QED Q's

(Service-Science & -Engineering of)

Quality- and Efficiency-Driven Queues

(Call/Contact Centers)

Avishai Mandelbaum

Technion, Haifa, Israel

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

CMU, Probability in (the **Service**) Industry, August 2007

Based on joint work with **Sergey Zeltyn,** . . .

Technion SEE Center / Lab: Paul Feigin, Valery Trofimov, RA's, . . .

Contents

- **Data-Based** Introduction
  - **Simple Models at the Service of Complex Realities:**
  - Courts, Banks, Hospitals, Call Centers, . . .

Contents

- **Data-Based** Introduction
  - **Simple Models at the Service of Complex Realities:**
  - Courts, Banks, Hospitals, Call Centers, . . .
- The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**

Contents

- **Data-Based** Introduction
  - **Simple Models at the Service of Complex Realities:**
  - Courts, Banks, Hospitals, Call Centers, . . .
- The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?** Robustness via asymptotic analysis that reveals operational regimes: **QED, ED, ED+QED**

Contents

- **Data-Based** Introduction
  - **Simple Models at the Service of Complex Realities:**
  - Courts, Banks, Hospitals, Call Centers, . . .
- The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?** Robustness via asymptotic analysis that reveals operational regimes: **QED, ED, ED+QED**
- And many (“stochastically-challenged") call centers work as well - Why? **“Right Answers for the Wrong Reasons"**

Contents

- **Data-Based** Introduction
  - **Simple Models at the Service of Complex Realities:**
  - Courts, Banks, Hospitals, Call Centers, . . .
- The Basic Call-Center Model: **Palm/Erlang-A** (M/M/N+M)
- Validating Erlang-A? All **Assumptions Violated**
- But **Erlang-A Works! Why?** Robustness via asymptotic analysis that reveals operational regimes: **QED, ED, ED+QED**
- And many (“stochastically-challenged") call centers work as well - Why? **“Right Answers for the Wrong Reasons"**
- “Appendix": Demo of **DataMOCCA**

Background Material (Downloadable)

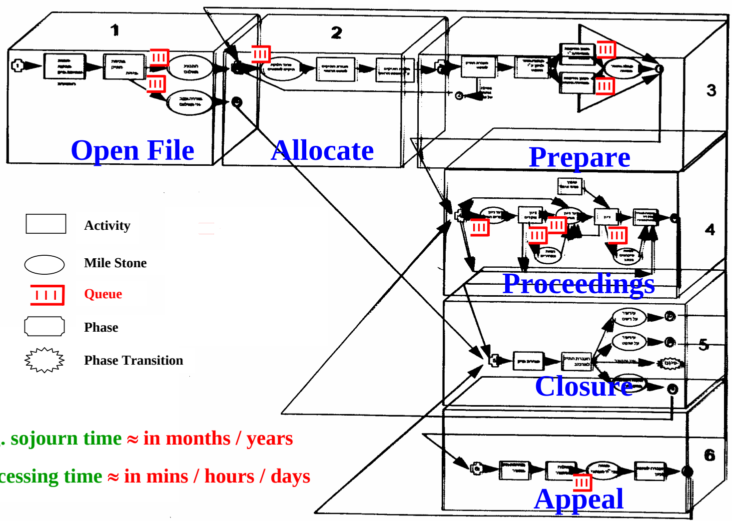
- Technion's "**Service-Engineering" Course** ($\geq$ 1995): http://ie.technion.ac.il/serveng

Background Material (Downloadable)

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

“Production of Justice" (Administrative) Network

Skills-Based-Routing at the Labor-Court in Haifa, Israel

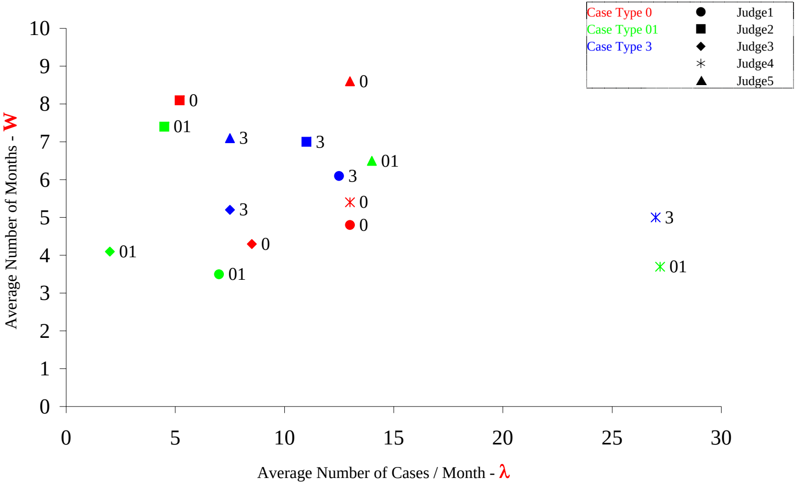


Avg. sojourn time ≈ in months / years

Processing time ≈ in mins / hours / days

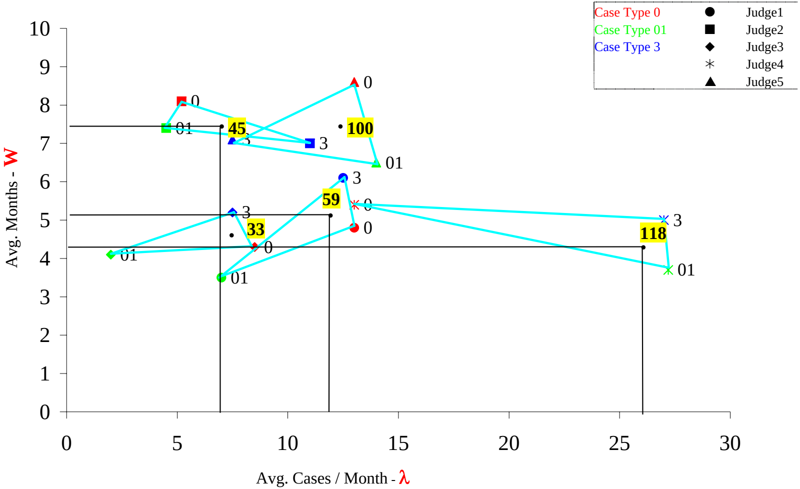
Operational Performance: 5 Judges, 3 Case-Types

Judges: The Best/Worst (Operational) Performer



Little's Law in Court (Creative Averaging)

Judges: The Best/Worst (Operational) Performer



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.

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.

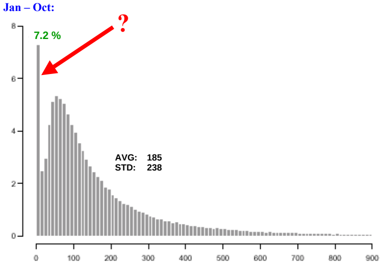
Sources: "Service-floor" (vs. Industry-level, Surveys, . . .)

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

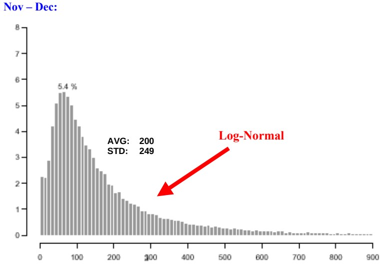
Beyond Averages: Service Times in a Call Center

