) arrival functions common

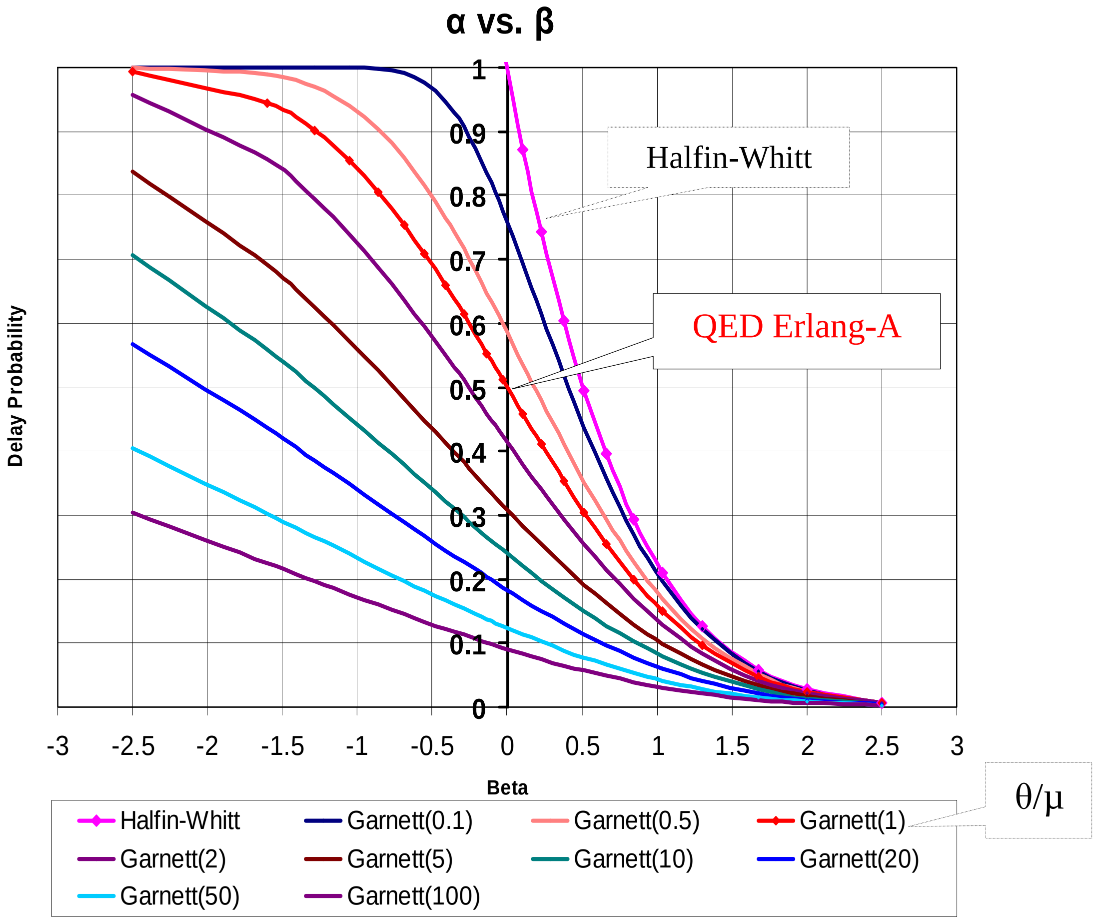
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