Histogram of Service Times in an Israeli Call Center

January-October



November-December

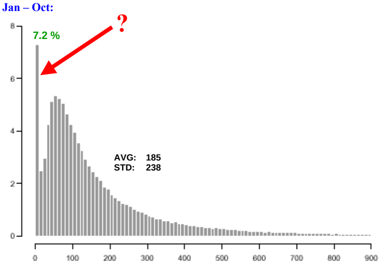


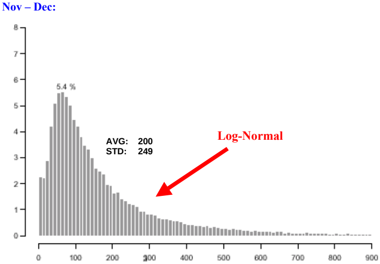
- **7.2% Short-Services:**

Beyond Averages: Service Times in a Call Center

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

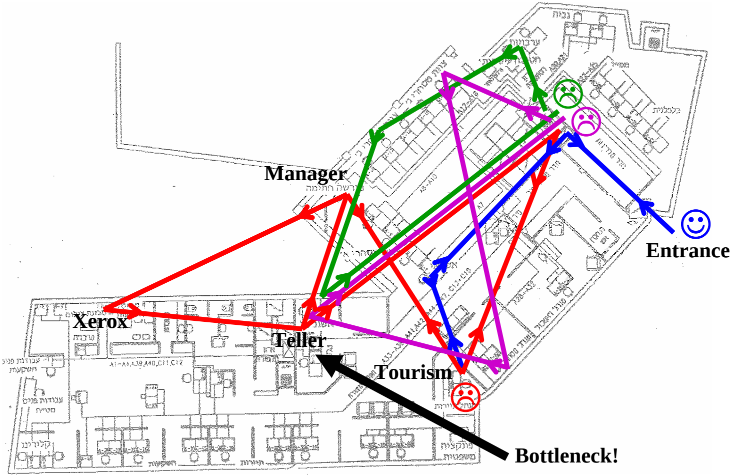
Measurements: Face-to-Face Services

23 Bar-Code Readers at a Bank Branch

Bank – 2nd Floor Measurements

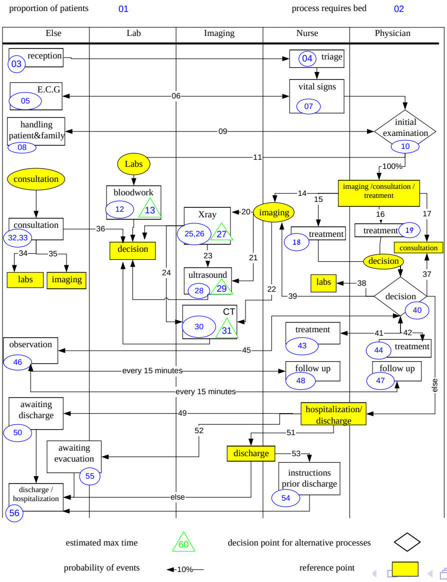
“Face-to-Face Services" Network

Bank Branch = Jackson Network



Hospital Network (Marmur, Sinreich)

Generic Emergency Department (RFID)



Present Focus: Call Centers

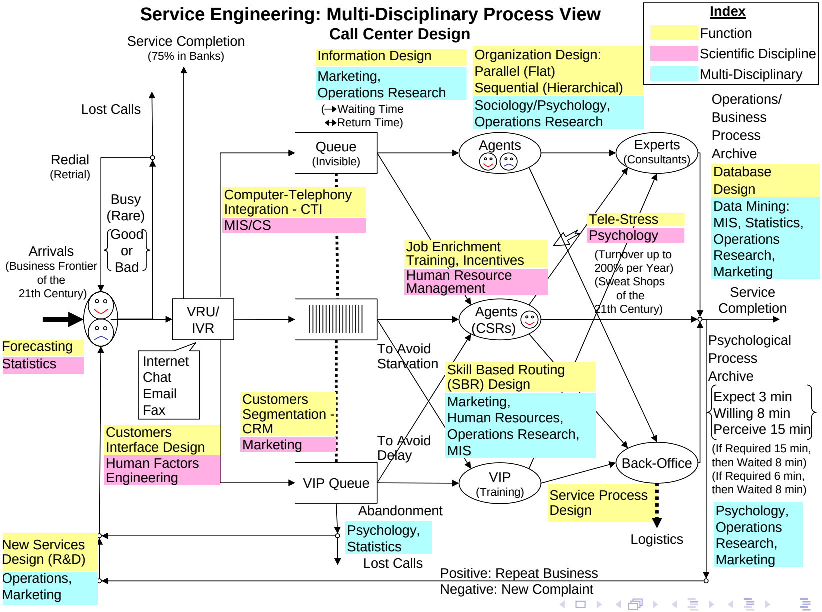
U.S. Statistics (Relevant Elsewhere)

- 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

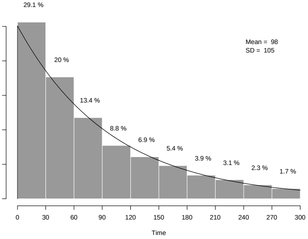
Call-Centers: “Sweat-Shops of the 21st Century"

Call-Center Network: Gallery of Models

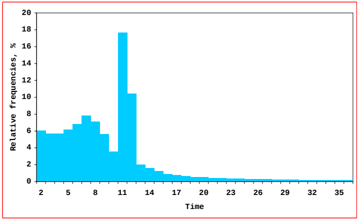


Beyond Averages: Waiting Times in a Call Center

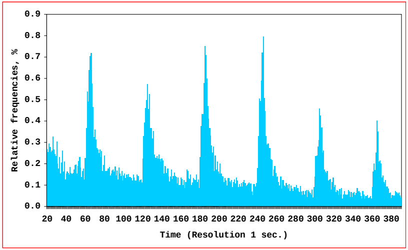
Small Israeli Bank



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.

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.	

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.	

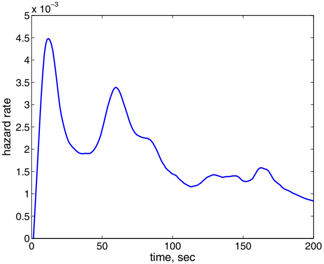
Experienced customers	$\Rightarrow$	Expected = Required
"Rational" customers	$\Rightarrow$	Perceived = Actual.

Then left with (τ, V).

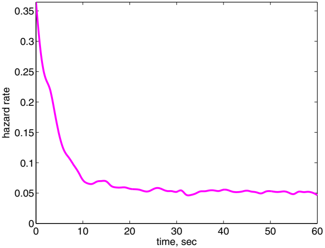
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time τ

Israel



U.S.

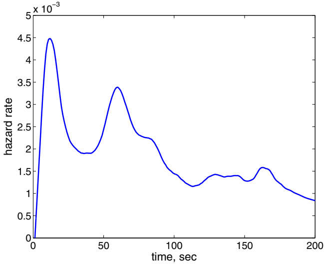


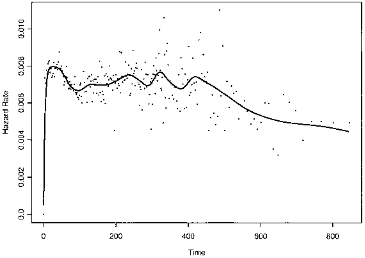
Call Center Data: Hazard Rates (Un-Censored)

(Im)Patience Time τ

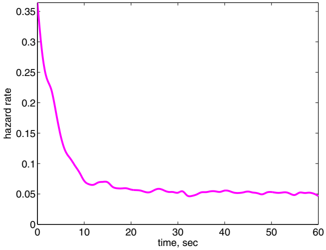
Required/Offered Wait V

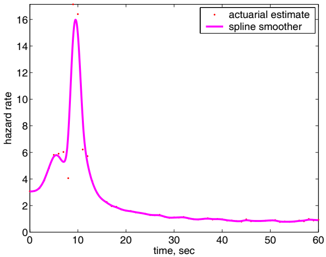
Israel





U.S.

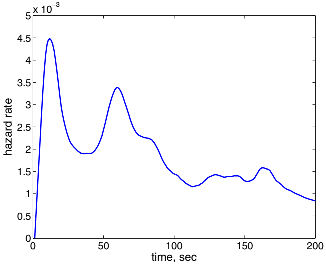


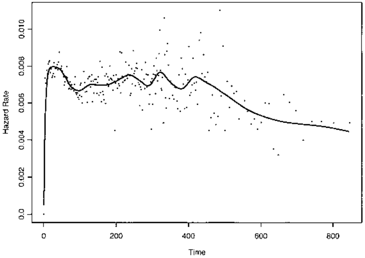


Call Center Data: Hazard Rates (Un-Censored)

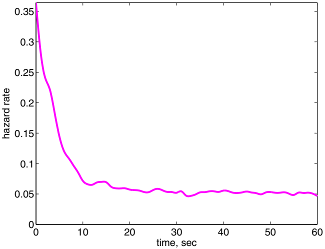
(Im)Patience Time τ | **Required/Offered Wait V**

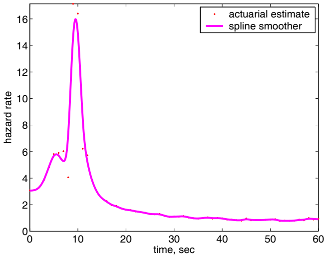
Israel





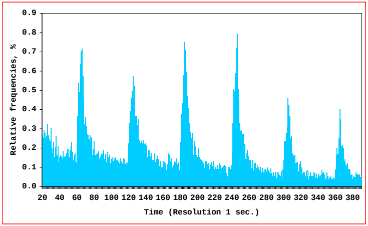
U.S.





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

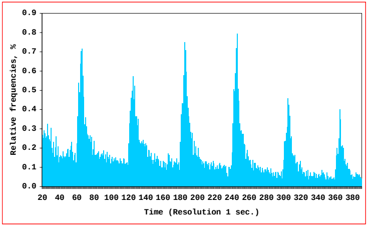
A “Waiting-Times" Puzzle at a Large Israeli Bank



Peaks Every 60 Seconds. Why?

- Human: **Voice-announcement** every 60 seconds.

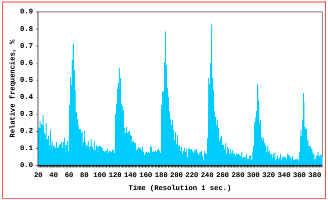
A “Waiting-Times" Puzzle at a Large Israeli Bank



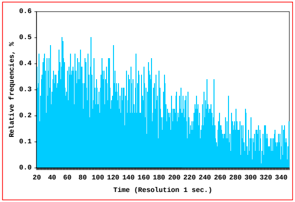
Peaks Every 60 Seconds. Why?

- 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** $W_q = \min\{\tau, V\}$.

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** $W_q = \min\{\tau, V\}$.

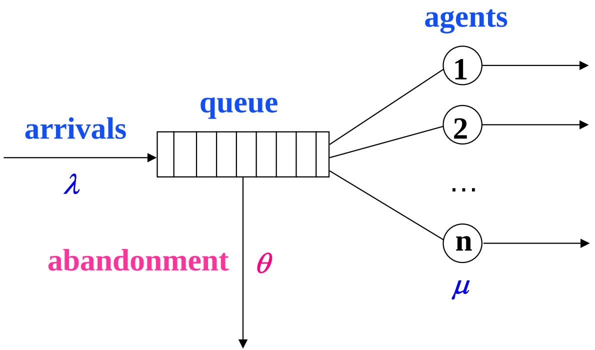
Modeling: τ = **input** to the model, V = **output**.

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**.

Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?

Testing the Erlang-A Primitives

- **Arrivals**: Poisson?
- **Service-durations**: Exponential?
- **(Im)Patience**: Exponential?

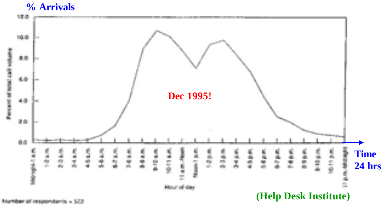
- Primitives independent?
- Customers / Servers Heterogeneous?
- Service discipline FCFS?
- . . . ?

Validation: Support? Refute?

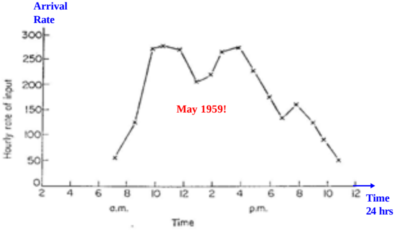
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

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



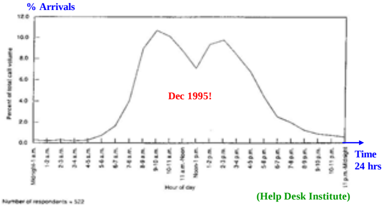
May **1959** (England)



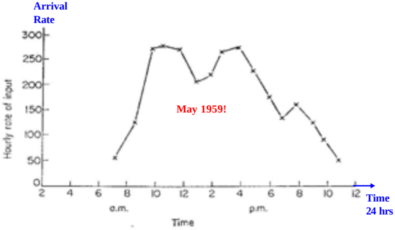
Arrivals to Service: only Poisson-Relatives

Arrival Rate to Three Call Centers

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



May **1959** (England)



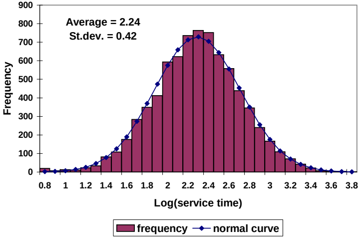
November **1999** (Israel)

Observation:

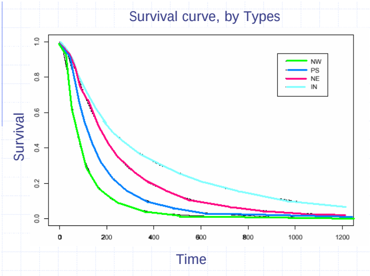
Peak Loads at **10:00 & 15:00**

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



Survival-Functions by Service-Class

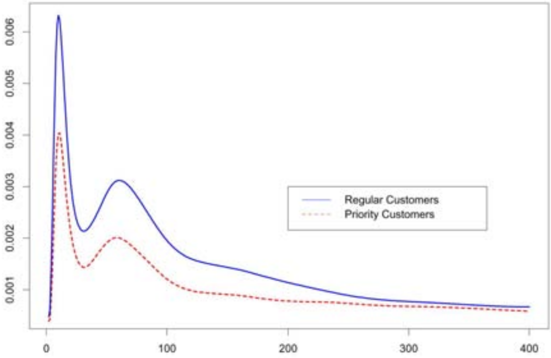


- **New** Customers: 2 min (NW);
- **Regulars**: 3 min (PS);
- **Stock**: 4.5 min (NE);
- Tech-Support: 6.5 min (IN).

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

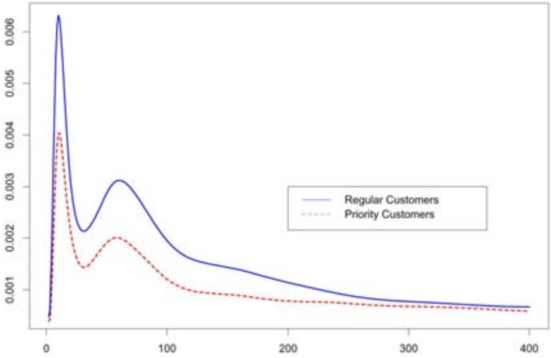
(Im)Patience while Waiting (Palm 1943-53)

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



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

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



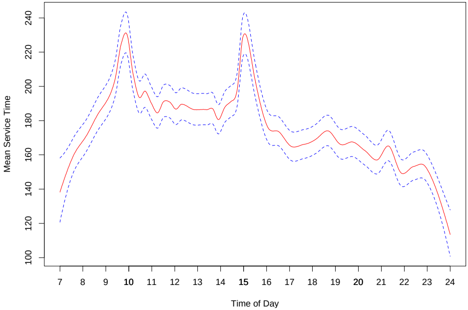
- **Peaks** of abandonment at times of **Announcements**
- **Call-by-Call Data (DataMOCCA)** required (& Un-Censoring).

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

A “Service-Time" Puzzle at an Israeli Bank

Inter-related Primitives

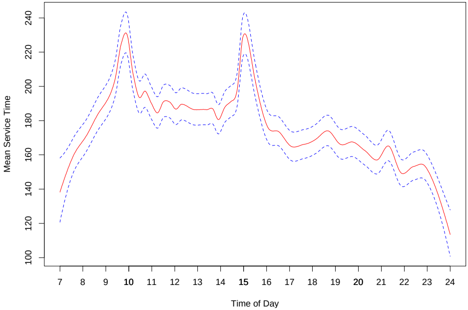
Average Service Time over the Day – Israeli Bank



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

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?**

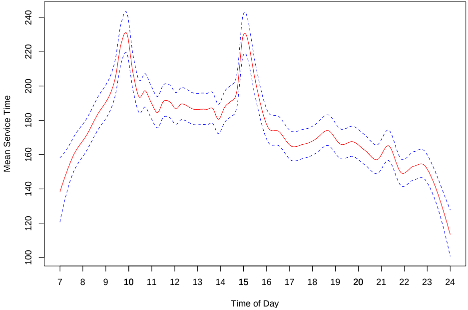
Explanations:

- Common: Service protocol different (longer) during peak times.

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.

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).

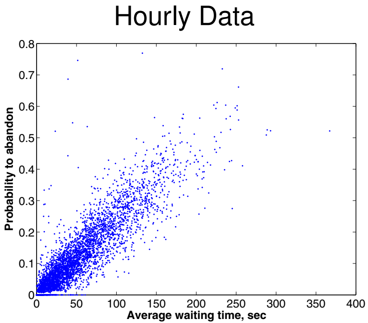
- Building Blocks need **not be independent** (eg. long wait possibly implies long service)
- Customers and Servers **not homogeneous** (classes, skills)
- Customers return for service (after busy, abandonment)
- $\cdots$, and more.

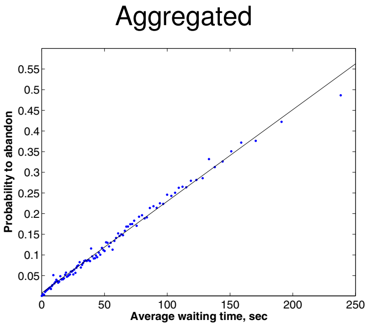
Question: **Is Erlang-A Practically Relevant?**

Estimating (Im)Patience: via P{Ab} $\propto$ E[W_q]

Assume **Exp(θ)** (im)patience. Then, $P\{Ab\} = \theta \cdot E[W_q]$.