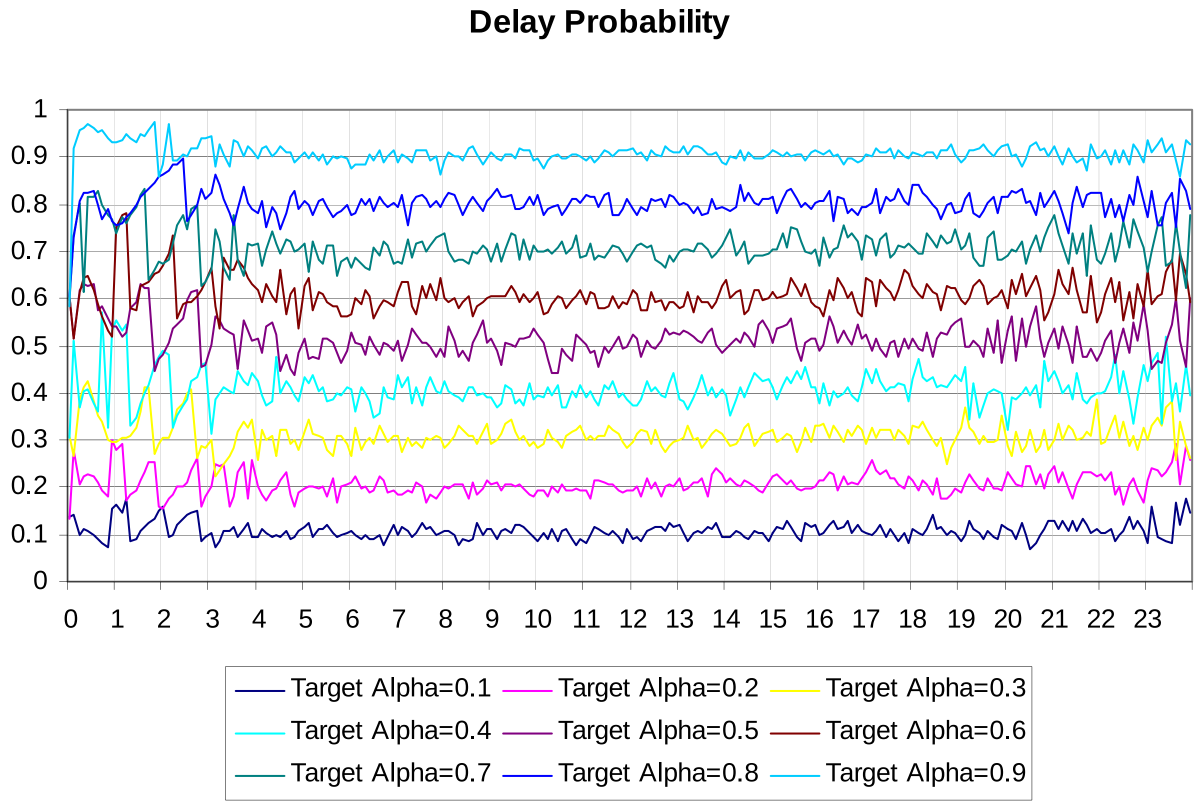
Assume: Service and abandonment times are both

Exponential, with mean 0.1 (6 min.)