Israeli Bank: Yearly Data





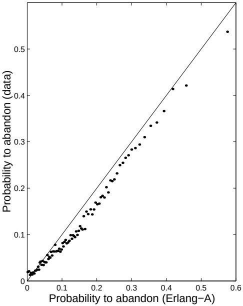
Graphs based on 4158 hour intervals.

Estimate of mean (im)patience: 250/0.55 $\approx$ **450 seconds**.

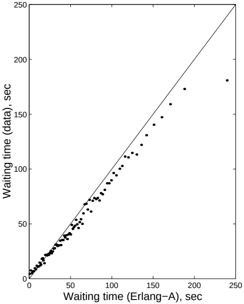
Erlang-A: Fitting a Simple Model to a Complex Reality

- **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).

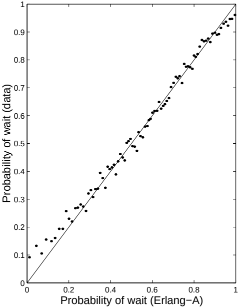
P{Ab}



$E[W_q]$



$P\{W_q > 0\}$



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 , . . .

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.

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** $R = \frac{\lambda}{\mu}$ $(= \lambda \times \mathrm{E}[S])$ not too small.

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

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

ED Regime: $N \approx R - \gamma R$

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

Operational Regimes: Conceptual Framework

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

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

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

ED Regime: $N \approx R - \gamma R$

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

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

- Erlang 1924, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

Operational Regimes: Conceptual Framework

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

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

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

ED Regime: $N \approx R - \gamma R$

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

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

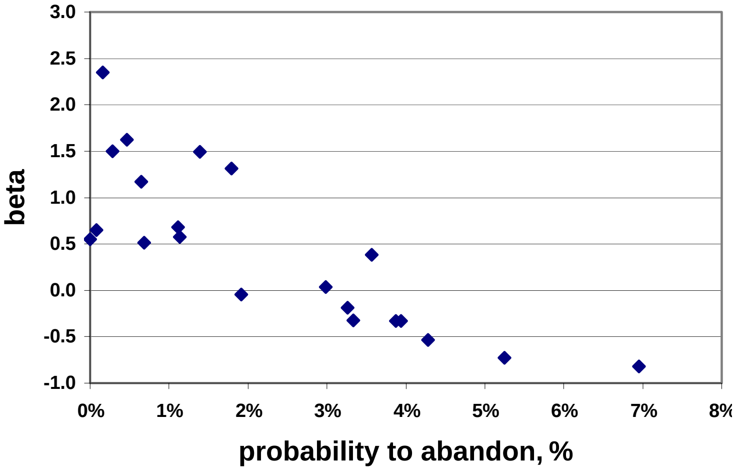
- Erlang 1924, Halfin & Whitt 1981
- %Delayed between 25% and 75%
- Wait one-order below service-time (sec vs. min); 1-5% Abandon.

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

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

QED: Practical Support

QOS parameter $\beta = (N - R)/\sqrt{R}$ vs. %Abandonment



QED: Theoretical Support (Garnett, M., 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.

QED Approximations (Zeltyn, M. '06)

G – patience distribution,

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

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

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

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

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

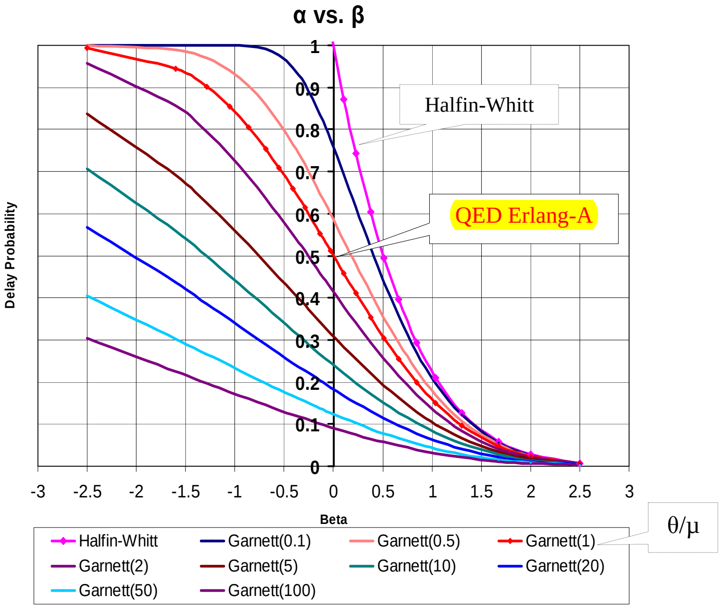
Here

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

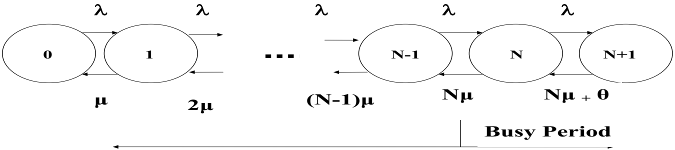
$$\bar{\Phi}(x) = 1 - \Phi(x),$$

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

Garnett / Halfin-Whitt Functions: $P\{W_q > 0\}$



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 $$T_{N-1,N} = \frac{1}{\lambda_N E_{1,N-1}} \sim \frac{1}{N\mu \times h(-\beta)/\sqrt{N}} \sim \frac{1}{\sqrt{N}} \cdot \frac{1/\mu}{h(-\beta)}$$

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

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

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

QED Intuition via Excursions: Asymptotics

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

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

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 cases:

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

QED Intuition via Excursions: Asymptotics

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

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

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 cases:

- $\boldsymbol{\mu = \theta}$: $Q \overset{d}{=} M/M/\infty$, since sojourn-time always $\exp(\mu = \theta)$.
- $\boldsymbol{\beta = 0}$ ($N \approx R$): $P\{Wait > 0\} \approx [1 + \sqrt{\theta/\mu}]^{-1}$.

QED Intuition via Excursions: Asymptotics

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

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

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 cases:

- $\mu = \theta$: $Q \stackrel{d}{=} M/M/\infty$, since sojourn-time always $\exp(\mu = \theta)$.
- $\beta = 0$ ($N \approx R$): $P\{Wait > 0\} \approx [1 + \sqrt{\theta/\mu}]^{-1}$.
- **Both** of the above: $P\{Wait > 0\} \approx 1/2$.

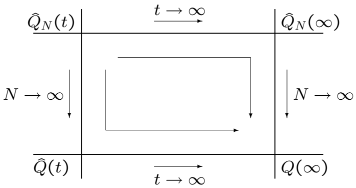
Process Limits (Queueing, Waiting)

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



Approximating (Virtual) Waiting Time

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

Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server **Asymptotic Analysis** (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are **extremely accurate**.

Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server **Asymptotic Analysis** (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are **extremely accurate.**
2. **Dimensioning**:
   - **Cost / Profit Optimization**: eg. Min costs of Staffing + Congestion.
   - **Constraint Satisfaction**: eg. **Min. N , s.t. QOS constraints** .

Dimensioning a Service System

Operational Regimes provide a **conceptual framework**.

Questions:

1. How accurate are QD/ED/QED approximations?
2. How to determine the regime? QOS parameters?
3. Is there a regime robust enough to cover the others?

Answers, via many-server **Asymptotic Analysis** (w/ Borst & Reiman, 2004; Zeltyn, 2006):

1. Approximations are **extremely accurate.**
2. **Dimensioning**:
   - **Cost / Profit Optimization**: eg. Min costs of Staffing + Congestion.
   - **Constraint Satisfaction**: eg. **Min. N , s.t. QOS constraints**.
3. Robustness depends:
   - **Without Abandonment: QED** covers all, at amazing accuracy.
   - **With Abandonment: ED, QED, ED+QED** all have a role.

Operational Regimes: Rules-of-Thumb

Constraint	P{Ab}		E[W]		P{$W > T$}	
Offered Load	Tight 1-10%	Loose $\geq$ 10%	Tight $\leq 10\%\text{E}[\tau]$	Loose $\geq 10\%\text{E}[\tau]$	Tight $0 \leq T \leq 10\%\text{E}[\tau]$ $5\% \leq \alpha \leq 50\%$	Loose $T \geq 10\%\text{E}[\tau]$ $5\% \leq \alpha \leq 50\%$
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.

Back to "Why does Erlang-A Work?"

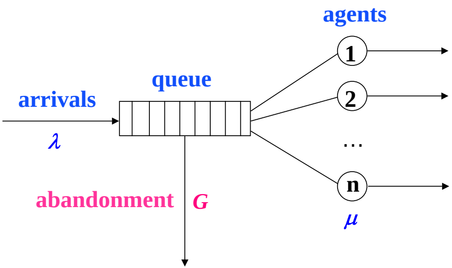
Theoretical Answer: $M_t^J/G/N_t + G \overset{d}{\approx} (M/M/N+M)_t\ , \quad t \geq 0.$

- **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 (stochastically-challenged) Call Centers work?

"The right answer for the wrong reason"

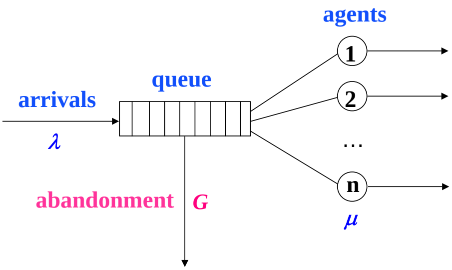
“Why does Erlang-A Work?" General Patience



(Im)Patience times **Generally Distributed**: M/M/$\boldsymbol{n}$+G

Exact analysis in steady-state (Baccelli & Hebuterne, 1981): solve Kolmogorov’s PDE’s (semi-Markov) for the offered-wait $\boldsymbol{V}$.

"Why does Erlang-A Work?" General Patience



(Im)Patience times **Generally Distributed**: M/M/n+G

Exact analysis in steady-state (Baccelli & Hebuterne, 1981): solve Kolmogorov's PDE's (semi-Markov) for the offered-wait V.

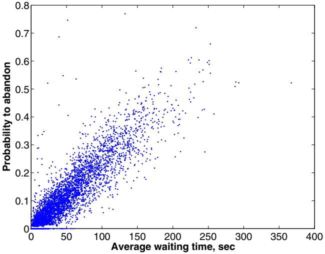
QED analysis (w/ Zeltyn, 2006): $n \approx R + \beta\sqrt{R}$.

- Assume (Im)Patience density $g(0) > 0$.
- V asymptotics ($\lambda \uparrow \infty$): Laplace Method, leading to
- **QED Approximations: Use Erlang-A** as is, with $\theta \leftrightarrow g(0)$.

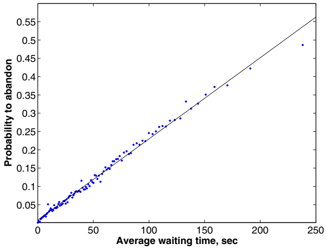
General Patience: Fitting Erlang-A

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Theory:

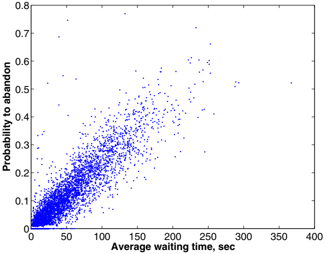
Erlang-A: P{Ab} $= \theta \cdot$ E[W_q];

M/M/N+G: P{Ab} $\approx g(0) \cdot$ E[W_q].

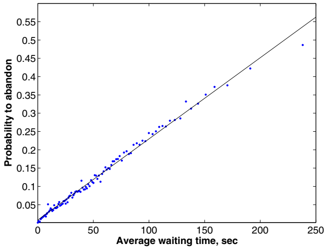
General Patience: Fitting Erlang-A

Israeli Bank: Yearly Data

Hourly Data



Aggregated



Theory:

Erlang-A: $\mathsf{P}\{\mathsf{Ab}\} = \theta \cdot \mathsf{E}[W_q]$; **M/M/*N*+G:** $\mathsf{P}\{\mathsf{Ab}\} \approx g(0) \cdot \mathsf{E}[W_q]$.

Recipe:

In both cases, use Erlang-A, with $\hat{\theta} = \widehat{\mathsf{P}\{\mathsf{Ab}\}} / \widehat{\mathsf{E}[W_q]}$ (slope above).

Why Does Erlang-A Work? General Services

Established: M/M/N+**G** $\approx$ M/M/N+M ($\theta = g(0)$).

Why Does Erlang-A Work? General Services

Established: M/M/N+**G** $\approx$ M/M/N+M ($\theta = g(0)$).

Now: M/**G**/N+G $\approx$ M/M/N+G (E[S] same in both).

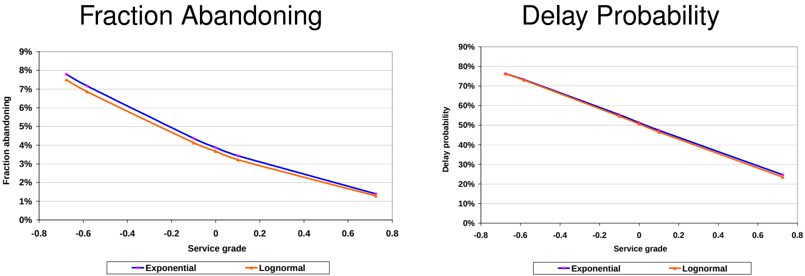
Why Does Erlang-A Work? General Services

Established: M/M/N+**G** $\approx$ M/M/N+M ($\theta = g(0)$).

Now: M/**G**/N+G $\approx$ M/M/N+G (E[S] same in both).

Numerical Experiments: Whitt (2004), Rosenshmidt (2006) demonstrate a **useful fit** for typical call-center parameters.

Lognormal (CV=1) vs. Exponential Service Times, QED Regime; 100 agents, average patience = average service



Why Does Erlang-A Work? General Services

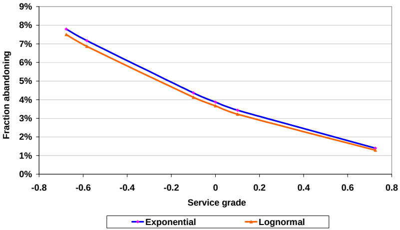
Established: M/M/N+**G** $\approx$ M/M/N+M ($\theta = g(0)$).

Now: M/**G**/N+G $\approx$ M/M/N+G (E[S] same in both).

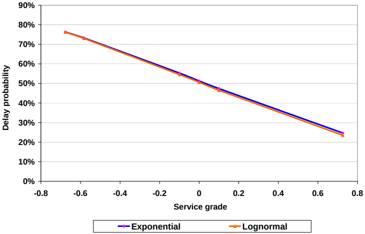
Numerical Experiments: Whitt (2004), Rosenshmidt (2006) demonstrate a **useful fit** for typical call-center parameters.

Lognormal (CV=1) vs. Exponential Service Times, QED Regime; 100 agents, average patience = average service

Fraction Abandoning



Delay Probability



QED G-Services: G/D_K/N+G (w/ Momčilović, ongoing).

Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: $M/G/N+\mathrm{G} \approx M/M/N+\mathrm{M}$ $(\theta = g(0))$.

Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: $M/G/N$+G $\approx M/M/N$+M $(\theta = g(0))$.

Now: $M_t/G/N_t + G \approx (M/G/N + G)_t$ $(N_t, \lambda$ well chosen).

Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: $M/G/N$+G $\approx M/M/N$+M $(\theta = g(0))$.

Now: $M_t/G/N_t + G \approx (M/G/N + G)_t$ $(N_t, \lambda$ well chosen).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. Modified Offered-Load: $\boldsymbol{\lambda}$
   - Consider $M_t/G/N_t + G$ with arrival rate $\boldsymbol{\lambda(t), t \geq 0}$.
   - Approximate its **time-varying** performance **at time *t*** with a **stationary** $\boldsymbol{M/G/N_t + G}$, in which $\boldsymbol{\lambda = \mathrm{E}\lambda(t - S_e)}$.
     ($\boldsymbol{S_e \stackrel{d}{=}}$ residual-service: congestion-lag behind peak-load.)

Time-Varying Arrival Rates

Established: $M/G/N$+G $\approx M/M/N$+M $(\theta = g(0))$.

Now: $M_t/G/N_t + G \approx (M/G/N + G)_t$ $(N_t, \lambda$ well chosen).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. Modified Offered-Load: $\boldsymbol{\lambda}$
   - Consider $M_t/G/N_t + G$ with arrival rate $\boldsymbol{\lambda(t), t \geq 0}$.
   - Approximate its **time-varying** performance **at time *t*** with a **stationary** $\boldsymbol{M/G/N_t + G}$, in which $\boldsymbol{\lambda = \mathrm{E}\lambda(t - S_e)}$.
     ($\boldsymbol{S_e} \stackrel{d}{=}$ residual-service: congestion-lag behind peak-load.)
2. Square-Root Staffing: $\boldsymbol{N_t}$
   - Let $\boldsymbol{R_t = \mathrm{E}\lambda(t - S_e) \times \mathrm{E}S}$ be the **Offered-Load** at time $\boldsymbol{t}$
     ($\boldsymbol{R_t}$ = Number-in-system in a corresponding $\boldsymbol{M_t/G/\infty}$.)
   - Staff $\boldsymbol{N_t = R_t + \beta\sqrt{R_t}}$.