HW/GMR Delay Functions

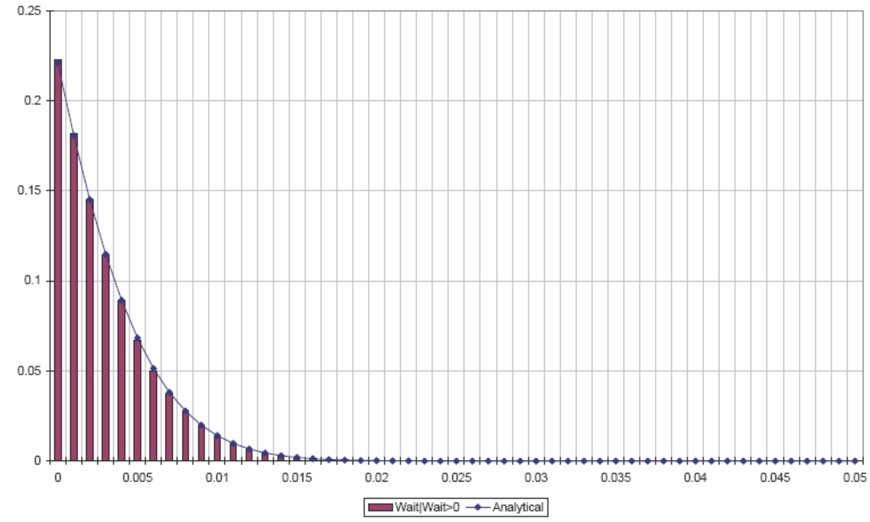


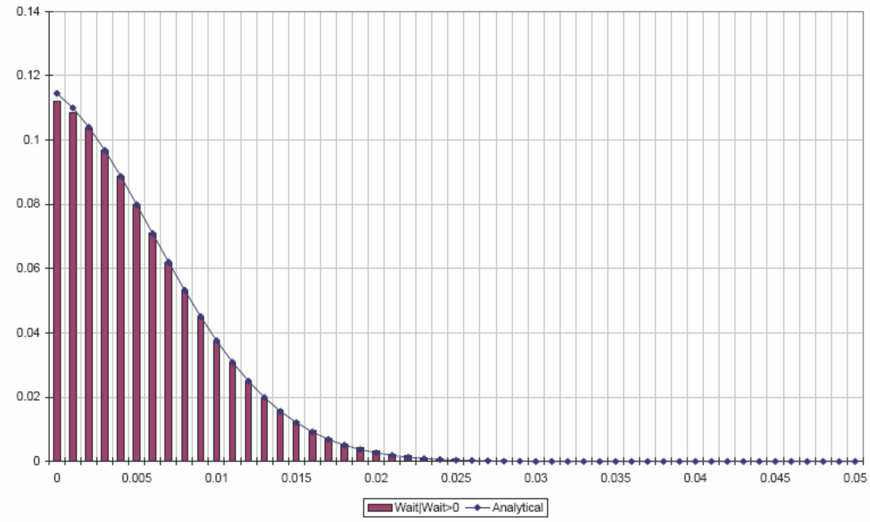
Delay Probability α

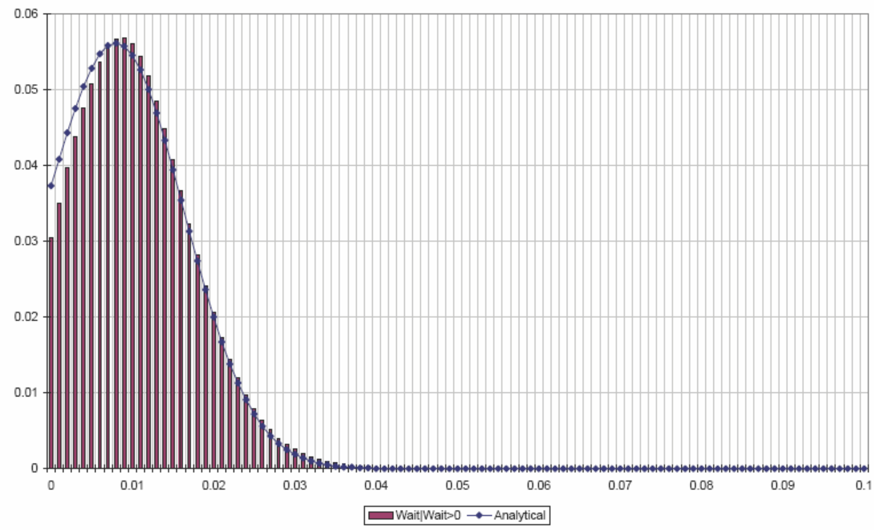


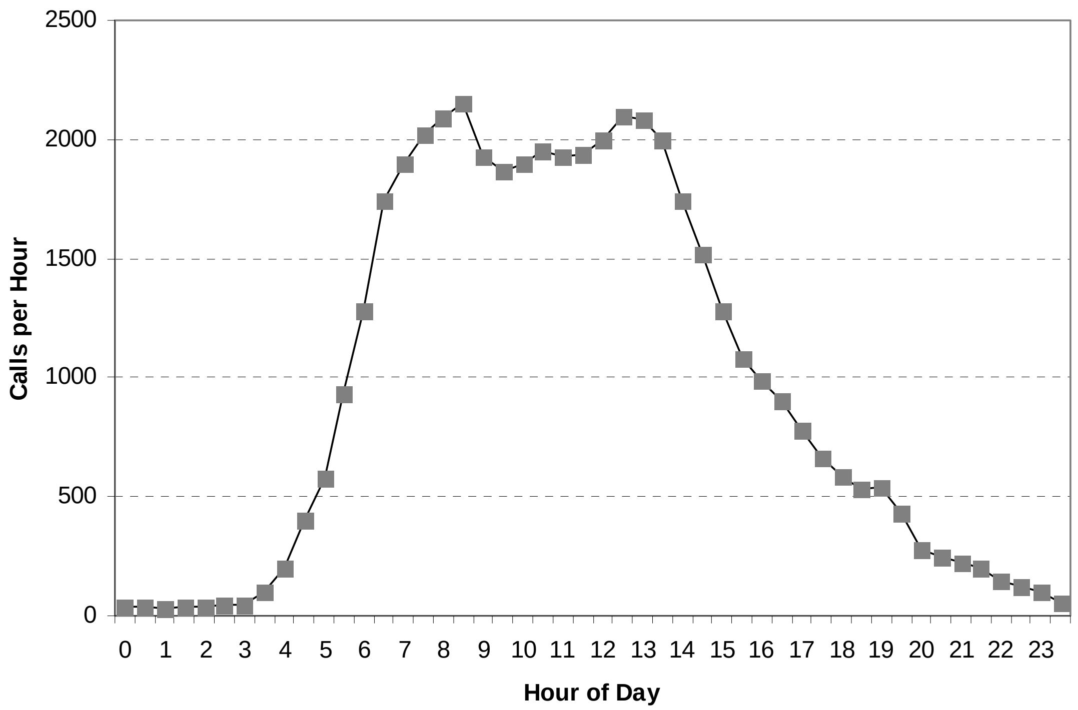
Real Call Center: Empirical waiting time, given positive wait