Why Does Erlang-A Work? Time-Varying Arrival Rates

Established: $M/G/N{+}G \approx M/M/N{+}M \quad (\theta = g(0))$.

Now: $M_t/G/N_t + G \approx (M/G/N + G)_t \quad (N_t, \lambda$ well chosen).

Two steps (Feldman, M., Massey & Whitt, 2006):

1. Modified Offered-Load: λ
   - Consider $M_t/G/N_t + G$ with arrival rate $\lambda(t), t \geq 0$.
   - Approximate its **time-varying** performance **at time** t with a **stationary** $M/G/N_t + G$, in which $\lambda = E\lambda(t - S_e)$.
     ($S_e \stackrel{d}{=}$ residual-service: congestion-lag behind peak-load.)
2. Square-Root Staffing: N_t
   - Let $R_t = E\lambda(t - S_e) \times ES$ be the **Offered-Load** at time t
     (R_t = Number-in-system in a corresponding $M_t/G/\infty$.)
   - Staff $N_t = R_t + \beta\sqrt{R_t}$.

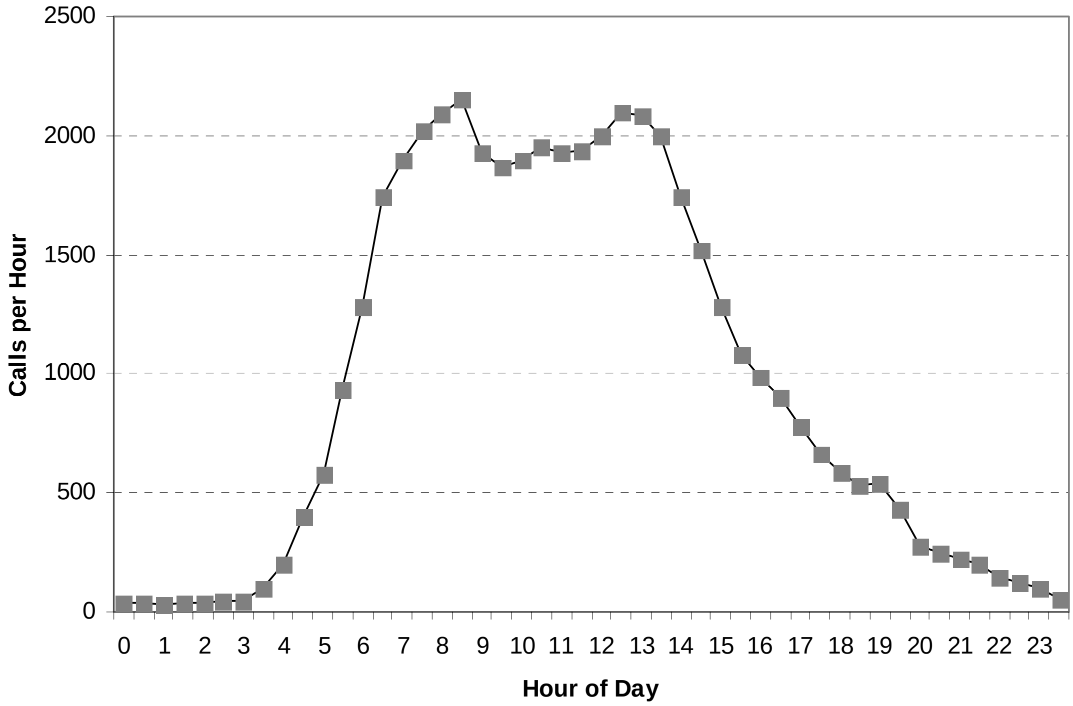
Serendipity: Time-stable performance, supported by **ISA** = Iterative Staffing Algorithm, and QED diffusion limits ($M_t/M/N + M, \ \mu = \theta$).

Example: "Real" Call Center

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

Time-Varying (two-hump) arrival functions common

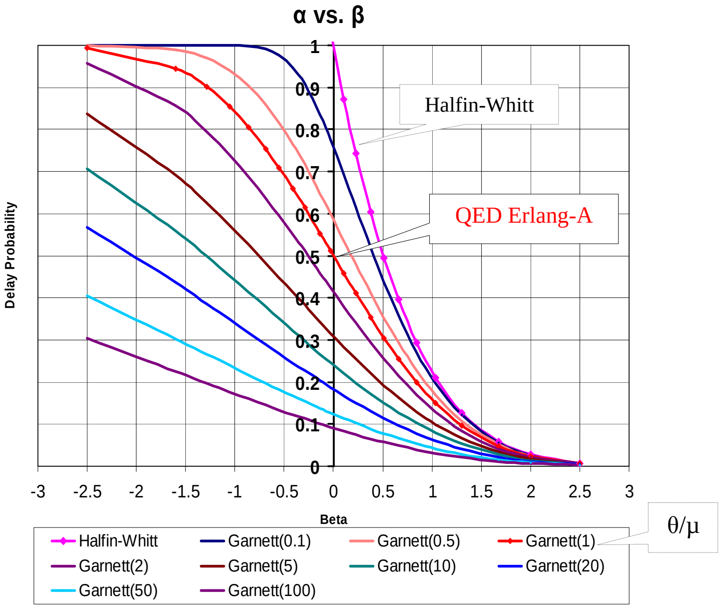
(Adapted from Green L., Kolesar P., Soares J. for benchmarking.)



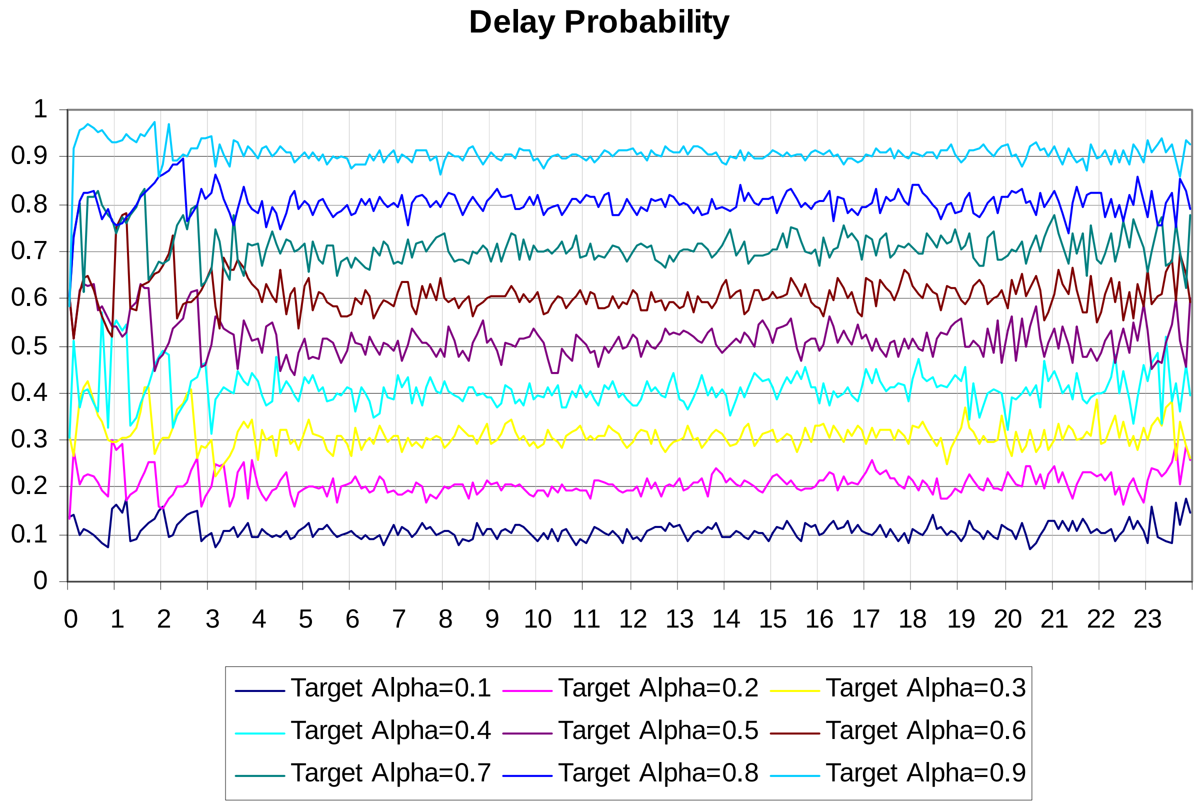
Assume: Service and abandonment times are both

Exponential, with mean 0.1 (6 min.)

Garnett / Halfin-Whitt Functions: $P\{W_q > 0\}$

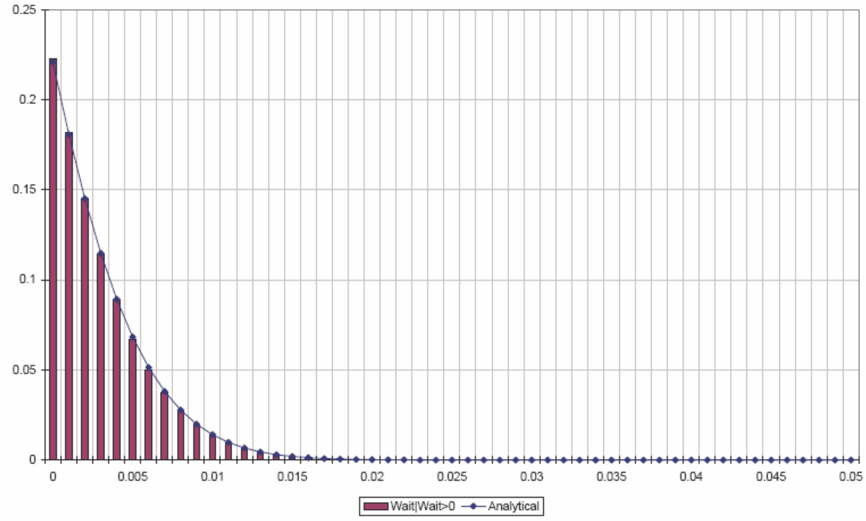


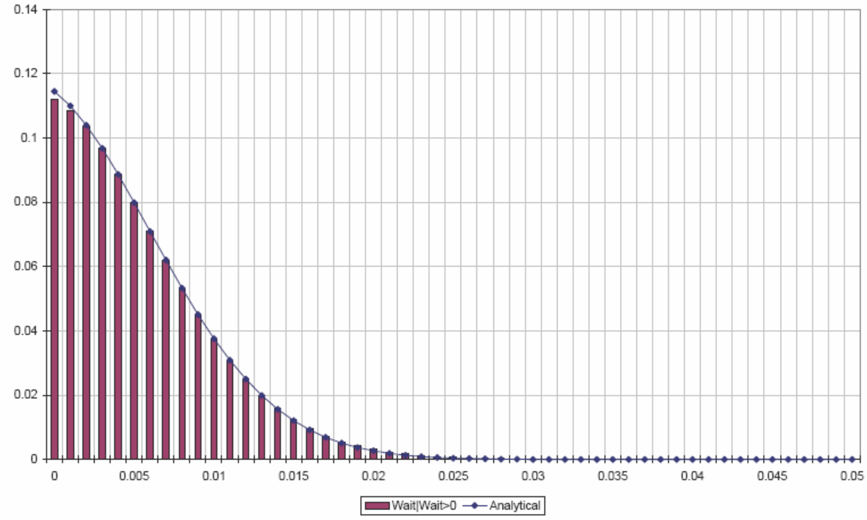
Delay Probability α

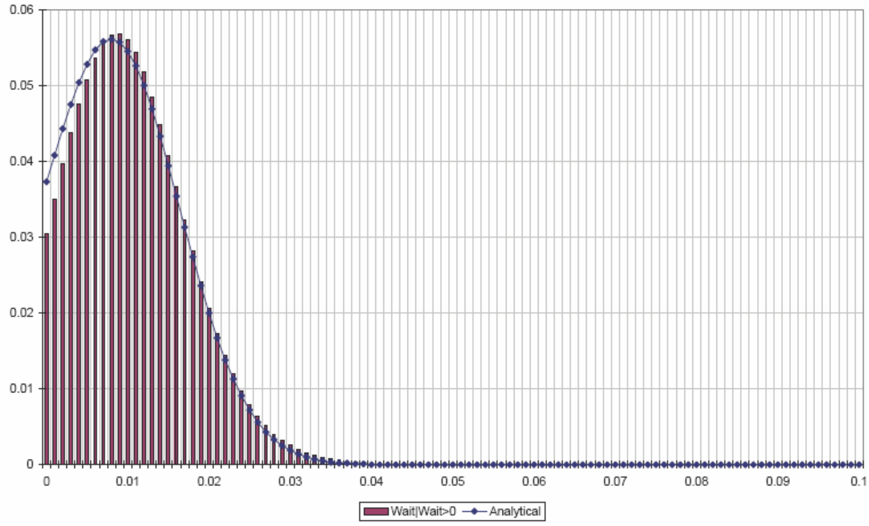


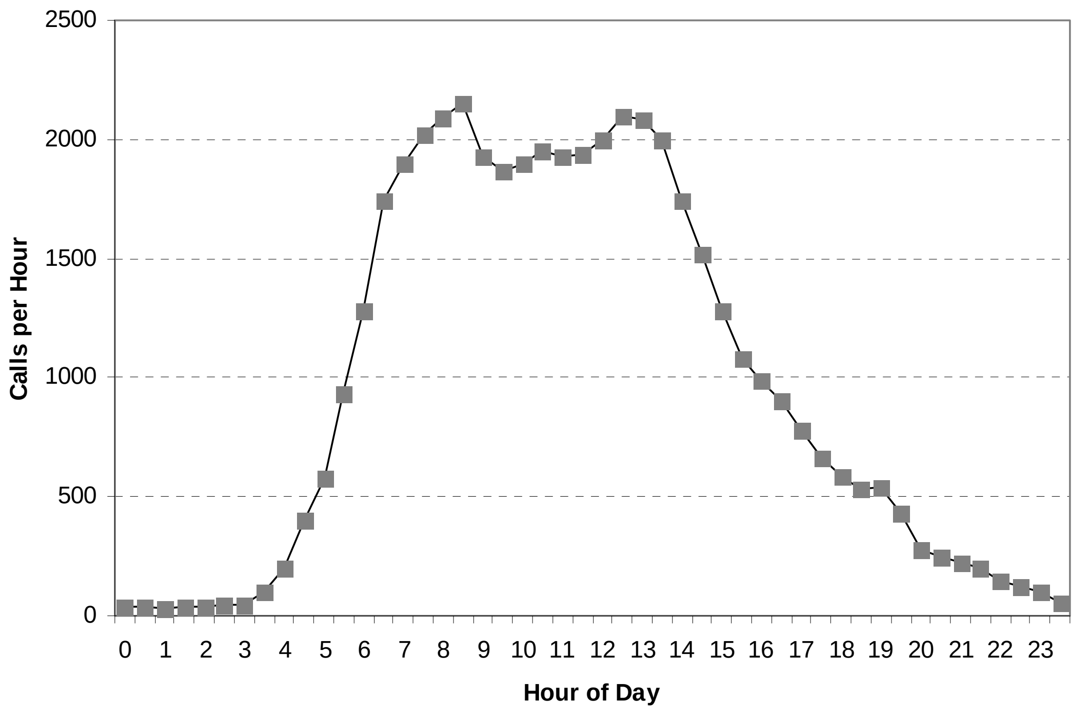
Real Call Center: Empirical waiting time, given positive wait

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

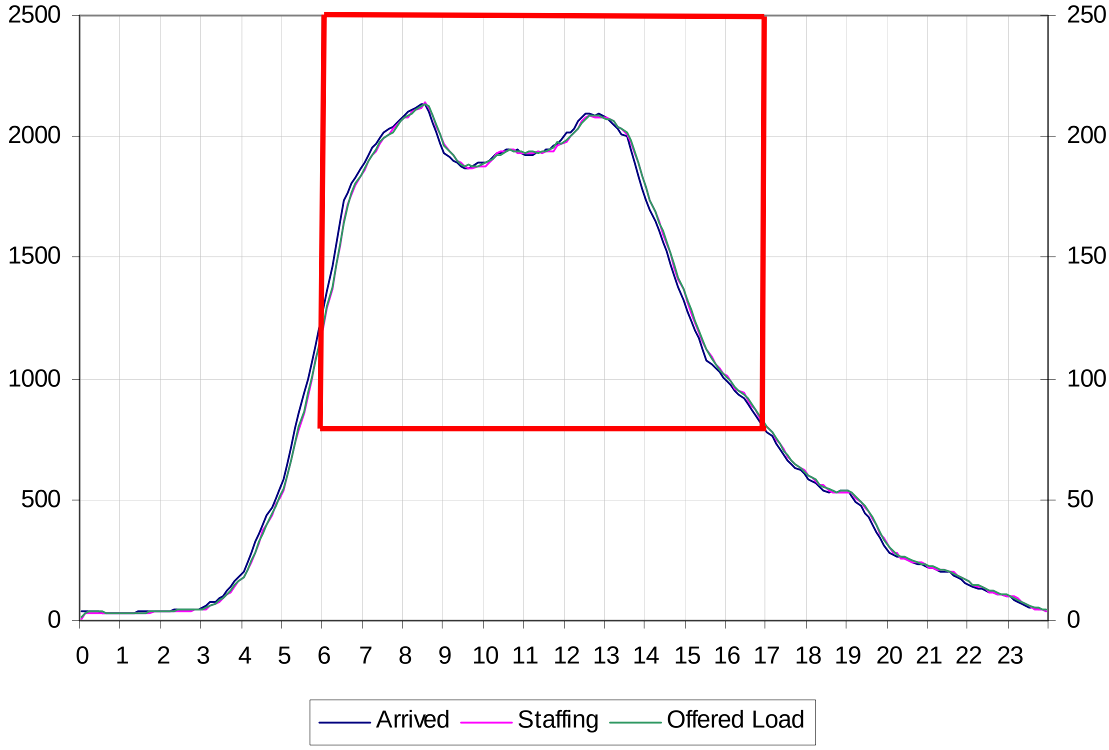








QED Staffing (β=0 iff α=0.5)