(1) α=0.1 (*QD*) **(2)** α=0.5 (*QED*) **(3)** α=0.9 (*ED*)

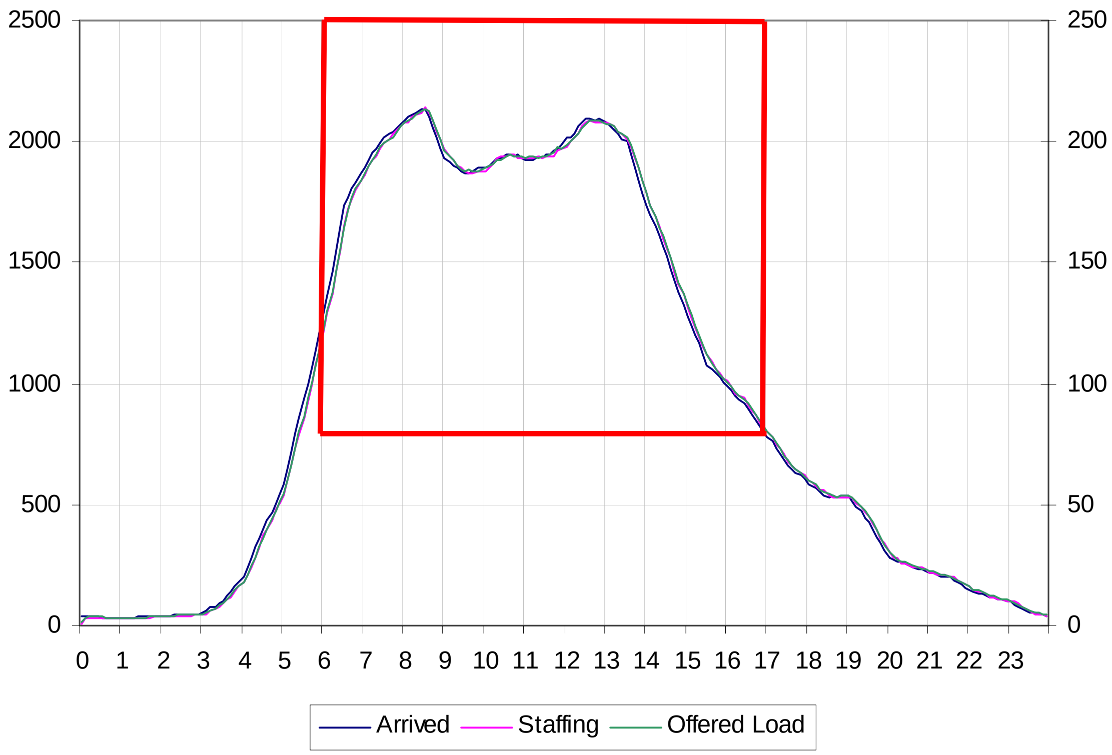








QED Staffing (β=0 iff α=0.5)



The Technion SEE Center / Laboratory

Technion S.E.E Center

Service Enterprise Engineering

DataMOCCA = Data MOdels for Call Center Analysis

- **Technion**: P. Feigin, V. Trofimov, Statistics / SEE Laboratory.
- **Wharton**: L. Brown, N. Gans, H. Shen (UNC).
- **industry**:
  - U.S. Bank: **2.5 years, 220M calls, 40M by 1000 agents**.
  - Israeli Cellular: **2.5 years, 110M calls, 25M calls by 750 agents; ongoing**.

Project Goal: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing and displaying **Call-by-Call-based Data / Information**.

System Components:

- **Clean Databases**: operational-data of individual calls / agents.
- **Graphical Online Interface**: easily generates graphs and tables, at varying resolutions (seconds, minutes, hours, days, months).

Free for academic adoption: **ask for a DVD (3GB)**.