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

Prevalent Practice $$N_t = \lceil \lambda(t) \cdot E(S) \rceil$$ (PSA)

"Right Answer" $$N_t \approx R_t + \beta \cdot \sqrt{R_t}$$ (MOL)

$$R_t = E\lambda(t-S) \cdot E(S)$$

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

Prevalent Practice $$N_t = \lceil \lambda(t) \cdot E(S) \rceil$$ (PSA)

"Right Answer" $$N_t \approx R_t + \beta \cdot \sqrt{R_t}$$ (MOL)

$$R_t = E\lambda(t - S) \cdot E(S)$$

Practice ≈ "Right" $\beta \approx 0$ (QED)

and $\lambda(t) \approx$ stable over service-durations

Practice Improved $$N_t = \lceil \lambda[t - E(S)] \cdot E(S) \rceil$$

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

Prevalent Practice $N_t \;=\; \lceil \lambda(t) \cdot E(S) \rceil$ (PSA)

"Right Answer" $N_t \;\approx\; R_t + \beta \cdot \sqrt{R_t}$ (MOL)

$$R_t = E\lambda(t-S) \cdot E(S)$$

Practice ≈ "Right" $\beta \approx 0$ (QED)

and $\lambda(t) \approx$ stable over service-durations

Practice Improved $N_t \;=\; \lceil \lambda[t - E(S)] \cdot E(S) \rceil$

When Optimal ? for moderately-patient customers:

1. Satisfization ⇔ At least 50% to be serve immediately

2. Optimization ⇔ Customer-Time = 2 x Agent-Salary

Why Does Erlang-A Work? Multi-Class Customers

Now: $M_t^J/G/N_t + G \approx (M^J/G/N + G)_t$ (well staffed & controlled).

Service Levels: Class 1 = **VIP** , . . ., Class J = **best-effort**.

Why Does Erlang-A Work? Multi-Class Customers

Now: $M_t^J/G/N_t + G \approx (M^J/G/N+G)_t$ (well staffed & controlled).

Service Levels: Class 1 = **VIP** , . . ., Class J = **best-effort**.

Staffing, Control (w/ Gurvich & Armony 2005; Feldman & Gurvich):

- Consider $M_t^J/G/N_t + G$ with arrival rates $\lambda_j(t), t \geq 0$.
- Assume **i.i.d. servers.**
- Let $R_t = \mathrm{E}\sum_j \lambda_j(t - S_e) \times \mathrm{E}S$ be the **Offered-Load** at time t.

Why Does Erlang-A Work? Multi-Class Customers

Now: $M_t^J/G/N_t + G \approx (M^J/G/N+G)_t$ (well staffed & controlled).
Service Levels: Class 1 = **VIP** , . . ., Class J = **best-effort**.

Staffing, Control (w/ Gurvich & Armony 2005; Feldman & Gurvich):

- Consider $M_t^J/G/N_t + G$ with arrival rates $\lambda_j(t), t \geq 0$.
- Assume **i.i.d. servers.**
- Let $R_t = \mathrm{E}\sum_j \lambda_j(t - S_e) \times \mathrm{E}S$ be the **Offered-Load** at time t.
- **Staff** $N_t = R_t + \beta\sqrt{R_t}$, with β determined by a desired QED performance for the lowest-priority class J.
- **Control** via threshold priorities, where the thresholds are determined by ISA according to desired service levels.
- Approximate **time-varying** performance **at time t** with a **stationary** threshold-controlled $M^J/G/N_t + G$, in which $\lambda_j = \mathrm{E}\lambda_j(t - S_e)$.

Why Does Erlang-A Work? Multi-Class Customers

Now: $M_t^J/G/N_t + G \approx (M^J/G/N+G)_t$ (well staffed & controlled).
Service Levels: Class 1 = **VIP** , . . ., Class J = **best-effort**.

Staffing, Control (w/ Gurvich & Armony 2005; Feldman & Gurvich):

- Consider $M_t^J/G/N_t + G$ with arrival rates $\lambda_j(t), t \geq 0$.
- Assume **i.i.d. servers.**
- Let $R_t = \mathrm{E} \sum_j \lambda_j(t - S_e) \times \mathrm{E}S$ be the **Offered-Load** at time t.
- **Staff** $N_t = R_t + \beta\sqrt{R_t}$, with β determined by a desired QED performance for the lowest-priority class ***J***.
- **Control** via threshold priorities, where the thresholds are determined by ISA according to desired service levels.
- Approximate **time-varying** performance **at time *t*** with a **stationary** threshold-controlled $\boldsymbol{M^J/G/N_t + G}$, in which $\boldsymbol{\lambda_j = \mathrm{E}\lambda_j(t - S_e)}$.

Serendipity: Multi-Class Multi-Skill, w/ **class-dependent** services.
Support: ISA, QED diffusion limits (Atar, M. & Shaikhet, 2007).

Additional Simple (QED) Models of Complex Realities: Exponential Services; i.i.d. Customers, i.i.d. Servers

- **Performance Analysis**:
  - Khudiakova, Feigin, M. (Semi-Open): Call-Center + IVR/VRU;
  - De Véricourt, Jennings (Closed + Delay), then w/ Yom-Tov (Semi-Open): **Nurse staffing** (ratios), bed sizing;
  - Randhawa, Kumar (Closed + Loss): Subscriber queues.
- **Optimal Staffing**: Accurate to **within 1**, even with very small *n*'s, for both **constraint-satisfaction** and cost/revenue **optimization** (staffing, abandonment and waiting costs).
  - Armony, Maglaras: (M_x/M/N) Delay information (Equilibrium);
  - Borst, M., Reiman (M/M/N): Asymptotic framework;
  - Zeltyn, M. (M/M/N+G): Optimization still ongoing.
- **Time-Varying Queues**, via 2 approaches:
  - Jennings, M., Massey, Whitt, then w/ Feldman: **Time-Stable Performance** (ISA, leading to Modified Offered Load);
  - M., Massey, Reiman, Rider, Stolyar: Unavoidable **Time-Varying Performance** (Fluid & Diffusion models, via Uniform Acceleration).

Less-Simple (QED) Models: General Service-Times

The Challenge: Must keep track of the state of ***n* individual servers, as *n*** $\uparrow \infty$. (Recall Kiefer & Wolfowitz).

- Shwartz, M. (M/G/N), Rosenshmidt, M. (M/G/N+G): Simulations; **LogNormal better then Exp**, 2-valued same as D.
- Whitt (GI/M+0/N): Covering $CV \geq 1$;
- Puhalskii, Reiman (GI/PH/N): Markovian process-limits (no steady-state); also priorities;
- Jelencović, M., Momčilović (GI/D/N): steady-state (via round-robin); then M., Momčilović (G/D_K/N): process-limits, via "Lindley-Trees"; G/D_K/N+G ongoing.
- Kaspi, Ramanan (G/G/N): Fluid, next Diffusion (measure-valued ages, following Kiefer & Wolfowitz);
- Reed (GI/GI/N): Fluid, Diffusion (**Skorohod-Like Mapping**).

Complex (QED) Models: Skills-Based Routing (Heterogeneous Customers or/and Servers - Theory)

- **V-Model**: Harrison, Zeevi; Atar, M., Reiman; Gurvich, M., Armony;
  then **Class-dependent** services: Atar, M., Shaikhet;
- **Reversed-V**: Armony, M.;
  then **Pool-dependent** services: Dai, Tezcan; Gurvich, Whitt (**G-$c\mu$**); Atar, M., Shaikhet (Abandonment);
- **General**: Atar, then w/ Shaikhet (Null-controllability, Throughput-suboptimality); Gurvich, Whitt (FQR);
- **Distributed Networks**: Tezcan;
- **Random Service Rates**: Atar (Fastest or longest-idle server).

The Technion SEE Center / Laboratory

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**.

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**.

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).

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)**.